

EMNLP

2024



MIAMI, FLORIDA
November 12–16

**The 2024 Conference on
Empirical Methods in
Natural Language Processing**

Table of contents

	Page
1 Conference Information	1
Message from the General Chair	1
Message from the Program Chairs	3
Message from the Local Chair	7
Organizing Committee	9
Senior Program Committee	11
Conference Organizers & Vendors	15
2 Anti-Harassment Policy	16
3 Meal Info	18
Overview	18
4 Welcome Reception	20
5 Social Event Gala Dinner	21
6 Keynotes	25
7 Panel	29
8 Birds-of-a-Feather and Affinity Group Meetup	32
9 Main Conference Overview	33
Monday, November 11 - Registration & Welcome Reception	33
Tuesday, November 12 - Main Conference	33
Wednesday, November 13 - Main Conference	36
Thursday, November 14 - Main Conference	37
Friday, November 15 - Workshop/Tutorial Day	39
Saturday, November 16 - Workshop/Tutorial Day	39

10 Oral Presentations	40
Session 02 - Nov 12 (Tue) 11:00-12:30	40
Language Modeling 1	40
Interpretability and Analysis of Models for NLP 1	41
Interpretability and Analysis of Models for NLP 2	42
Interpretability and Analysis of Models for NLP 1	42
Low-resource Methods for NLP 1	42
Human-centered NLP 1	43
Human-centered NLP 2	44
Human-centered NLP 1	44
Machine Translation 1	45
Session 03 - Nov 12 (Tue) 14:00-15:30	46
Generation and Summarization	46
Dialogue and Interactive Systems 1	47
Dialogue and Interactive Systems 2	48
Dialogue and Interactive Systems 1	48
Computational Social Science and Cultural Analytics 1	48
Special Theme: Efficiency in Model Algorithms, Training, and Inference 1	49
Resources and Evaluation 1	50
Session 04 - Nov 12 (Tue) 16:00-17:30	51
Ethics, Bias, and Fairness 2	52
Information Retrieval and Text Mining 2	53
Multimodality and Language Grounding to Vision, Robotics and Beyond 2	54
Multimodality and Language Grounding to Vision, Robotics and Beyond 1	54
Linguistic Theories, Cognitive Modeling and Psycholinguistics 2	55
Industry	56
Session 06 - Nov 13 (Wed) 10:30-12:00	57
Multimodality and Language Grounding to Vision, Robotics and Beyond 3	57
Ethics, Bias, and Fairness 3	59
Discourse + Phonology + Syntax 2	60
Question Answering 2	61
Question Answering 1	61
Question Answering 2	62
Industry	62
Session 09 - Nov 13 (Wed) 16:00-17:30	64
Resources and Evaluation 4	64
Interpretability and Analysis of Models for NLP 4	65
NLP Applications 3	66
Information Extraction 1	67
Machine Learning for NLP 2	68
Session 11 - Nov 14 (Thu) 10:30-12:00	69
NLP Applications 4	69
Computational Social Science and Cultural Analytics 3	71
Sentiment and Semantics	72
Language Modeling 4	73
Language Modeling 2	73
Language Modeling 4	74
Multilinguality and Language Diversity 2	74
Multilinguality and Language Diversity 1	75
Multilinguality and Language Diversity 2	75
Session 12 - Nov 14 (Thu) 14:00-15:30	75
Interpretability and Analysis of Models for NLP 6	75
Speech Processing and Spoken Language Understanding 2	76

Resources and Evaluation 6	77
Resources and Evaluation 2	78
Resources and Evaluation 6	78
Generation 3	79
Machine Learning for NLP 4	80
11 Posters and Demos	82
Session 02 - Nov 12 (Tue) 11:00-12:30	82
Demo	82
Generation 1	83
Industry	88
Information Retrieval and Text Mining 1	92
Linguistic Theories, Cognitive Modeling and Psycholinguistics 1	96
Multimodality and Language Grounding to Vision, Robotics and Beyond 1	100
NLP Applications 1	106
Session 03 - Nov 12 (Tue) 14:00-15:30	114
Demo	114
Discourse + Phonology + Syntax 1	115
Ethics, Bias, and Fairness 1	119
Interpretability and Analysis of Models for NLP 2	124
Language Modeling 2	131
Machine Learning for NLP 1	136
Multilinguality and Language Diversity 1	141
Session 04 - Nov 12 (Tue) 16:00-17:30	148
Computational Social Science and Cultural Analytics 2	148
Demo	153
Machine Translation 2	154
Question Answering 1	158
Resources and Evaluation 2	163
Sentiment Analysis, Stylistic Analysis, and Argument Mining	172
Special Theme: Efficiency in Model Algorithms, Training, and Inference 2	176
Summarization	182
Session 06 - Nov 13 (Wed) 10:30-12:00	185
Demo	185
Human-centered NLP 2	186
Interpretability and Analysis of Models for NLP 3	190
Low-resource Methods for NLP 2	196
NLP Applications 2	202
Resources and Evaluation 3	210
Speech Processing and Spoken Language Understanding 1	219
Session 09 - Nov 13 (Wed) 16:00-17:30	222
Demo	222
Dialogue and Interactive Systems 2	223
Industry	228
Information Retrieval and Text Mining 3	232
Language Modeling 3	236
Multimodality and Language Grounding to Vision, Robotics and Beyond 4	242
Question Answering 3	248
Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas	252
TACL + CL	255
Session 11 - Nov 14 (Thu) 10:30-12:00	257
Demo	257
Ethics, Bias, and Fairness 4	258

Generation 2	262
Interpretability and Analysis of Models for NLP 5	268
Machine Learning for NLP 3	275
Resources and Evaluation 5	279
Special Theme: Efficiency in Model Algorithms, Training, and Inference 3	288
Session 12 - Nov 14 (Thu) 14:00-15:30	294
Computational Social Science and Cultural Analytics 4	294
Dialogue and Interactive Systems 3	300
Industry	304
Information Extraction 2	308
Multimodality and Language Grounding to Vision, Robotics and Beyond 5	314
NLP Applications 5	320
Virtual Poster Session 1 - (Nov 12): 17:45:18:45 (Evening)	329
Computational Social Science and Cultural Analytics	329
Demo	329
Dialogue and Interactive Systems	330
Ethics, Bias, and Fairness	330
Generation 1	331
Information Extraction	331
Information Retrieval and Text Mining	332
Interpretability and Analysis of Models for NLP	332
Language Modeling	333
Linguistic Theories, Cognitive Modeling and Psycholinguistics	334
Low-resource Methods for NLP	334
Machine Learning for NLP 1	334
Machine Translation	335
Multilinguality and Language Diversity	336
Multimodality and Language Grounding to Vision, Robotics and Beyond	336
NLP Applications 1	337
Question Answering	338
Resources and Evaluation	339
Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas	340
Sentiment Analysis, Stylistic Analysis, and Argument Mining	341
Special Theme: Efficiency in Model Algorithms, Training, and Inference	341
Speech Processing and Spoken Language Understanding	342
Syntax: Tagging, Chunking and Parsing	342
Virtual Poster Session 2 - (Nov 13): 7:45:18:45 (Morning)	342
Computational Social Science and Cultural Analytics	342
Demo	343
Dialogue and Interactive Systems	346
Discourse and Pragmatics	348
Ethics, Bias, and Fairness	348
Generation	350
Industry	352
Information Extraction	358
Information Retrieval and Text Mining	359
Interpretability and Analysis of Models for NLP	360
Language Modeling	363
Linguistic Theories, Cognitive Modeling and Psycholinguistics	366
Low-resource Methods for NLP	367
Machine Learning for NLP	370
Machine Translation	371
Multilinguality and Language Diversity	371

Multimodality and Language Grounding to Vision, Robotics and Beyond	372
NLP Applications	377
Question Answering	383
Resources and Evaluation	386
Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas	389
Sentiment Analysis, Stylistic Analysis, and Argument Mining	391
Special Theme: Efficiency in Model Algorithms, Training, and Inference	391
Speech Processing and Spoken Language Understanding	393
Summarization	394
Syntax: Tagging, Chunking and Parsing	395
NLP Applications	395
Virtual Poster Session 3 - (Nov 14): 13:0014:00 (Afternoon)	395
Computational Social Science and Cultural Analytics	395
Demo	397
Ethics, Bias, and Fairness	398
Generation	399
Industry	399
Information Extraction	400
Information Retrieval and Text Mining	401
Interpretability and Analysis of Models for NLP	401
Language Modeling	402
Linguistic Theories, Cognitive Modeling and Psycholinguistics	402
Low-resource Methods for NLP	403
Machine Learning for NLP	403
Machine Translation	403
Multimodality and Language Grounding to Vision, Robotics and Beyond	404
NLP Applications	404
Question Answering	406
Resources and Evaluation	406
Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas	407
Sentiment Analysis, Stylistic Analysis, and Argument Mining	408
Special Theme: Efficiency in Model Algorithms, Training, and Inference	408
Speech Processing and Spoken Language Understanding	409
Syntax: Tagging, Chunking and Parsing	409
12 Tutorials: Friday, November 15, 2024	410
Overview	410
13 Tutorials: Saturday, November 16, 2024	411
Overview	411
14 Tutorials Details	412
Tutorial Message	412
T1 - Countering Hateful and Offensive Speech Online - Open Challenges	413
T2 - Enhancing LLM Capabilities Beyond Scaling Up	415
T3 - Reasoning with Natural Language Explanation	417
T4 - Language Agents: Foundations, Prospects, and Risks	418
T5 - AI for Science in the Era of Large Language Models	419
T6 - Human-Centered Evaluation of Language Technologies	420

15 Workshops	421
Overview	421
W1 - BlackboxNLP 2024: Analyzing and interpreting neural networks for NLP	423
W2 - Seventh Workshop on Computational Models of Reference, Anaphora and Coreference	424
W3 - Seventh Workshop on Fact Extraction and VERification (FEVER)	426
W4 - Workshop on the Future of Event Detection	430
W5 - The Sixth Workshop on Narrative Understanding	431
W6 - Third Workshop on NLP for Positive Impact	432
W7 - The Third Workshop on Text Simplification, Accessibility and Readability	433
W8 - The Eighth Widening NLP Workshop (WiNLP 2024)	435
W9 - The SIGNLL Conference on Computational Natural Language Learning (CoNLL)	436
W10 - Ninth Conference on Machine Translation (WMT24)	439
W11 - Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)	449
W12 - The 4th International Workshop on Natural Language Processing for Digital Humanities (NLP4DH)	451
W13 - GenBench: The second workshop on generalisation (benchmarking) in NLP	455
W14 - Natural Legal Language Processing (NLLP) Workshop 2024	457
W15 - The 4th Workshop on Multilingual Representation Learning	460
W16 - NLP4Science: The First Workshop on Natural Language Processing for Science	461
W17 - The Second Workshop on Social Influence in Conversations (SICon 2024)	462
W18 - The 11th Workshop on Asian Translation (WAT2024)	463
W19 - The First Workshop on Advancing Natural Language Processing for Wikipedia (NLP for Wikipedia)	465
16 Local Guide	466
Conference Venue	466
About Miami	468
Useful Information	471
Local Customs	472
Food Options	473
17 Venue Map	475
Author Index	479
Sponsorship	521

Conference Information

Message from the General Chair

I'm over the moon excited to welcome you to the 2024 edition of the Conference on Empirical Methods in Natural Language Processing! This year marks the 29th edition of EMNLP, at least according to ACL Anthology proceedings. I counted 14 papers in that first edition; how times have changed and how much our community has grown!

The organization's logistics have become much more complex, with a growing number of submissions and attendees. The effort and time from the many volunteers poured into making this meeting happen are tremendous and worthy of recognition. My first acknowledgments go to the entire organization committee. My heartfelt thank you goes to each one of them:

- **Program Chairs:** Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung (Vivian) Chen
- **ARR Guest Program Chair:** Vincent Ng
- **Local Chairs:** Mark Finlayson, and Zoey Liu
- **Industry Track Chairs:** Franck Dernoncourt, Daniel Preoțiuc-Pietro, and Anastasia Shimorina
- **Workshop Chairs:** David Vilar, Xiaodan Zhu, and Marta R. Costa-Jussa
- **Tutorial Chairs:** Jessy Li, and Fei Liu
- **Ethics Chairs:** Luciana Benotti, Snigdha Chaturvedi, and Sunipa Dev
- **Internal Communications Chairs:** Jing Li, Yixin Cao
- **Demonstration Chairs:** Delia Irazu Hernandez Farias, Tom Hope, and Manling Li
- **Publication Chairs:** Milad Alshomary, Danilo Croce, and Gözde Gülahin
- **Handbook Chairs:** Marco Polignano
- **Publicity Chairs:** Shruti Rijhwani, and Elias Stengel-Eskin

-
- **Student Volunteer Chairs:** Shubhra Kanti (Santu) Karmaker, Nafise Sadat Moosavi, and Emily Prud'hommeaux
 - **Diversity/Inclusion Chairs:** Christos Christodoulopoulos, Veronica Perez-Rosas, and Danish Pruthi
 - **Sponsorship Chairs:** Heba Elfardy, and Leonardo Neves
 - **Website Chairs:** Raj Dabre, and Tiago Torrent
 - **Virtual Infrastructure Chairs:** Lianhui Qin, and Vladimir Araujo
 - **Past Program Chair Advisors:** Juan Pino, and Kevin Duh

Thank you to the wonderful Jenn Rachford, our amazing business manager, and the SIGDAT board, Isabelle Augenstein, Kai-Wei Chang, Alice Oh, and Juan Pino, for all their support and flexibility.

I hope this conference brings you inspiration, motivation, energy, and stronger connections; enjoy your time in Miami!

Deseo que ésta conferencia les sea muy productiva!

Thamar Solorio
General Chair

Message from the Program Chairs

Welcome to EMNLP 2024! We are excited to welcome you to one of the most prominent conferences in the field of Natural Language Processing. This year, EMNLP 2024 is being held in a hybrid format, offering both virtual and in-person participation in beautiful Miami. Due to a record-breaking number of submissions, we've expanded the total number of accepted papers to accommodate more cutting-edge research from around the globe.

Submission and Acceptance

EMNLP 2024 received an unprecedented 6,395 committed paper submissions, marking a significant milestone in the conference's history. There are 220 papers desk-rejected due to issues such as anonymity violations, multiple submissions, or formatting discrepancies and 70 papers withdrawn before completing the review process. After these exclusions, 6,105 submissions were fully reviewed. This represents a substantial increase of 1,196 submissions compared to last year, further underscoring the growing interest in the field of NLP. Following a rigorous review process, including feedback from reviewers, area chairs, and senior area chairs, we have carefully balanced the number of accepted papers to maintain an acceptance rate consistent with previous years.

We ultimately accepted 1,271 papers for the main conference and 1,029 papers as Findings of EMNLP. Among these, 168 will be presented orally, while 1,745 will be presented as poster presentations, both in-person and virtually. Additionally, 122 submissions were accepted for the Industry track and 52 for the Demo track. The conference will also feature 13 papers from Computational Linguistics (CL) and 32 from Transactions of the Association for Computational Linguistics (TACL). Detailed statistics on the accepted papers are provided below.

- **Number of Papers Submitted:** 6,105
- **Accepted to the Main Conference:** 1,271
- **Acceptance Rate (Main):** 20.8%
- **Accepted to Findings:** 1,029
- **Acceptance Rate (Main + Findings):** 37.7%

Limitations Statement

Continuing the practice from previous conferences, every submitted paper was required to include an explicitly named Limitations section, discussing the limitations of the work. Importantly, this section does not count toward the page limit. While this rule was in place, we chose not to enforce desk rejections for papers that did not strictly comply during this submission phase.

Tracks

For a smooth submission process, EMNLP 2024 papers were categorized into 26 tracks, closely mirroring the structure of previous EMNLP conferences and reflecting the established divisions within the field. Among these tracks, NLP Applications, Resources and Evaluation, Interpretability and Analysis of Models for NLP, and Multimodality and Language Grounding to Vision, Robotics, and Beyond were the most popular, each receiving over 200 submissions.

ACL Rolling Review

The ACL Rolling Review (ARR) is an initiative of the Association for Computational Linguistics that introduces a two-step process for reviewing and accepting papers: (1) a centralized rolling review and (2) the opportunity for authors to commit their reviewed papers to a specific publication venue. For EMNLP 2024, we continued the practice from previous years where authors first submit their papers to ARR, then commit the reviewed papers by a specified deadline. As part of the process, we served as ARR Editors and worked in collaboration with the ARR Editors-in-Chief for the June cycle, overseeing the review processgathering reviews and meta-reviews, and coordinating with reviewers and meta-reviewers. The committed papers underwent additional review by SACs, who provided recommendations. Final decisions were made based on SAC recommendations and all information collected from the reviewing phase, taking into account not just review scores, but also the quality of reviews, author responses, discussions, meta-reviews, and SAC/AC recommendations.

Best Paper Selection

This year, we included best paper awards to recognize a broader range of exceptional work as previous events:

- **Best Papers and Outstanding Papers** present fascinating, controversial, surprising, impressive, and/or potentially field-changing ideas.
- **Senior Area Chair's awards** are similar to Outstanding papers, but specific to this research track.
- **Best Theme Papers** (= Senior Area Chair's awards for the special theme track) make significant new contributions to efficiency in model algorithms, training, and inference.
- **Social Impact Papers** have the potential for significant positive societal impact.
- **Resource Papers** announce, describe, and share a fascinating, valuable, or potentially field-changing new resource.

Based on nominations from SACs and ACs, 114 candidates have been shortlisted for consideration for the above awards. The final selection is made by the Best Paper Award Committee, and the winners will be announced and will present their work during the closing ceremony.

Presentation Mode

When deciding between oral and poster presentations, our goal was not to base the choice solely on the perceived quality or merit of the papers. Instead, we also considered the authors' preferences for their presentation mode, as well as our assessment of which format would best suit the content of each individual paper for optimal engagement and clarity.

Keynotes and Panel

A major highlight of this years program is the lineup of three outstanding keynote talks:

- Prof. Percy Liang (Stanford University) on Open-Source and Science in the Era of Foundation Models.
- Prof. Anca Dragan (University of California Berkeley and Google Deepmind) on My Journey in AI Safety and Alignment.
- Prof. Tom Griffiths (Princeton University) on Bayes in the Age of Intelligent Machines.

Moreover, we will have an exciting panel to discuss the importance of NLP in the era of LLMs.

Gratitude

We would like to thank the following people for their support and contributions:

- The General Chair, Thamar Solorio;
 - The ARR Editors-in-Chief of the June 2024 cycle (Vincent Ng), Technical Staff (Jonathan K. Kummerfeld), and the entire team (Mausam, Viviane Moreira, Lilja Øvreliid, Anna Rogers, Jun Suzuki, Jing Jiang, Michael White);
 - The OpenReview team, especially Celeste Martinez for multiple rounds of technical help in setting up EMNLP 2024 on the OR platform;
 - The 99 Senior Area Chairs;
 - The 1,458 Area Chairs and the 10,309 reviewers;
 - The awards committee chairs, Luke Zettlemoyer, Ivan Titov, and Claire T. Cardie, and 36 awards committee members;
 - The ethics chairs, Luciana Benotti, Snigdha Chaturvedi, and Sunipa Dev;
 - The industry track chairs, Franck Dernoncourt, Daniel Preoṭiuc-Pietro, Daniel Preoṭiuc-Pietro, and Anastasia Shimorina;
 - The demonstration chairs, Delia Irazu Hernandez Farias, Tom Hope, and Manling Li;
 - The internal communications chairs, Jing Li and Yixin Cao;
 - The website chairs, Raj Dabre and Tiago Torrent;
 - The publication chairs, Milad Alshomary, Danilo Croce, and Gözde Gül ahin;
 - The handbook chair, Marco Polignano
 - The local organization chairs, Mark Finlayson and Zoey Liu , and their team;
 - The publicity chairs, Shruti Rijhwani and Elias Stengel-Eskin;
 - The student volunteer chairs, Shubhra Kanti (Santu) Karmaker, Nafise Sadat Moosavi, and Emily Prud'hommeaux;
 - The diversity/inclusion chairs, Christos Christodoulopoulos, Veronica Perez-Rosas, and Danish Pruthi;
 - The virtual infrastructure chairs, Lianhui Qin and Vladimir Araujo;
 - The ACL Anthology Director Matt Post and his team;
 - The TACL editors-in-chief (Asli Celikyilmaz, Roi Reichart, Dilek Hakkani Tur) and CL Editor in-Chief Wei Lu for coordinating TACL and CL presentations with us;
 - The NAACL 2024 Program Chairs (Kevin Duh, Helena Gomez, and Steven Bethard) and the ACL 2024 Program Chairs (Lun-Wei Ku, André F. T. Martins, Vivek Srikumar);
-

-
- Damira Mrsic and Underline Team;
 - Jennifer Rachford and entire conference support staff;
 - All the authors of papers submitted for review and committed to the conference.

We hope that you will enjoy this years program and hybrid conference!

Yaser Al-Onaizan (Saudi Data and AI Authority, National Center for AI)

Mohit Bansal (University of North Carolina at Chapel Hill)

Yun-Nung (Vivian) Chen (National Taiwan University)

EMNLP 2024 Program Co-Chairs

Message from the Local Chair

Dear EMNLP 2024 Participants,

It is our great pleasure to welcome you to EMNLP 2024, held in the lovely tropical city of Miami, Florida, which is North America's gateway to South America and the Caribbean.

Miami's unique language characteristics are at the heart of it's identity. Spanish is the dominant language here, but not central American Spanish, as is often found elsewhere in the United States. The dominant dialect is Cuban Spanish, with extensive local enclaves of Venezuelans, Colombians, and Argentinians. Indeed, every nationality and cultural group from South American and Caribbean is well represented here. You will find neighborhoods which speak primarily Haitian Creole, as well as Brazilian Portuguese. When you add in the large populations of Europeans from France, Spain, and the Balkans, the local language picture becomes quite rich indeed. Miami even has it's own dialect of English, which was identified by Florida International University socio-linguist Phillip Carter: this dialect overlays standard English with Spanish syntax and Cuban vocabulary and slang, and comes with it's own distinctive accent. So Miami has truly a distinctive language mix!

Given Miami's differences from much of the rest of the United States—in language, population, and climate—locals joke that we are not really in the United States at all, but rather in Latin America. The joke continues with the observation that one of Miami's most convenient attributes, as a supposedly independent Latin American nation, is its proximity to the United States and the fact that we share an open border and a currency. While tongue-in-cheek, if you have traveled elsewhere in the United States you will see a grain of truth in this, and we hope you enjoy and appreciate Miami's unusual character.

While you are here we hope you take full advantage of the cultural richness Miami has to offer, and the many fun things to do. This includes our vibrant, burgeoning, world-class food scene (with many Michelin starred restaurants), our world-renowned nightlife, and our beautiful beaches. Try the widely available Latin American food, such as *pastelitos* or *croquetas*, or our many varieties of speciality coffee, such as *cafecitas* (cuban coffee) or *cortaditos*. Enjoy a night out salsa dancing on *Calle Ocho* (8th Street) in Little Havana, or catch a Latin music or concert at the many concert halls in Downtown, Brickell, or Wynwood. Dance the night away at our dance clubs that feature world-class electronic music DJs, or visit some of our fantastic museums, such as the Perez Art Museum Miami (the PAMM), the Frost Museum of Science (where the social event will be held), or the Viscaya Museum and Gardens. Indulge your dark desire for conspicuous consumption of luxury goods in our high-end stores in Brickell City Center or the Miami Design District.

In the past Miami has been famous for its nightlife and Latin flavors, as a place to visit, relax, and have fun. This is still true, but Miami has grown tremendously as a city in recent years in many other ways. For example, we now boast two major research universities: Florida International University, a public research university, is home to over 50,000 students and has recently been ranked as a top-50 public university in the United states and a top-100 university overall. The University of Miami, also ranked in the top 100, boasts a beautiful campus in Coral Gables, nearly 20,000 students, and a major research hospital. Miami is also home to a fast-growing tech startup and cryptocurrency scene, with a variety of startup accelerators, incubators, funders, and networking organizations, including Endeavor Miami, The Knight Foundation, The Lab, Rokk3r Labs, eMerge Americas, the Miami Angels and Wyncode. Finally, Miami continues to solidify its standing a major hub of international finance for Latin America, with many banks and other financial firms opening major branches here or even moving their headquarters to Miami.

Returning to EMNLP, we would like to extend our thanks to Jennifer Rachford and Megs Haddad, both of the ACL business office, who provided quick, gracious, and ever-informative help in the quite laborious process of issuing many hundreds of visa invitation letters for those coming from abroad. If you were one

of those who needed a visa letter, and see them at the conference, please take a moment to thank them for their hard work.

In closing, we hope that you will thoroughly enjoy your stay in Miami, exploring its rich culture, taking advantage of the many opportunities for fun, all the while getting the most out the extensive technical program of EMNLP.

¡Bienvenido a Miami!

Mark Finlayson
Florida International University, Miami, FL

Zoey Liu
University of Florida, Gainesville, FL

Local Chairs, EMNLP 2024

Organizing Committee

General Chair

Thamar Solorio, Mohamed bin Zayed University of Artificial Intelligence
and University of Houston

Program Chairs

Yaser Al-Onaizan, Saudi Data and AI Authority, National Center for AI
Mohit Bansal, University of North Carolina at Chapel Hill
Yun-Nung (Vivian) Chen, National Taiwan University

Local Chair

Mark Finlayson, Florida International University
Zoey Liu, University of Florida

Industry Track Chairs

Franck Dernoncourt, Adobe Research
Daniel Preotiuc-Pietro, Bloomberg
Anastasia Shimorina, Orange Innovation

Workshop Chairs

David Vilar, Google Inc.
Xiaodan Zhu, Queen's University
Marta R. Costa-Jussa, Meta AI

Tutorial Chairs

Jessy Li, The University of Texas at Austin
Fei Liu, Emory University

Ethics Chairs

Luciana Benotti, Facultad de Matemática, Astronomía, Física y Computación
Snigdha Chaturvedi, University of North Carolina at Chapel Hill
Sunipa Dev, Google Research

Demonstration Chairs

Delia Irazu Hernandez Farias, Instituto Nacional de Astrofísica, Óptica y Electrónica
Tom Hope, AI2
Manling Li, Northwestern University

Publication Chairs

Milad Alshomary, Columbia University
Danilo Croce, University of Rome Tor Vergata
Gözde Gülahin, Koç University

Handbook Chair

Marco Polignano, University of Bari Aldo Moro

Publicity Chairs

Shruti Rijhwani, Google DeepMind
Elias Stengel-Eskin, University of North Carolina

Student Volunteer Chairs

Shubhra Kanti (Santu) Karmaker, University of Central Florida
Nafise Sadat Moosavi, University of Sheffield
Emily Prud'hommeaux, Boston College

Diversity and Inclusion Chairs

Christos Christodoulopoulos, Amazon
Veronica Perez-Rosas, University of Michigan
Danish Pruthi, Indian Institute of Science (IISc), Bangalore

Sponsorship Chairs

Heba Elfardy, Amazon
Leonardo Neves, Snap Inc.

Website Chairs

Raj Dabre, National Institute of Information and Communications Technology (NICT), Japan
Tiago Torrent, Federal University of Juiz de Fora

Virtual Infrastructure Chair

Lianhui Qin, AI2
Vladimir Araujo, KU Leuven

Past Program Chair Advisors

Juan Pino, Meta
Kevin Duh, Johns Hopkins University

Senior Program Committee

Computational Social Science and Cultural Analytics

Chenhao Tan, University of Chicago
David Bamman, School of Information at UC Berkeley
Svitlana Volkova, Aptima Inc.
Tanmoy Chakraborty, Indian Institute of Technology Delhi

Dialogue and Interactive Systems

Larry P. Heck, Tech AI
Luis Fernando D'Haro, Universidad Politécnica de Madrid
Minlie Huang, Tsinghua University
Rebecca Passonneau, Indian Institute of Technology Delhi

Discourse and Pragmatics

Malihe Alikhani, Northeastern University
Vered Shwartz, University of British Columbia

Ethics, Bias, and Fairness

Malvina Nissim, University of Groningen
Monojit Choudhury, Mohamed bin Zayed University of Artificial Intelligence
Natalie Schluter, Apple
Kai-Wei Chang, University of California, Los Angeles

Generation

Mirella Lapata, Edinburgh University
Naoki Okazaki, Tokyo Institute of Technology
Sebastian Gehrmann, Bloomberg LP
Yangfeng Ji, University of Virginia

Human-Centered NLP

David Mimno, Cornell University
Jeff Bigham, Carnegie Mellon University
Marine Carpuat, University of Maryland

Information Extraction

Derry Tanti Wijaya, Boston University
Lifu Huang, University of California
Ndapa Nakashole, University of California
Ruihong Huang, Texas A&M University
Scott Yih, Facebook AI Research

Information Retrieval and Text Mining

Luca Soldaini, Allen Institute for AI

Pawan Goyal, Indian Institute of Technology
Wenhu Chen, University of Waterloo

Interpretability, Interactivity and Analysis of Models for NLP

Grzegorz Chrupala, Tilburg University
Lingpeng Kong, University of Hong Kong
Nadir Durrani, Arabic Language Technologies (ALT)
Tal Linzen, New York University
Ren Chen, University of Southern California
Xuanjing Huang, Fudan University

Language Modeling

Anna Rumshisky, UMass Lowell
Nanyun Peng, University of California, Los Angeles
Swabha Swayamdipta, University of Southern California
Tatsunori Hashimoto, Stanford University

Linguistic Theories, Cognitive Modeling and Psycholinguistics

Frank Keller, University of Edinburgh
Najoung Kim, Boston University

Low-resource Methods for NLP

Kareem Darwish, Qatar Computing Research Institute
Miryam De Lhoneux, KU Leuven in Belgium
Shafiq Joty, Salesforce
Wenpeng Yin, Penn State University
Yue Dong, University of California

Machine Learning for NLP

Gunhee Kim, Seoul National University
Partha Talukdar, Google Research and IISc Bangalore
Shashank Srivastava, UNC Chapel Hill
Taylor Berg-Kirkpatrick, University of California

Machine Translation

Alexander Fraser, Technical University of Munich
Lei Li, Carnegie Mellon University
Paco Guzmán, Meta AI

Multilinguality and Language Diversity

Antonis Anastasopoulos, George Mason Computer Science
Manaal Faruqui, Google
Steven Bird, Charles Darwin University
Zornitsa Kozareva, SliceX AI

Multimodality and Language Grounding to Vision, Robotics and Beyond

Gabriel Stanovsky, Hebrew University of Jerusalem
Hao Tan, Nottingham University Business School China
Jack Hessel, Samaya.ai
Jesse Thomason, University of Southern California
Roma Patel, Brown University
Zhe Gan, Apple

NLP Applications

Avirup Sil, IBM Research AI
Gholamreza Haffari, Monash University
Gokhan Tur, University of Illinois Urbana-Champaign
Joel Tetreault, University of Rochester
Kevin Small, Amazon
Makoto Miwa, Toyota Technological Institute
Parisa Kordjamshidi, Michigan State University
Roman Klinger, University of Bamberg
Sudha Rao, Microsoft Research
Wei Lu, University of Michigan

Phonology, Morphology and Word Segmentation

Brian Roark, Google

Question Answering

Eunsol Choi, New York University
Huan Sun, The Ohio State University
Mrinmaya Sachan, ETH Zürich
Siva Reddy, McGill University

Resources and Evaluation

Adina Williams, Facebook AI Research
Alane Suhr, UC Berkeley
Eduardo Blanco, Univiersity of Arizona
Jimmy Lin, University of Waterloo
Masayuki Asahara, National Institute for Japanese Language and Linguistics
Sujian Li, Peking University
Wei Xu, Georgia Institute of Technology
Yue Zhang, Westlake University
Yufang Hou, IBM Research

Semantics: Lexical, Sentence level, Document Level, Textual Inference, etc.

Marianna Apidianaki, University of Pennsylvania
Sujith Ravi, Temple University

Sentiment Analysis, Stylistic Analysis, and Argument Mining

Saif Mohammad, National Research Council Canada

Veronique Hoste, Ghent University
Zhongyu Wei, Fudan University

Special Theme: Efficiency in Model Algorithms, Training, and Inference

Emma Strubell, Carnegie Mellon University
Nafise Sadat Moosavi, University of Sheffield
Sara Hooker, Cohere For AI

Speech Processing and Spoken Language Understanding

Julia Hirschberg, Columbia University
Preethi Jyothi, IIT Bombay

Summarization

Ramakanth Pasunuru, FAIR at Meta
Xiaojun Wan, Peking University

Syntax, Parsing and their Applications

Lingpeng Kong, University of Hong Kong
Najoung Kim, Boston University

Conference Organizers & Vendors

I would like to extend my sincere gratitude to David Yarowsky for his invaluable assistance in the submission and selection process for our EMNLP 2024 venue. I also appreciate the support of the SIGDAT & ACL Boards for their efforts in reviewing bids and selecting Miami, Florida, as our conference location. Special thanks to Thamar Solorio and Mark Finlayson for their on-site planning and assisting in solidifying the logistics for our conference activities.

Additionally, I would like to acknowledge our partners for EMNLP 2024:

ACL Onsite Team

- Megan Haddad, Event Assistant
- Lina Staggs, Support Staff
- Sally Stevenson, Support Staff

Lee Hartman & Sons

- Jon Dorsey, AV Director
- Trevor Laffoon, AV Tech

Underline

- Damira Mrsic
- Luka Simic
- Borna Bevanda
- Rafael Grabovica
- Lucijan Prpic
- Felipe Salazar
- Petra Vizintin

Vista South

- Anthony Montanaro

Sir Speedy

- Manny Pose

Venues

- **Hyatt Regency Miami:** Steven Encinosa, Maria Corona, Heather Eubank, Cori Cramsey, Chef Robert Allum
- SpringHill Suites Miami Downtown/Medical Center
- Aloft Miami - Brickell
- AC Hotel Miami

Thank you to the entire organizing and program committees for your hard work, dedication, and countless hours of effort—sometimes at the expense of sleep—to make this conference a success.

Sincerely,
Jennifer Rachford
ACL Director of Events/Business Manager

2

Anti-Harassment Policy

EMNLP 2024 adheres to the ACL Anti-Harassment Policy. Any participant who experiences harassment or hostile behavior may contact any current member of the ACL Professional Conduct Committee or Jennifer Rachford, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken. The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of a ACL conference. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for participants at our events and in our programs. Harassment and hostile behavior are unwelcome at any ACL conference. This includes speech or behavior (including in public presentations and on-line discourse) that intimidates, creates discomfort, or interferes with a persons participation or opportunity for participation in the conference. We aim for ACL conferences to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention.

The ACL board members are listed at <https://www.aclweb.org/portal/about>. The full policy and its implementation is defined at https://aclweb.org/adminwiki/index.php/Anti-Harassment_Policy

Code of Ethics Policy

ACL adopts the ACM Code of Ethics (<https://www.acm.org/code-of-ethics>) in the version adopted June 22nd, 2018, by the ACM Council. In its application to ACL, it is to be read in the contextually appropriate interpretation, e.g., ACM member is to be read as ACL member. Sec 4.2 should be read as follows: 4.2 Treat violations of the Code as inconsistent with membership in the ACL. Each ACL member should encourage and support adherence by all members of the CL/NLP community regardless of ACL membership. ACL members who recognize a breach of the Code should consider reporting the violation to the ACL, which may result in remedial action.

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of the ACL. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for all the members, as well as participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference, associated event, or in ACL-affiliated online discussions. This includes speech or behavior that intimidates, creates discomfort, or interferes with a persons participation or opportunity for participation in a conference or an event. We aim for ACL-related activities to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, appearance, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention. The policy is not intended to inhibit challenging scientific debate, but rather to promote it by ensuring that all are welcome to participate in the shared spirit of scientific inquiry. Vexatious complaints and willful misuse of this procedure will render the complainant subject to the same sanctions as a violation of the anti-harassment policy.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, anyone who experiences harassment or hostile behavior may contact any member of the Professional Conduct Committee. Members of this committee are instructed to keep any such contact in strict confidence, and those who approach the committee will be consulted before any actions are taken.



Meal Info

Overview

Breakfast

Nov 11 - Nov 16

Breakfast is not provided, the hotel has a market open 24hrs that has breakfast sandwiches, pastries and snacks for purchase.

Breaks*

Nov 11 - Nov 16

Coffee, tea, pastry, and fruit are provided late morning and midafternoon.

Lunch

Nov 11 - Nov 16

On your own. Lunch is not provided, the hotel has a market open 24hrs that has breakfast sandwiches, pastries and snacks for purchase. See detailed agenda for times.

Dinner

Nov 11 - Nov 16

On your own. Dinner is not provided, but there are plenty of cafes, and restaurants within walking distance.

Social Events **

Nov 11, 18:30 - 21:00 - Welcome Reception

Light canapes, and a drink ticket will be provided on Monday Evening, November 11, 2024, at the Welcome Reception. It will be held on the lower level of the Hyatt Regency Miami. From the lobby take the escalator down to the Terrace Level.

Nov 13, 19:00 - 22:30 - Social Event Gala Dinner

International Buffet Dinner and a drink ticket will be provided on Wednesday Evening, November 13, 2024, at the Social Event Gala Dinner (Social Gala). Held at the Frost Science Museum.

Our primary goal for EMNLP 2024 is to make it an exceptional annual meeting. We want this conference to be remembered not just for the outstanding lineup of speakers and the impressive conference venue but also for the vibrant Social Programs for our Full Conference Attendees. Our aim is to provide an experience that goes beyond the academic sessions, ensuring that every delegate gets a taste of the best we have to offer. Allow us to provide you with a detailed glimpse into the exciting activities and events that we have in store to enhance networking opportunities and foster a strong sense of community among participants.

Please note the following:

* Denotes Included in Main Conference, Tutorial & Workshop Registrations

** Denotes included in the Main Conference Ticket only

Registration and can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk or through your YesEvents registration login. No admission without an entry ticket.

4

Welcome Reception



CONFERENCE BADGE REQUIRED FOR ENTRY

Monday 11th, November, 2024

Lite Appetizers & Drink Ticket (Dinner not provided)

Venue: **Hyatt Regency Miami Riverfront** - 400 South East Second Ave - Miami, FL 33131

Lower Level Terrace (take the lobby escalator down to the Terrace level)

<https://www.hyatt.com/>

Time: **18.30 - 21.00**

Dress code: **Smart Casual**

Kick off EMNLP 2024 with our Welcome Reception, an evening of networking, refreshments, and engaging conversations. This event offers a great opportunity to meet fellow attendees, reconnect with colleagues, and engage with the natural language processing community before the conference begins in full. We look forward to welcoming you to an evening that promises to set the stage for an inspiring and productive conference experience.

Please Note: Attendance to the Welcome Reception is included for Main Conference attendees. If you are not registered for the Main Conference but would like to attend, you may add this event to your registration at the Registration Solution Desk located on the lobby level.

Warm regards, *The EMNLP 2024 Organizing Committee*

5

Social Event Gala Dinner



CONFERENCE BADGE REQUIRED FOR ENTRY

Wednesday 13th, November, 2024

Venue: **Frost Science Museum** - 1101 Biscayne Blvd, Miami, FL 33132

Schedule:

19:00 - Doors open Explore the Exhibits, Planetarium and Aquarium

19:30 - Buffets open

21:00 - Dancing

22:00 - Last Call

22:15 - End

Dress code: **Casual**

The Frost Science Museum is a prominent science museum and planetarium. It features a variety of interactive exhibits focusing on science, technology, engineering, and mathematics (STEM) topics. Here are some highlights:

- **Exhibits:** The museum offers exhibits on diverse topics such as the human body, the physics of flight, marine biology, and outer space exploration.
- **Aquarium:** It includes a 500,000-gallon Gulf Stream Aquarium showcasing the diverse marine life found in Florida's ecosystems.

- **Planetarium:** The Frost Planetarium hosts astronomy shows and immersive experiences about space exploration and the universe.
- **Interactive Learning:** Many exhibits are hands-on, encouraging visitors to engage directly with scientific concepts and phenomena.

The Frost Science Museum aims to inspire curiosity and a passion for science through its interactive exhibits and educational programs. It's a popular destination for both locals and tourists interested in exploring science and technology in an engaging way.

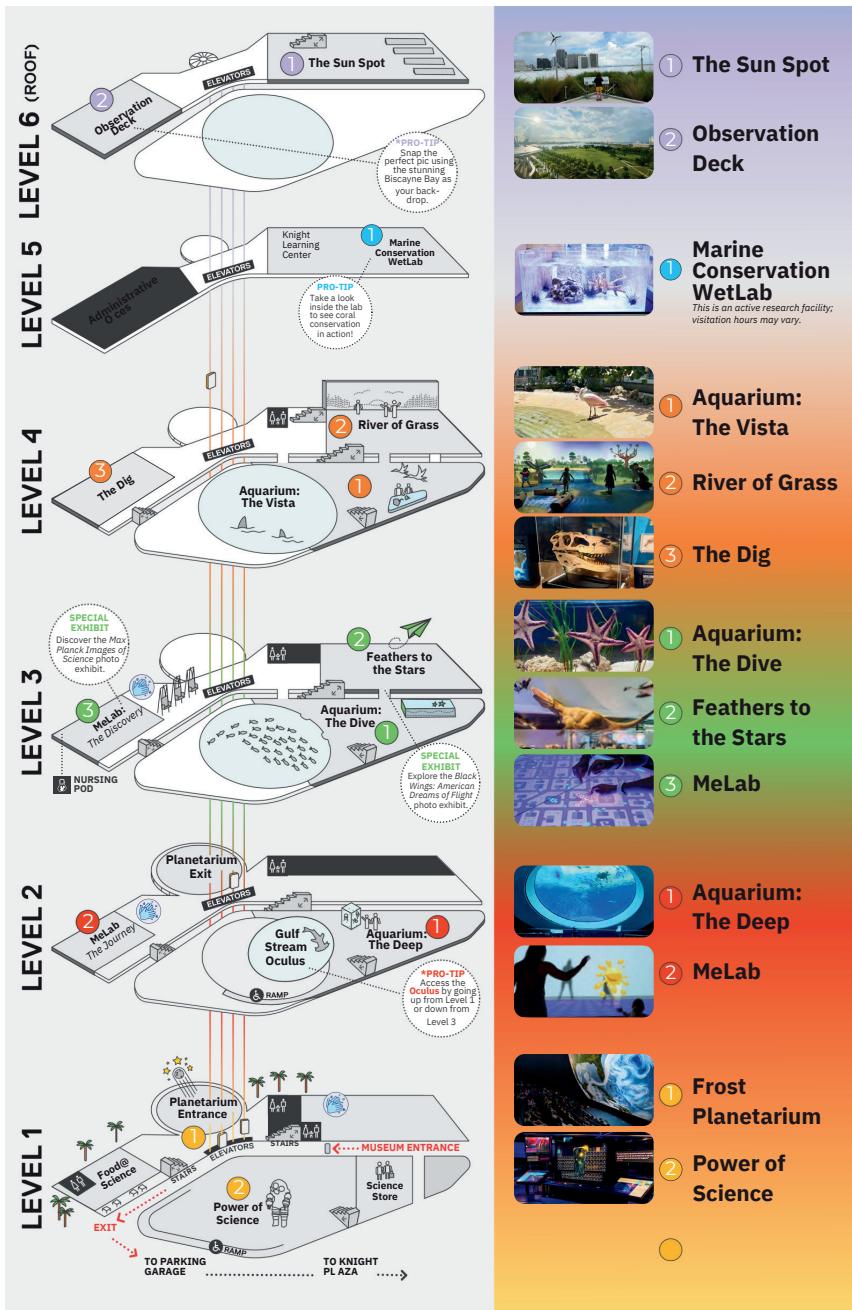
Transportation: Miami Metro Mover James Knight Center (Omni Loop)

Hours: 05:00am - 12:00am

Hop on Station: Knight Center station (100 SE Second Street)

Hop off Station: Museum Park station (1191 Biscayne Blvd.)

FROST SCIENCE MUSEUM MAP



6

Keynotes

Open-Source and Science in the Era of Foundation Models



Percy Liang

Stanford University, Stanford, California

Tuesday, November 12th – Time: from 09:30 to 10:30 – Room: James Knight Center

Abstract: As capabilities of foundation models skyrocket, openness plummets. In this talk, I argue that open-source models are essential for the long-term goal of building a rigorous foundation for AI. Greater access—from API to open-weight to open-source—enables deeper forms of research. API access allows us to push the frontier of agents, and I will present our recent work on simulation and problem-solving agents. Open weights enables reproducible research on safety, interpretability, and more generally, model forensics. Open-source unlocks fundamental innovations in architectures, training procedures, and data curation methods. Of course, the key obstacle for building open-source models is the resources required (data, compute, and research/engineering). I will conclude with some promising directions that leverage the community that bring us closer to the vision of open-source foundation models.

Bio: Percy Liang is an Associate Professor of Computer Science at Stanford University (B.S. from MIT, 2004; Ph.D. from UC Berkeley, 2011) and the director of the Center for Research on Foundation Models (CRFM). He is currently focused on making foundation models (in particular, language models) more accessible through open-source and understandable through rigorous benchmarking. In the past, he has worked on many topics centered on machine learning and natural language processing, including robustness, interpretability, human interaction, learning theory, grounding, semantics, and reasoning. He is also a strong proponent of reproducibility through the creation of CodaLab Worksheets. His awards include the Presidential Early Career Award for Scientists and Engineers (2019), IJCAI Computers and Thought Award (2016), an NSF CAREER Award (2016), a Sloan Research Fellowship (2015), a Microsoft Research Faculty Fellowship (2014), and paper awards at ACL, EMNLP, ICML, COLT, ISMIR, CHI, UIST, and RSS.

My Journey in AI Safety and Alignment



Anca Dragan

University of California, Berkeley, California

Wednesday, November 13th – Time: from 09:00 to 10:00 – Room: James Knight Center

Abstract: For nearly a decade now, the problem that has been top of mind for me is how we might enable AI systems to robustly optimize for what people want, and to avoid causing harm from robots and self-driving cars, to assistive devices and deep brain stimulation, to theory and toy models, to large language models and now Gemini. In this talk, I'll take the opportunity to share a bit about my journey in this space, what lessons I've learned, and how we're approaching the safety and alignment of frontier models at Google DeepMind.

Bio: Anca Dragan is an Associate Professor in the EECS Department at UC Berkeley, currently on leave to head AI Safety and Alignment at Google DeepMind. The goal of her research at UC Berkeley has been to enable AI agents (from robots to cars to LLMs to recommender systems) to work with, around, and in support of people. Anca runs the InterACT Lab, where they focus on algorithms for human-AI and human-robot interaction. One of the core problems the Lab has worked on since its inception is AI alignment: getting AI agents to do what people actually want – this has meant learning reward functions interactively, from diverse human feedback forms, across different modalities, while maintaining uncertainty. They have also contributed to algorithms for human-AI collaboration and coordination, like agents fluently working together with human-driven avatars in games, assistance and adaption in brain-machine interfaces, and autonomous cars sharing the road with human drivers.

Bayes in the age of intelligent machines



Tom Griffiths

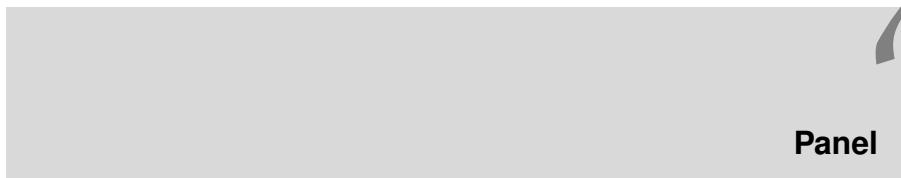
Princeton University, Princeton, New Jersey

Thursday, November 14th – Time: from 09:00 to 10:00 – Room: James Knight Center

Abstract: Recent rapid progress in the creation of artificial intelligence (AI) systems has been driven in large part by innovations in architectures and algorithms for developing large scale artificial neural networks. As a consequence, it's natural to ask what role abstract principles of intelligence such as Bayes rule might play in developing intelligent machines. In this talk, I will argue that there is a new way in which Bayes can be used in the context of AI, more akin to how it is used in cognitive science: providing an abstract description of how agents should solve certain problems and hence a tool for understanding their behavior. This new role is motivated in large part by the fact that we have succeeded in creating intelligent systems that we do not fully understand, making the problem for the machine learning researcher more closely parallel that of the cognitive scientist. I will talk about how this perspective can help us think about making machines with better informed priors about the world and give us insight into their behavior by directly creating cognitive models of neural networks.

Bio: Tom Griffiths is the Henry R. Luce Professor of Information Technology, Consciousness and Culture in the Departments of Psychology and Computer Science at Princeton University, where he is also the Director of the Princeton Laboratory for Artificial Intelligence. His research explores connections between human and machine learning, using ideas from statistics and artificial intelligence to understand how people solve the challenging computational problems they encounter in everyday life. Tom completed his PhD in Psychology at Stanford University in 2005, and taught at Brown University and the University of California, Berkeley before moving to Princeton. He has received awards for his research from organizations ranging from the American Psychological Association to the National Academy of Sciences and is a co-author of the book *Algorithms to Live By*, introducing ideas from computer science and cognitive science to a general audience.

7



Increasing significance of NLP in the age of Large Language Models (LLMs)

- *Prof. Heng Ji, University of Illinois Urbana-Champaign*
- *Prof. Rada Mihalcea, University of Michigan*
- *Prof. Alice Oh, Korea Advanced Institute of Science and Technology*
- *Prof. Sasha Rush, Cornell University; HuggingFace*

Wednesday, November 13, 2024

Time: 14:30 - 15:30 followed by Q&A during the break
Room: James Knight Center

This diverse group of panelists will provide a comprehensive view of the latest trends and challenges in NLP and the interactions with the LLM era. Panelist details can be found on the following Platforms: Underline, Whova, EMNLP 24 Website

Panelist Short Bio

Prof. Heng Ji, University of Illinois Urbana-Champaign

Heng Ji is a professor at Computer Science Department, and an affiliated faculty member at Electrical and Computer Engineering Department and Coordinated Science Laboratory of University of Illinois Urbana-Champaign. She is an Amazon Scholar. She is the Founding Director of Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE). She received her B.A. and M. A. in Computational Linguistics from Tsinghua University, and her M.S. and Ph.D. in Computer Science from New York University. Her research interests focus on Natural Language Processing, especially on Multimedia Multilingual Information Extraction, Knowledge-enhanced Large Language Models and Vision-Language Models. She was selected as a "Young Scientist" by the World Laureates Association in 2023 and 2024. She was selected as "Young Scientist" and a member of the Global Future Council on the Future of Computing by the World Economic Forum in 2016 and 2017. She was named as part of Women Leaders of Conversational AI (Class of 2023) by Project Voice. The other awards she received include two Outstanding Paper Awards at NAACL2024, "AI's 10 to Watch" Award by IEEE Intelligent Systems in 2013, NSF CAREER award in 2009, PACLIC2012 Best paper runner-up, "Best of ICDM2013" paper award, "Best of SDM2013" paper award, ACL2018 Best Demo paper nomination, ACL2020 Best Demo Paper Award, NAACL2021 Best Demo Paper Award, Google Research Award in 2009 and 2014, IBM Watson Faculty Award in 2012 and 2014 and Bosch Research Award in 2014-2018. She was invited to testify to the U.S. House Cybersecurity, Data Analytics, & IT Committee as an AI expert in 2023. She was selected to participate in DARPA AI Forward in 2023. She was invited by the Secretary of the U.S. Air Force and AFRL to join Air Force Data Analytics Expert Panel to inform the Air Force Strategy 2030, and invited to speak at the Federal Information Integrity R&D Interagency Working Group (IIRD IWG) briefing in 2023. She is the lead of many multi-institution projects and tasks, including the U.S. ARL projects on information fusion and knowledge networks construction, DARPA ECOLE MIRACLE team, DARPA KAIROS RESIN team and DARPA DEFT Tinker Bell team. She has coordinated the NIST TAC Knowledge Base Population task 2010-2020. She was the associate editor for IEEE/ACM Transaction on Audio, Speech, and Language Processing, and served as the Program Committee Co-Chair of many conferences including NAACL-HLT2018 and AACL-IJCNLP2022. She was elected as the North American Chapter of the Association for Computational Linguistics (NAACL) secretary 2020-2023. Her research has been widely supported by the U.S. government agencies (DARPA, NSF, DoE, ARL, IARPA, AFRL, DHS) and industry (Amazon, Google, Bosch, IBM, Disney).

Prof. Rada Mihalcea, University of Michigan

Rada Mihalcea is a Professor with the Computer Science and Engineering Department, the University of Michigan. Her research interests include computational linguistics, multimodal behavior analysis, and computational social sciences. She received the Ph.D. degree in computer science and engineering from Southern Methodist University, Dallas, TX, USA, in 2001, and the Ph.D. degree in linguistics from Oxford University, Oxford, U.K., in 2010. She was the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009). In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.

Prof. Alice Oh, Korea Advanced Institute of Science and Technology

Hello, I am a professor at KAIST in the School of Computing with joint appointment in the Graduate School of AI. My research interests are in developing and applying machine learning models for natural language processing. Please read through the pages for my research group (<http://uilab.kr/>) for the latest updates.

Prof. Sasha Rush, Cornell University; HuggingFace

My research (<https://rush-nlp.com/papers/>) aims to build and improve generative AI. We are interested primarily in tasks that involve text generation, historically translation, summarization, and data-to-text generation. Methodologically, we study data-driven probabilistic methods that combine deep-learning based models with probabilistic controls. I am also interested in open-source (<https://rush-nlp.com/pro>

jects/) NLP and deep learning, and develop projects to make deep learning systems safer, more clear, and easier to use. I work part-time at Hugging Face (<http://huggingface.co/>) and like to release various software projects to support NLP and DL research.

8

Birds-of-a-Feather and Affinity Group Meetup

All BoF & Affinity Session will run Parallel with the Oral/Poster Sessions

Tuesday, Nov 12

- 11:00 - 12:30 **LLMs for Embodied Agents** Organizer: Manling Li
Room: Foster (Convention Center 2nd level)
- 11:00 - 12:30 **Law Law Land: Legal Language Meets Large Language Models** Organizer: Santosh Tokala - Room: Johnson (Convention Center 2nd level)
- 12:30 - 14:00 **Large Multimodal Models for Biomedical Research** Organizer: Tianyu Liu
Room: Miami Lecture Hall (Convention Center Level 2)
- 14:00 - 15:30 **LLM Agents for Acoustics and Continuous Signals** Organizer: Huck Yang
Room: Foster (Convention Center Level 2)
- 16:00 - 17:30 **Queer in AI** More details at: <https://www.queerinai.com/>
- 16:00 - 17:30 **Southeast Asian NLP** Organizer: Genta Winata
Room: Johnson (Convention Center Level 2)
- 16:00 - 17:30 **NLP for Structured Data** Organizer: Vivek Gupta
Room: Miami Lecture Hall (Convention Center Level 2)

Wednesday, Nov 13

- 10:30 - 12:00 **Fostering Native and Cultural Inclusivity in LLMs** Organizer: Firoj Alam
Room: Foster
- 10:30 - 12:00 **NLP Tools for Community-Owned Religious Texts in Low-Resourced Languages** Organizer: Inam Ullah - Room: Johnson

Thursday, Nov 14

- 10:30 - 12:00 **2112: The AI Odyssey** Organizer: Prasson Bajpai
Room: Foster
- 10:30 - 12:00 **Embeddings, Reranker, Small LM for Better Search** Organizer: Han Xiao
Room: Miami Lecture Hall

More details can be found on the following Platforms: Underline, Whova, EMNLP 24 Website

Main Conference Overview

Monday, November 11 - Registration & Welcome Reception

- **14:00 - 20:30** Registration - Hyatt Regency Miami Lobby (Riverfront Foyer)
- **18:30 - 21:00** Welcome Reception (Terrace Level - Take lobby escalator down to Terrace Level)

Tuesday, November 12 - Main Conference

- **7:30 - 16:30** Registration - Hyatt Regency Miami Lobby (Riverfront Foyer)
- **09:00 - 09:30** Session 1: **Opening Session**
James Knight Center - 2nd level Convention Center
- **09:30 - 10:30** Session 1: **Invited Speaker - Percy Liang**
James Knight Center - 2nd level Convention Center
Title: Open-Source and Science in the Era of Foundation Models
- **10:30 - 11:00** Break
Riverfront Hall (Lobby Level of James Knight Convention Center)
- **11:00 - 12:30** Session 2: Orals/Posters/Demos Session A
Poster Presentation Tracks & Demos:
Riverfront Hall Lobby Level
 - Generation
 - NLP Applications
 - Information Retrieval

- Linguistic Theories
- All Demos located in Riverfront Hall

Jasmine Lower Terrace Level

- Multimodality
- Industry Track

Oral Presentations:

- Language Modeling 1 (Ashe Auditorium 2nd Floor Convention Level)
- Interpretability and Analysis of Models for NLP 1 (Brickell Lower Terrace Level)
- Low-resource Methods for NLP (Flagler Lower Terrace Level)
- Human-centered NLP (Monroe Lower Terrace Level)
- Machine Translation (Tuttle Lower Terrace Level)

Birds-of-a-Feather and Affinity Group Meetup

- LLMs for Embodied Agents (Foster - Convention Center 2nd level) - Organizer: Manling Li
- Law Law Land: Legal Language Meets Large Language Models (Johnson - Convention Center 2nd level) - Organizer: Santosh Tokala

- **12:30 - 14:00** Lunch Break

Birds-of-a-Feather and Affinity Group Meetup

- Large Multimodal Models for Biomedical Research (Miami Lecture Hall - Convention Center Level 2) - Organizer: Tianyu Liu

- **14:00 - 15:30** Session 3: Orals/Posters/Demos Session B

Poster Presentation Tracks & Demos:

Riverfront Hall Lobby Level

- Language Modeling
- Ethics
- Multilinguality
- Discourse + Phonology + Syntax
- All Demos located in Riverfront Hall

Jasmine Lower Terrace Level

- Interpretability
- Machine Learning for NLP

Oral Presentations:

- Generation and Summarization (Ashe Auditorium 2nd Floor Convention Level)
- Dialogue and Interactive Systems (Brickell Lower Terrace Level)
- Computational Social Science and Cultural Analytics 1 (Flagler Lower Terrace Level)

- Special Theme: Efficiency in Model Algorithms, Training, and Inference (Monroe Lower Terrace Level)
- Resources and Evaluation 1 (Tuttle Lower Terrace Level)

Birds-of-a-Feather and Affinity Group Meetup

- LLM Agents for Acoustics and Continuous Signals (Foster - Convention Center Level 2) - Organizer: Huck Yang
- **15:30 - 16:00 Break**
Riverfront Hall (Lobby Level of James Knight Convention Center)
- **16:00 - 17:30 Session 4: Orals/Posters/Demos Session C**
Poster Presentation Tracks & Demos:
Riverfront Hall Lobby Level

- Resources and Evaluation
- Question Answering
- Computational Social Science
- Machine Translation
- All Demos located in Riverfront Hall

Jasmine Lower Terrace Level

- Special Theme: Efficiency
- Sentiment Analysis
- Summarization

Oral Presentations:

- Ethics, Bias, and Fairness 1 (Ashe Auditorium 2nd Floor Convention Level)
- Information Retrieval and Text Mining (Brickell Lower Terrace Level)
- Multimodality and Language Grounding to Vision, Robotics and Beyond 1 (Flagler Lower Terrace Level)
- Linguistic Theories, Cognitive Modeling and Psycholinguistics (Monroe Lower Terrace Level)
- Industry Track 1 (Tuttle Lower Terrace Level)

Birds-of-a-Feather and Affinity Group Meetup

- Southeast Asian NLP (Johnson - Convention Center Level 2) - Organizer: Genta Winata
 - NLP for Structured Data (Miami Lecture Hall - Convention Center Level 2) - Organizer: Vivek Gupta
 - **17:45 - 18:45 Virtual Poster Session 1**
-

Wednesday, November 13 - Main Conference

- **07:45 - 08:45** Virtual Poster Session 2
- **08:30 - 16:30** Registration - Hyatt Regency Miami (Lobby Riverfront Foyer)
- **09:00 - 10:00** Session 5: **Invited Speaker - Anca Dragan**
James Knight Center
Title: My Journey in AI Safety and Alignment
- **10:00 - 10:30** Break
Riverfront Hall (Lobby Level of James Knight Convention Center)
- **10:30 - 12:00** Session 6: Orals/Posters/Demos Session D
Poster Presentation Tracks & Demos:
Riverfront Hall Lobby Level

- Human-centered NLP
- Resources and Evaluation
- Speech Processing
- NLP Applications

Jasmine Lower Terrace Level

- Low-resource
- Interpretability

Oral Presentations:

- Multimodality and Language Grounding to Vision, Robotics and Beyond (Ashe Auditorium 2nd Floor Convention Level)
- Ethics, Bias, and Fairness (Brickell Lower Terrace Level)
- Discourse, Phonology, and Syntax (Flagler Lower Terrace Level)
- Question Answering (Monroe Lower Terrace Level)
- Industry Track 2 (Tuttle Lower Terrace Level)

Birds-of-a-Feather and Affinity Group Meetup

- Fostering Native and Cultural Inclusivity in LLMs (Foster - Convention Center Level 2) - Organizer: Firoj Alam
- NLP Tools for Community-Owned Religious Texts in Low-Resourced Languages (Johnson - Convention Center Level 2) - Organizer: Inam Ullah
- **12:00 - 13:30** Lunch Break
- **13:30 - 14:15** Session 7: Business Meeting - James knight Center (All Attendees Welcome)
- **14:30 - 15:30** Session 8: Panel (see online schedule for details)
- **15:30 - 16:00** Break
Riverfront Hall (Lobby Level of James Knight Convention Center)

- **16:00 - 17:30 Session 9: Orals/Posters/Demos Session E**

Poster Presentation Tracks & Demos:

Riverfront Hall Lobby Level

- Dialogue
- Multimodality
- Semantics
- Information Retrieval
- Industry Track

Jasmine Lower Terrace Level

- Language Modeling
- Question Answering
- TACL + CL

Oral Presentations:

- Resources and Evaluation 2 (Ashe Auditorium 2nd Floor Convention Level)
- Interpretability and Analysis of Models for NLP 2 (Brickell Lower Terrace Level)
- NLP Applications (Flagler Lower Terrace Level)
- Information Extraction (Monroe Lower Terrace Level)
- Machine Learning for NLP 1 (Tuttle Lower Terrace Level)

- **19:00 - 22:15 Social Event Gala Dinner - Frost Science Museum**

Thursday, November 14 - Main Conference

- **08:30 - 17:00 Registration - Hyatt Regency Miami Lobby (Riverfront Foyer)**

- **09:00 - 10:00 Session 10: Invited Speaker - Tom Griffiths**

James Knight Center (2nd Floor Convention Level)

Title: "Bayes in the Age of Intelligent Machines"

- **10:00 - 10:30 Break**

Riverfront Hall (Lobby Level of James Knight Convention Center)

- **10:30 - 12:00 Session 11: Orals/Posters/Demos Session F**

Poster Presentation Tracks & Demos:

Riverfront Hall Lobby Level

- Generation
- Machine Learning for NLP
- Special Theme: Efficiency
- Resources and Evaluation

Jasmine Lower Terrace Level

- Interpretability
- Ethics

Oral Presentations:

- NLP Applications 2 (Ashe Auditorium 2nd Floor Convention Level)
- Computational Social Science and Cultural Analytics 2 (Brickell Lower Terrace Level)
- Sentiment and Semantics (Flagler Lower Terrace Level)
- Language Modeling 2 (Monroe Lower Terrace Level)
- Multilinguality and Language Diversity (Tuttle Lower Terrace Level)

Birds-of-a-Feather and Affinity Group Meetup

- 2112: The AI Odyssey (Foster - Convention Center Level 2) - Organizer: Prasson Bajpai
- Embeddings, Reranker, Small LM for Better Search (Miami Lecture Hall - Convention Center Level 2) - Organizer: Han Xiao
- **12:00 - 13:00** Lunch Break
- **13:00 - 14:00** Virtual Poster Session 3
- **14:00 - 15:30** Session 12: Orals/Posters/Demos Session G

Poster Presentation Tracks & Demos:

Riverfront Hall Lobby Level

- Dialogue
- NLP Applications
- Information Extraction
- Industry Track

Jasmine Lower Terrace Level

- Computational Social Science
- Multimodality

Oral Presentations:

- Interpretability and Analysis of Models for NLP 3 (Ashe Auditorium 2nd Floor Convention Level)
- Speech Processing and Spoken Language Understanding (Brickell Lower Terrace Level)
- Resources and Evaluation 3 (Flagler Lower Terrace Level)
- Generation (Monroe Lower Terrace Level)
- Machine Learning for NLP 2 (Tuttle Lower Terrace Level)

- **15:30 - 16:00** Break

Riverfront Hall (Lobby Level of James Knight Convention Center)

- **16:00 - 17:00** Session 13: Best Paper Awards

James Knight Center (2nd Floor Convention Level)

- **17:00 - 17:30** Session 13: Closing Session *James Knight Center (2nd Floor Convention Level)*

Friday, November 15 - Workshop/Tutorial Day

- **08:00 - 16:00** Registration - Hyatt Regency Miami Lobby (Riverfront Foyer)
- SEE WORKSHOP & TUTORIAL FOR DETAILS SCHEDULE

Saturday, November 16 - Workshop/Tutorial Day

- **08:00 - 16:00** Registration - Hyatt Regency Miami Lobby (Riverfront Foyer)
- SEE WORKSHOP & TUTORIAL FOR DETAILS SCHEDULE

10

Oral Presentations

Session 02 - Nov 12 (Tue) 11:00-12:30

Language Modeling 1

Nov 12 (Tue) 11:00-12:30 - Room: Ashe Auditorium

11:00 - 11:15 - Ashe Auditorium

Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, Jonathan Herzig

When large language models are aligned via supervised fine-tuning, they may encounter new factual information that was not acquired through pre-training. It is often conjectured that this can teach the model the behavior of hallucinating factually incorrect responses, as the model is trained to generate facts that are not grounded in its pre-existing knowledge. In this work, we study the impact of such exposure to new knowledge on the capability of the fine-tuned model to utilize its pre-existing knowledge. To this end, we design a controlled setup, focused on closed-book QA, where we vary the proportion of the fine-tuning examples that introduce new knowledge. We demonstrate that large language models struggle to acquire new factual knowledge through fine-tuning, as fine-tuning examples that introduce new knowledge are learned significantly slower than those consistent with the model's knowledge. However, we also find that as the examples with new knowledge are eventually learned, they linearly increase the model's tendency to hallucinate. Taken together, our results highlight the risk in introducing new factual knowledge through fine-tuning, and support the view that large language models mostly acquire factual knowledge through pre-training, whereas fine-tuning teaches them to use it more efficiently.

11:15 - 11:30 - Ashe Auditorium

Extracting Prompts by Inverting LLM Outputs

Collin Zhang, John Xavier Morris, Vitaly Shmatikov

We consider the problem of language model inversion: given outputs of a language model, we seek to extract the prompt that generated these outputs. We develop a new black-box method, output2prompt, that extracts prompts without access to the models logits and without adversarial or jailbreaking queries. Unlike previous methods, output2prompt only needs outputs of normal user queries. To improve memory efficiency, output2prompt employs a new sparse encoding technique. We measure the efficacy of output2prompt on a variety of user and system prompts and demonstrate zero-shot transferability across different LLMs.

11:30 - 11:45 - Ashe Auditorium

LLM See, LLM Do: Leveraging Active Inheritance to Target Non-Differentiable Objectives

Luisa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, Sara Hooker

The widespread adoption of synthetic data raises new questions about how models generating the data can influence other large language models (LLMs). To start, our work exhaustively characterizes the impact of passive inheritance of model properties by systematically studying how the source of synthetic data shapes models' internal biases, calibration and preferences, and their generations' textual attributes, providing one of the most comprehensive studies-to-date. We find that models are surprisingly sensitive towards certain attributes even when the synthetic data prompts appear "neutral" which invites the question: can we explicitly steer the distilled data towards desired properties? We demonstrate how such active inheritance can steer the generation profiles of models towards desirable non-differentiable attributes in both directions, e.g. increasing lexical diversity or reducing toxicity. Overall, our study broadens the understanding of the implicit biases inherited by LLMs and explores how we can leverage them to positive effect.

11:45 - 12:00 - Ashe Auditorium

Prompt Optimization in Multi-Step Tasks (PROMST): Integrating Human Feedback and Heuristic-based Sampling

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, Chuichi Fan

Prompt optimization aims to find the best prompt to a large language model (LLM) for a given task. LLMs have been successfully used to help find and improve prompt candidates for single-step tasks. However, realistic tasks for agents are multi-step and introduce new challenges: (1) Prompt content is likely to be more extensive and complex, making it more difficult for LLMs to analyze errors, (2) the impact of an individual step is difficult to evaluate, and (3) different people may have varied preferences about task execution. While humans struggle to optimize

prompts, they are good at providing feedback about LLM outputs; we therefore introduce a new LLM-driven discrete prompt optimization framework PROMST that incorporates human-designed feedback rules to automatically offer direct suggestions for improvement. We also use an extra learned heuristic model that predicts prompt performance to efficiently sample from prompt candidates. This approach significantly outperforms both human-engineered prompts and several other prompt optimization methods across 11 representative multi-step tasks (an average 10.6%-29.3% improvement to current best methods on five LLMs respectively). We believe our work can serve as a benchmark for automatic prompt optimization for LLM-driven multi-step tasks.

12:00 - 12:15 - Ashe Auditorium

Retrieval-Pretrained Transformer: Long-range Language Modeling with Self-retrieval

Ohad Rubin, Jonathan Berant

Retrieval-augmented language models (LMs) have received much attention recently. However, typically the retriever is not trained jointly as a native component of the LM, but added post-hoc to an already-pretrained LM, which limits the ability of the LM and the retriever to adapt to one another. In this work, we propose the emphRetrieval-Pretrained Transformer (RPT), an architecture and training procedure for jointly training a retrieval-augmented LM from scratch and apply it to the task of modeling long texts. Given a recently generated text chunk in a long document, the LM computes query representations, which are then used to retrieve earlier chunks in the document, located potentially tens of thousands of tokens before. Information from retrieved chunks is fused into the LM representations to predict the next target chunk. We train the retriever component with a semantic objective, where the goal is to retrieve chunks that increase the probability of the next chunk, according to a reference LM. We evaluate RPT on four long-range language modeling tasks, spanning books, code, and mathematical writing, and demonstrate that RPT improves retrieval quality and subsequently perplexity across the board compared to strong baselines.

12:15 - 12:30 - Ashe Auditorium

When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs

Ryo Kamoi, Yufan Zhang, Nan Zhang, Jiawei Han, Rui Zhang

Self-correction is an approach to improving responses from large language models (LLMs) by refining the responses using LLMs during inference. Prior work has proposed various self-correction frameworks using different sources of feedback, including self-evaluation and external feedback. However, there is still no consensus on the question of when LLMs can correct their own mistakes, as recent studies also report negative results. In this work, we critically survey broad papers and discuss the conditions required for successful self-correction. We first find that prior studies often do not define their research questions in detail and involve impractical frameworks or unfair evaluations that over-evaluate self-correction. To tackle these issues, we categorize research questions in self-correction research and provide a checklist for designing appropriate experiments. Our critical survey based on the newly categorized research questions shows that (1) no prior work demonstrates successful self-correction with feedback from prompted LLMs, except for studies in tasks that are exceptionally suited for self-correction, (2) self-correction works well in tasks that can use reliable external feedback, and (3) large-scale fine-tuning enables self-correction.

Interpretability and Analysis of Models for NLP 1

Nov 12 (Tue) 11:00-12:30 - Room: Brickell

11:00 - 11:15 - Brickell

Adapters Mixup: Mixing Parameter-Efficient Adapters to Enhance the Adversarial Robustness of Fine-tuned Pre-trained Text Classifiers

Tuc Van Nguyen, Thai Le

Existing works show that augmenting the training data of pre-trained language models (PLMs) for classification tasks fine-tuned via parameter-efficient fine-tuning methods (PEFT) using both clean and adversarial examples can enhance their robustness under adversarial attacks. However, this adversarial training paradigm often leads to performance degradation on clean inputs and requires frequent re-training on the entire data to account for new, unknown attacks. To overcome these challenges while still harnessing the benefits of adversarial training and the efficiency of PEFT, this work proposes a novel approach, called AdpMixup, that combines two paradigms: (1) fine-tuning through adapters and (2) adversarial augmentation via mixup to dynamically leverage existing knowledge from a set of pre-known attacks for robust inference. Intuitively, AdpMixup fine-tunes PLMs with multiple adapters with both clean and pre-known adversarial examples and intelligently mixes them up in different ratios during prediction. Our experiments show AdpMixup achieves the best trade-off between training efficiency and robustness under both pre-known and unknown attacks, compared to existing baselines on five downstream tasks across six varied black-box attacks and 2 PLMs. The code is available at https://github.com/nguyentuc/adapters_mixup.

11:15 - 11:30 - Brickell

An Unsupervised Approach to Achieve Supervised-Level Explainability in Healthcare Records

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, Tuukka Ruotsalo

Electronic healthcare records are vital for patient safety as they document conditions, plans, and procedures in both free text and medical codes. Language models have significantly enhanced the processing of such records, streamlining workflows and reducing manual data entry, thereby saving healthcare providers significant resources. However, the black-box nature of these models often leaves healthcare professionals hesitant to trust them. State-of-the-art explainability methods increase model transparency but rely on human-annotated evidence spans, which are costly. In this study, we propose an approach to produce plausible and faithful explanations without needing such annotations. We demonstrate on the automated medical coding task that adversarial robustness training improves explanation plausibility and introduce AttInGrad, a new explanation method superior to previous ones. By combining both contributions in a fully unsupervised setup, we produce explanations of comparable quality, or better, to that of a supervised approach. We release our code and model weights.

11:30 - 11:45 - Brickell

Evaluating n -Gram Novelty of Language Models Using Rusty-DAWG

William Merrill, Noah A. Smith, Yanai Elazar

How novel are texts generated by language models (LMs) relative to their training corpora? In this work, we investigate the extent to which modern LMs generate n -grams from their training data, evaluating both (i) the probability LMs assign to complete training n -grams and (ii) n -novelty, the proportion of n -grams generated by an LM that did not appear in the training data (for arbitrarily large n). To enable arbitrary-length n -gram search over a corpus in constant time w.r.t. corpus size, we develop Rusty-DAWG, a novel search tool inspired by indexing of genomic data. We compare the novelty of LM-generated text to human-written text and explore factors that affect generation novelty, focusing on the Pythia models. We find that, for $n > 4$, LM-generated text is less novel than human-written text, though it is more novel for smaller n . Larger LMs and more constrained decoding strategies both decrease novelty. Finally, we show that LMs complete n -grams with

lower loss if they are more frequent in the training data. Overall, our results reveal factors influencing the novelty of LM-generated text, and we release Rusty-DAWG to facilitate further pretraining data research.

Interpretability and Analysis of Models for NLP 2

Nov 12 (Tue) 11:00-12:30 - Room: Brickell

11:45 - 12:00 - Brickell

DA³: A Distribution-Aware Adversarial Attack against Language Models

Yibo Wang, Xiangjue Dong, James Caverlee, Philip S. Yu

Language models can be manipulated by adversarial attacks, which introduce subtle perturbations to input data. While recent attack methods can achieve a relatively high attack success rate (ASR), we've observed that the generated adversarial examples have a different data distribution compared with the original examples. Specifically, these adversarial examples exhibit reduced confidence levels and greater divergence from the training data distribution. Consequently, they are easy to detect using straightforward detection methods, diminishing the efficacy of such attacks. To address this issue, we propose a Distribution-Aware Adversarial Attack (DA³) method. DA³ considers the distribution shifts of adversarial examples to improve attacks' effectiveness under detection methods. We further design a novel evaluation metric, the Non-detectable Attack Success Rate (NASR), which integrates both ASR and detectability for the attack task. We conduct experiments on four widely used datasets to validate the attack effectiveness and transferability of adversarial examples generated by DA³ against both the white-box BERT-base and RoBERTa-base models and the black-box LLaMA2-7b model.

Interpretability and Analysis of Models for NLP 1

Nov 12 (Tue) 11:00-12:30 - Room: Brickell

12:00 - 12:15 - Brickell

Ranking Manipulation for Conversational Search Engines

Samuel Pfrommer, Yatong Bai, Tamay Gautam, Somayeh Sojoudi

Major search engine providers are rapidly incorporating Large Language Model (LLM)-generated content in response to user queries. These *conversational search engines* operate by loading retrieved website text into the LLM context for summarization and interpretation. Recent research demonstrates that LLMs are highly vulnerable to jailbreaking and prompt injection attacks, which disrupt the safety and quality goals of LLMs using adversarial strings. This work investigates the impact of prompt injections on the ranking order of sources referenced by conversational search engines. To this end, we introduce a focused dataset of real-world consumer product websites and formalize conversational search ranking as an adversarial problem. Experimentally, we analyze conversational search rankings in the absence of adversarial injections and show that different LLMs vary significantly in prioritizing product name, document content, and context position. We then present a tree-of-attacks-based jailbreaking technique which reliably promotes low-ranked products. Importantly, these attacks transfer effectively to state-of-the-art conversational search engines such as *perplexity.ai*. Given the strong financial incentive for website owners to boost their search ranking, we argue that our problem formulation is of critical importance for future robustness work.

12:15 - 12:30 - Brickell

The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, Yulan He

Understanding in-context learning (ICL) capability that enables large language models (LLMs) to excel in proficiency through demonstration examples is of utmost importance. This importance stems not only from the better utilization of this capability across various tasks, but also from the proactive identification and mitigation of potential risks, including concerns regarding truthfulness, bias, and toxicity, that may arise alongside the capability. In this paper, we present a thorough survey on the interpretation and analysis of in-context learning. First, we provide a concise introduction to the background and definition of in-context learning. Then, we give an overview of advancements from two perspectives: 1) a theoretical perspective, emphasizing studies on mechanistic interpretability and delving into the mathematical foundations behind ICL; and 2) an empirical perspective, concerning studies that empirically analyze factors associated with ICL. We conclude by discussing open questions and the challenges encountered, and suggesting potential avenues for future research. We believe that our work establishes the basis for further exploration into the interpretation of in-context learning. To aid this effort, we have created a repository containing resources that will be continually updated.

Low-resource Methods for NLP 1

Nov 12 (Tue) 11:00-12:30 - Room: Flagler

11:00 - 11:15 - Flagler

A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression

Alessio Devoto, Yu Zhao, Simone Scardapane, Pasquale Minervini

The deployment of large language models (LLMs) is often hindered by the extensive memory requirements of the Key-Value (KV) cache, especially as context lengths increase. Existing approaches to reduce the KV cache size involve either fine-tuning the model to learn a compression strategy or leveraging attention scores to reduce the sequence length. We analyse the attention distributions in decoder-only Transformers-based models and observe that attention allocation patterns stay consistent across most layers. Surprisingly, we find a clear correlation between the L_2 norm and the attention scores over cached KV pairs, where a low L_2 norm of a key embedding usually leads to a high attention score during decoding. This finding indicates that the influence of a KV pair is potentially determined by the key embedding itself before being queried. Based on this observation, we compress the KV cache based on the L_2 norm of key embeddings. Our experimental results show that this simple strategy can reduce the KV cache size by 50% on language modelling and needle-in-a-haystack tasks and 90% on passkey retrieval tasks without losing accuracy. Moreover, without relying on the attention scores, this approach remains compatible with FlashAttention, enabling broader applicability.

11:15 - 11:30 - Flagler

Mixture-of-Subspaces in Low-Rank Adaptation*Taigiang Wu, Jiahao Wang, Zhe Zhao, Ngai Wong*

In this paper, we introduce a subspace-inspired Low-Rank Adaptation (LoRA) method, which is computationally efficient, easy to implement, and readily applicable to large language, multimodal, and diffusion models. Initially, we equivalently decompose the weights of LoRA into two subspaces, and find that simply mixing them can enhance performance. To study such a phenomenon, we revisit it through a fine-grained subspace lens, showing that such modification is equivalent to employing a fixed mixer to fuse the subspaces. To be more flexible, we jointly learn the mixer with the original LoRA weights, and term the method as Mixture-of-Subspaces LoRA (MoSLoRA). MoSLoRA consistently outperforms LoRA on tasks in different modalities, including commonsense reasoning, visual instruction tuning, and subject-driven text-to-image generation, demonstrating its effectiveness and robustness.

11:30 - 11:45 - Flagler

Model Balancing Helps Low-data Training and Fine-tuning*Zihang Liu, Yuanzhe Hu, Tianyu Pang, Yefan Zhou, Pu Ren, Yaoqing Yang*

Recent advances in foundation models have emphasized the need to align pre-trained models with specialized domains using small, curated datasets. Studies on these foundation models underscore the importance of low-data training and fine-tuning. This topic, well-known in natural language processing (NLP), has also gained increasing attention in the emerging field of scientific machine learning (SciML). To address the limitations of low-data training and fine-tuning, we draw inspiration from Heavy-Tailed Self-Regularization (HT-SR) theory, analyzing the shape of empirical spectral densities (ESDs) and revealing an imbalance in training quality across different model layers. To mitigate this issue, we adapt a recently proposed layer-wise learning rate scheduler, TempBalance, which effectively balances training quality across layers and enhances low-data training and fine-tuning for both NLP and SciML tasks. Notably, TempBalance demonstrates increasing performance gains as the amount of available tuning data decreases. Comparative analyses further highlight the effectiveness of TempBalance and its adaptability as an add-on method for improving model performance.

11:45 - 12:00 - Flagler

SciPrompt: Knowledge-Augmented Prompting for Fine-Grained Categorization of Scientific Topics*Zhiwen You, Kanyao Han, Huaonian Zhu, Bertram Ludeascher, Jana Diesner*

Prompt-based fine-tuning has become an essential method for eliciting information encoded in pre-trained language models for a variety of tasks, including text classification. For multi-class classification tasks, prompt-based fine-tuning under low-resource scenarios has resulted in performance levels comparable to those of fully fine-tuning methods. Previous studies have used crafted prompt templates and verbalizers, mapping from the label terms space to the class space, to solve the classification problem as a masked language modeling task. However, cross-domain and fine-grained prompt-based fine-tuning with an automatically enriched verbalizer remains unexplored, mainly due to the difficulty and costs of manually selecting domain label terms for the verbalizer, which requires humans with domain expertise. To address this challenge, we introduce SciPrompt, a framework designed to automatically retrieve scientific topic-related terms for low-resource text classification tasks. To this end, we select semantically correlated and domain-specific label terms within the context of scientific literature for verbalizer augmentation. Furthermore, we propose a new verbalization strategy that uses correlation scores as additional weights to enhance the prediction performance of the language model during model tuning. Our method outperforms state-of-the-art, prompt-based fine-tuning methods on scientific text classification tasks under few and zero-shot settings, especially in classifying fine-grained and emerging scientific topics.

12:00 - 12:15 - Flagler

The Zenos Paradox of Low-Resource Languages*Hellina Hailu Nigatu, Amafua Lambebo Tonja, Benjamin Rosman, Thamar Solorio, Monojit Choudhury*

The disparity in the languages commonly studied in Natural Language Processing (NLP) is typically reflected by referring to languages as low vs high-resourced. However, there is limited consensus on what exactly qualifies as a 'low-resource language.' To understand how NLP papers define and study 'low resource' languages, we qualitatively analyzed 150 papers from the ACL Anthology and popular speech-processing conferences that mention the keyword 'low-resource.' Based on our analysis, we show how several interacting axes contribute to 'low-resourcedness' of a language and why that makes it difficult to track progress for each individual language. We hope our work (1) elicits explicit definitions of the terminology when it is used in papers and (2) provides grounding for the different axes to consider when connoting a language as low-resource.

12:15 - 12:30 - Flagler

Unsupervised Named Entity Disambiguation for Low Resource Domains*Debarghya Datta, Soumajit Pramanik*

In the ever-evolving landscape of natural language processing and information retrieval, the need for robust and domain-specific entity linking algorithms has become increasingly apparent. It is crucial in a considerable number of fields such as humanities, technical writing and biomedical sciences to enrich texts with semantics and discover more knowledge. The use of Named Entity Disambiguation (NED) in such domains requires handling noisy texts, low resource settings and domain-specific KBs. Existing approaches are mostly inappropriate for such scenarios, as they either depend on training data or are not flexible enough to work with domain-specific KBs. Thus in this work, we present a unsupervised approach leveraging the concept of Group Steiner Trees (GST), which can identify the most relevant candidate for entity disambiguation using the contextual similarities across candidate entities for all the mentions present in a document. We outperform the state-of-the-art unsupervised methods by more than 40%(in avg) in terms of Precision@1 and Hit@5 across various domain-specific datasets.

Human-centered NLP 1

Nov 12 (Tue) 11:00-12:30 - Room: Monroe

11:00 - 11:15 - Monroe

A User-Centric Multi-Intent Benchmark for Evaluating Large Language Models*Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, Jian-Yun Nie*

Large language models (LLMs) are essential tools that users employ across various scenarios, so evaluating their performance and guiding users in selecting the suitable service is important. Although many benchmarks exist, they mainly focus on specific predefined model abilities, such as world knowledge, reasoning, etc. Based on these ability scores, it is hard for users to determine which LLM best suits their particular needs. To address these issues, we propose to evaluate LLMs from a user-centric perspective and design this benchmark to measure their efficacy in satisfying user needs under distinct intents. Firstly, we collect 1,846 real-world use cases from a user study with 712 participants from

23 countries. This first-hand data helps us understand actual user intents and needs in LLM interactions, forming the User Reported Scenarios (URS) dataset, which is categorized with six types of user intents. Secondly, based on this authentic dataset, we benchmark 10 LLM services with GPT-4-as-Judge. Thirdly, we show that benchmark scores align well with human preference in both real-world experience and pair-wise annotations, achieving Pearson correlations of 0.95 and 0.94, respectively. This alignment confirms that the URS dataset and our evaluation method establish an effective user-centric benchmark. The dataset, code, and process data are publicly available at <https://github.com/Al-ice1998/URS>.

11:15 - 11:30 - Monroe

ACE: A LLM-based Negotiation Coaching System

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, Zhou Yu

The growing prominence of LLMs has led to an increase in the development of AI tutoring systems. These systems are crucial in providing underrepresented populations with improved access to valuable education. One important area of education that is unavailable to many learners is strategic bargaining related to negotiation. To address this, we develop a LLM-based Assistant for Coaching nEgotiation (ACE). ACE not only serves as a negotiation partner for users but also provides them with targeted feedback for improvement. To build our system, we collect a dataset of negotiation transcripts between MBA students. These transcripts come from trained negotiators and emulate realistic bargaining scenarios. We use the dataset, along with expert consultations, to design an annotation scheme for detecting negotiation mistakes. ACE employs this scheme to identify mistakes and provide targeted feedback to users. To test the effectiveness of ACE-generated feedback, we conducted a user experiment with two consecutive trials of negotiation and found that it improves negotiation performances significantly compared to a system that doesn't provide feedback and one which uses an alternative method of providing feedback.

11:30 - 11:45 - Monroe

Do LLMs Plan Like Human Writers? Comparing Journalist Coverage of Press Releases with LLMs

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, Mark Dredze

Journalists engage in multiple steps in the news writing process that depend on human creativity, like exploring different "angles" (i.e. the specific perspectives a reporter takes). These can potentially be aided by large language models (LLMs). By affecting planning decisions, such interventions can have an outsize impact on creative output. We advocate a careful approach to evaluating these interventions to ensure alignment with human values. In a case study of journalistic coverage of press releases, we assemble a large dataset of 250k press releases and 650k articles covering them. We develop methods to identify news articles that _challenge_ and contextualize_ press releases. Finally, we evaluate suggestions made by LLMs for these articles and compare these with decisions made by human journalists. Our findings are three-fold: (1) Human-written news articles that challenge_ and contextualize_ press releases more take more creative angles and use more informational sources. (2) LLMs align better with humans when recommending angles, compared with informational sources. (3) Both the angles and sources LLMs suggest are significantly less creative than humans.

Human-centered NLP 2

Nov 12 (Tue) 11:00-12:30 - Room: Monroe

11:45 - 12:00 - Monroe

Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles

Ryan Louie, Anjanan Nandi, William Fang, Cheng Chang, Emma Brunskill, Diyi Yang

Recent works leverage LLMs to roleplay realistic social scenarios, aiding novices in practicing their social skills. However, simulating sensitive interactions, such as in the domain of mental health, is challenging. Privacy concerns restrict data access, and collecting expert feedback, although vital, is laborious. To address this, we develop Roleplay-doh, a novel human-LLM collaboration pipeline that elicits qualitative feedback from a domain-expert, which is transformed into a set of principles, or natural language rules, that govern an LLM-prompted roleplay. We apply this pipeline to enable senior mental health supporters to create customized AI patients as simulated practice partners for novice counselors. After uncovering issues with basic GPT-4 simulations not adhering to expert-defined principles, we also introduce a novel principle-adherence prompting pipeline which shows a 30% improvement in response quality and principle following for the downstream task. Through a user study with 25 counseling experts, we demonstrate that the pipeline makes it easy and effective to create AI patients that more faithfully resemble real patients, as judged by both creators and third-party counselors. We provide access to the code and data on our project website: <https://roleplay-doh.github.io/>.

Human-centered NLP 1

Nov 12 (Tue) 11:00-12:30 - Room: Monroe

12:00 - 12:15 - Monroe

Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method

Yang Trista Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle L. Mazurek, Katherine Shilton, Hal Daumé III

Extensive efforts in automated approaches for content moderation have been focused on developing models to identify toxic, offensive, and hateful content with the aim of lightening the load for moderators. Yet, it remains uncertain whether improvements on those tasks have truly addressed moderators' needs in accomplishing their work. In this paper, we surface gaps between past research efforts that have aimed to provide automation for aspects of content moderation and the needs of volunteer content moderators, regarding identifying violations of various moderation rules. To do so, we conduct a model review on Hugging Face to reveal the availability of models to cover various moderation rules and guidelines from three exemplar forums. We further put state-of-the-art LLMs to the test, evaluating how well these models perform in flagging violations of platform rules from one particular forum. Finally, we conduct a user survey study with volunteer moderators to gain insight into their perspectives on useful moderation models. Overall, we observe a non-trivial gap, as missing developed models and LLMs exhibit moderate to low performance on a significant portion of the rules. Moderator reports provide guides for future work on developing moderation assistant models.

12:15 - 12:30 - Monroe

Successfully Guiding Humans with Imperfect Instructions by Highlighting Potential Errors and Suggesting Corrections

Min-Hsuan Yeh, Ruyuan Wan, Ting-Hao Kenneth Huang

Language models will inevitably err in situations with which they are unfamiliar. However, by effectively communicating uncertainties, they can still guide humans toward making sound decisions in those contexts. We demonstrate this idea by developing HEAR, a system that can successfully guide humans in simulated residential environments despite generating potentially inaccurate instructions. Diverging from systems that provide users with only the instructions they generate, HEAR warns users of potential errors in its instructions and suggests corrections. This rich uncertainty information effectively prevents misguidance and reduces the search space for users. Evaluation with 80 users shows that HEAR achieves a 13% increase in success rate and a 29% reduction in final location error distance compared to only presenting instructions to users. Interestingly, we find that offering users possibilities to explore, HEAR motivates them to make more attempts at the task, ultimately leading to a higher success rate. To our best knowledge, this work is the first to show the practical benefits of uncertainty communication in a long-horizon sequential decision-making problem.

Machine Translation 1

Nov 12 (Tue) 11:00-12:30 - Room: Tuttle

11:00 - 11:15 - Tuttle

Neuron Specialization: Leveraging Intrinsic Task Modularity for Multilingual Machine Translation

Shaonu Tan, Di Wu, Christof Monz

Training a unified multilingual model promotes knowledge transfer but inevitably introduces negative interference. Language-specific modeling methods show promise in reducing interference. However, they often rely on heuristics to distribute capacity and struggle to foster cross-lingual transfer via isolated modules. In this paper, we explore intrinsic task modularity within multilingual networks and leverage these observations to circumvent interference under multilingual translation. We show that neurons in the feed-forward layers tend to be activated in a language-specific manner. Meanwhile, these specialized neurons exhibit structural overlaps that reflect language proximity, which progress across layers. Based on these findings, we propose Neuron Specialization, an approach that identifies specialized neurons to modularize feed-forward layers and then continuously updates them through sparse networks. Extensive experiments show that our approach achieves consistent performance gains over strong baselines with additional analyses demonstrating reduced interference and increased knowledge transfer.

11:15 - 11:30 - Tuttle

PsFuture: A Pseudo-Future-based Zero-Shot Adaptive Policy for Simultaneous Machine Translation

Libo Zhao, Jing Li, Ziqian Zeng

Simultaneous Machine Translation (SiMT) requires target tokens to be generated in real-time as streaming source tokens are consumed. Traditional approaches to SiMT typically require sophisticated architectures and extensive parameter configurations for training adaptive read/write policies, which in turn demand considerable computational power and memory. We propose PsFuture, the first zero-shot adaptive read/write policy for SiMT, enabling the translation model to independently determine read/write actions without the necessity for additional training. Furthermore, we introduce a novel training strategy, Prefix-to-Full (P2F), specifically tailored to adjust offline translation models for SiMT applications, exploiting the advantages of the bidirectional attention mechanism inherent in offline models. Experiments across multiple benchmarks demonstrate that our zero-shot policy attains performance on par with strong baselines and the P2F method can further enhance performance, achieving an outstanding trade-off between translation quality and latency.

11:30 - 11:45 - Tuttle

Simul-MuST-C: Simultaneous Multilingual Speech Translation Corpus Using Large Language Model

Mana Makinae, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe

Simultaneous Speech Translation (SiST) begins translating before the entire source input is received, making it crucial to balance quality and latency. In real interpreting situations, interpreters manage this simultaneity by breaking sentences into smaller segments and translating them while maintaining the source order as much as possible. SiST could benefit from this approach to balance quality and latency. However, current corpora used for simultaneous tasks often involve significant word reordering in translation, which is not ideal given that interpreters faithfully follow source syntax as much as possible. Inspired by conference interpreting by humans utilizing the salami technique, we introduce the Simul-MuST-C, a dataset created by leveraging the Large Language Model (LLM), specifically GPT-4o, which aligns the target text as closely as possible to the source text by using minimal chunks that contain enough information to be interpreted. Experiments on three language pairs show that the effectiveness of segmented-base monotonicity in training data varies with the grammatical distance between the source and the target, with grammatically distant language pairs benefiting the most in achieving quality while minimizing latency.

11:45 - 12:00 - Tuttle

Towards Cross-Cultural Machine Translation with Retrieval-Augmented Generation from Multilingual Knowledge Graphs

Simone Comia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Podar, Yanyao Li

Translating text that contains entity names is a challenging task, as cultural-related references can vary significantly across languages. These variations may also be caused by transcription, an adaptation process that entails more than transliteration and word-for-word translation. In this paper, we address the problem of cross-cultural translation on two fronts: (i) we introduce XC-Translate, the first large-scale, manually-created benchmark for machine translation that focuses on text that contains potentially culturally-nuanced entity names, and (ii) we propose KG-MT, a novel end-to-end method to integrate information from a multilingual knowledge graph into a neural machine translation model by leveraging a dense retrieval mechanism. Our experiments and analyses show that current machine translation systems and large language models still struggle to translate texts containing entity names, whereas KG-MT outperforms state-of-the-art approaches by a large margin, obtaining a 129% and 62% relative improvement compared to NLLB-200 and GPT-4, respectively.

12:00 - 12:15 - Tuttle

What do large language models need for machine translation evaluation?

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Dipesh Kanodia, Constantin Orasan, Tharindu Ranasinghe, Fred Blain

Leveraging large language models (LLMs) for various natural language processing tasks has led to superlative claims about their performance. For the evaluation of machine translation (MT), existing research shows that LLMs are able to achieve results comparable to fine-tuned multilingual pre-trained language models. In this paper, we explore what translation information, such as the source, reference, translation errors and annotation guidelines, is needed for LLMs to evaluate MT quality. In addition, we investigate prompting techniques such as zero-shot, Chain of Thought (CoT) and few-shot prompting for eight language pairs covering high-, medium- and low-resource languages, leveraging varying LLM variants. Our findings indicate the importance of reference translations for an LLM-based evaluation. While larger models do not necessarily fare better, they tend to benefit more from CoT prompting, than smaller models. We also observe that LLMs do not always provide a numerical score when generating evaluations, which poses a question on their reliability for the task. Our work presents a comprehensive analysis for resource-constrained and training-less LLM-based evaluation of machine translation. We release the accrued prompt

templates, code and data publicly for reproducibility.

12:15 - 12:30 - Tuttle

Word Alignment as Preference for Machine Translation

Qiuyu Wu, Masaaki Nagata, Zhongtao Miao, Yoshimasa Tsuruoka

The problem of hallucination and omission, a long-standing problem in machine translation (MT), is more pronounced when a large language model (LLM) is used in MT because an LLM itself is susceptible to these phenomena. In this work, we mitigate the problem in an LLM-based MT model by guiding it to better word alignment. We first study the correlation between word alignment and the phenomena of hallucination and omission in MT. Then we propose to utilize word alignment as preference to optimize the LLM-based MT model. The preference data are constructed by selecting chosen and rejected translations from multiple MT tools. Subsequently, direct preference optimization is used to optimize the LLM-based model towards the preference signal. Given the absence of evaluators specifically designed for hallucination and omission in MT, we further propose selecting hard instances and utilizing GPT-4 to directly evaluate the performance of the models in mitigating these issues. We verify the rationality of these designed evaluation methods by experiments, followed by extensive results demonstrating the effectiveness of word alignment-based preference optimization to mitigate hallucination and omission. On the other hand, although it shows promise in mitigating hallucination and omission, the overall performance of MT in different language directions remains mixed, with slight increases in BLEU and decreases in COMET.

Session 03 - Nov 12 (Tue) 14:00-15:30

Generation and Summarization

Nov 12 (Tue) 14:00-15:30 - Room: Ashe Auditorium

14:00 - 14:15 - Ashe Auditorium

Chain-of-Dictionary Prompting Elicits Translation in Large Language Models

Hongyuan Lu, HAORAN YANG, Haoyang Huang, Dongdong Zhang, Wai Lam, Furu Wei

Large language models (LLMs) have shown surprisingly good performance in multilingual neural machine translation (MNMT) even if not being trained explicitly for translation. Yet, they still struggle with translating low-resource languages. As supported by our experiments, a bilingual dictionary between the source and the target language could help. Motivated by the fact that multilingual training effectively improves cross-lingual performance, we show that a chained multilingual dictionary with words expressed in more languages can provide more information to better enhance the LLM translation. To this end, we present a novel framework, CoD, Chain-of-Dictionary Prompting, which augments LLMs with prior knowledge with the chains of multilingual dictionaries for a subset of input words to elicit translation abilities for LLMs. Experiments indicate that ChatGPT and InstructGPT still have room for improvement in translating many language pairs. And CoD elicits large gains by up to 13x chrF++ points for MNMT (3.08 to 42.63 for English to Serbian written in Cyrillic script) on FLORES-200 full devtest set. We demonstrate the importance of chaining the multilingual dictionaries, as well as the superiority of CoD to few-shot-in-context learning for low-resource languages. Using CoD helps ChatGPT to obviously surpass the SOTA translator NLLB 3.3B.

14:15 - 14:30 - Ashe Auditorium

Evaluating LLMs for Targeted Concept Simplification for Domain-Specific Texts

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, Mirella Lapata

One useful application of NLP models is to support people in reading complex text from unfamiliar domains (e.g., scientific articles). Simplifying the entire text makes it understandable but sometimes removes important details. On the contrary, helping adult readers understand difficult concepts in context can enhance their vocabulary and knowledge. In a preliminary human study, we first identify that lack of context and unfamiliarity with difficult concepts is a major reason for adult readers' difficulty with domain-specific text. We then introduce targeted concept simplification, a simplification task for rewriting text to help readers comprehend text containing unfamiliar concepts. We also introduce WikiDomains, a new dataset of 22k definitions from 13 academic domains paired with a difficult concept within each definition. We benchmark the performance of open-source and commercial LLMs and a simple dictionary baseline on this task across human judgments of ease of understanding and meaning preservation. Interestingly, our human judges preferred explanations about the difficult concept more than simplifications of the concept phrase. Further, no single model achieved superior performance across all quality dimensions, and automated metrics also show low correlations with human evaluations of concept simplification (~ 0.2), opening up rich avenues for research on personalized human reading comprehension support.

14:30 - 14:45 - Ashe Auditorium

Induct-Learn: Short Phrase Prompting with Instruction Induction

Po-Chun Chen, Sheng-Lun Wei, Hen-Hsen Huang, Hsin-Hsi Chen

Large Language Models (LLMs) have demonstrated capability in "instruction induction," generating instructions from demonstrations (input-output pairs). However, existing methods often rely on large datasets or numerous examples, which is impractical and costly in real-world scenarios. In this work, we propose a low-cost, task-level framework called Induct-Learn. It induces pseudo instructions from a few demonstrations and a short phrase, adding a CoT process into existing demonstrations. When encountering new problems, the learned pseudo instructions and demonstrations with the pseudo CoT process can be combined into a prompt to guide the LLM's problem-solving process. We validate our approach on the BBH-Induct and Eval-Induct datasets, and the results show that the Induct-Learn framework outperforms state-of-the-art methods. We also exhibit cross-model adaptability and achieve superior performance at a lower cost compared to existing methods.

14:45 - 15:00 - Ashe Auditorium

Learning to Generate Writing Feedback via Language Model Simulated Student Revisions

Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, Lu Wang

Providing feedback is widely recognized as crucial for refining students writing skills. Recent advances in language models (LMs) have made it possible to automatically generate feedback that is actionable and well-aligned with human-specified attributes. However, it remains unclear whether the feedback generated by these models is truly effective in enhancing the quality of student revisions. Moreover, prompting LMs with a precise set of instructions to generate feedback is nontrivial due to the lack of consensus regarding the specific attributes that can lead to improved revising performance. To address these challenges, we propose PROF that PROduces Feedback via learning from LM simulated student revisions. PROF aims to iteratively optimize the feedback generator by directly maximizing the effectiveness of students' overall revising performance as simulated by LMs. Focusing on an economic essay assignment, we empirically test the efficacy of PROF and observe that

our approach not only surpasses a variety of baseline methods in effectiveness of improving students writing but also demonstrates enhanced pedagogical values, even though it was not explicitly trained for this aspect.

15:00 - 15:15 - Ashe Auditorium

Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, Md Shad Akhtar

In mental health counseling, condensing dialogues into concise and relevant summaries (aka counseling notes) holds pivotal significance. Large Language Models (LLMs) exhibit remarkable capabilities in various generative tasks; however, their adaptation to domain-specific intricacies remains challenging, especially within mental health contexts. Unlike standard LLMs, mental health experts first plan to apply domain knowledge in writing summaries. Our work enhances LLMs' ability by introducing a novel planning engine to orchestrate structuring knowledge alignment. To achieve high-order planning, we divide knowledge encapsulation into two major phases: (i) holding dialogue structure and (ii) incorporating domain-specific knowledge. We employ a planning engine on Llama-2, resulting in a novel framework, PIECE. Our proposed system employs knowledge filtering-cum-scaffolding to encapsulate domain knowledge. Additionally, PIECE leverages sheaf convolution learning to enhance its understanding of the dialogue's structural nuances. We compare PIECE with 14 baseline methods and observe a significant improvement across ROUGE and Bleurt scores. Further, expert evaluation and analyses validate the generation quality to be effective, sometimes even surpassing the gold standard. We further benchmark PIECE with other LLMs and report improvement, including Llama-2 (+2.72%), Mistral (+2.04%), and Zephyr (+1.59%), to justify the generalizability of the planning engine.

15:15 - 15:30 - Ashe Auditorium

STORYSUMM: Evaluating Faithfulness in Story Summarization

Melanie Subbiah, Faisal Ladha, Akanksha Mishra, Griffin Thomas Adams, Lydia Chilton, Kathleen McKeown

Human evaluation has been the gold standard for checking faithfulness in abstractive summarization. However, with a challenging source domain like narrative, multiple annotators can agree a summary is faithful, while missing details that are obvious errors only once pointed out. We therefore introduce a new dataset, StorySumm, comprising LLM summaries of short stories with localized faithfulness labels and error explanations. This benchmark is for evaluation methods, testing whether a given method can detect challenging inconsistencies. Using this dataset, we first show that any one human annotation protocol is likely to miss inconsistencies, and we advocate for pursuing a range of methods when establishing ground truth for a summarization dataset. We finally test recent automatic metrics and find that none of them achieve more than 70% balanced accuracy on this task, demonstrating that it is a challenging benchmark for future work in faithfulness evaluation.

Dialogue and Interactive Systems 1

Nov 12 (Tue) 14:00-15:30 - Room: Brickell

14:00 - 14:15 - Brickell

Global Reward to Local Rewards: Multimodal-Guided Decomposition for Improving Dialogue Agents

Dong Won Lee, Hae Won Park, Yoon Kim, Cynthia Brozale, Louis-Philippe Morency

We describe an approach for aligning an LLM based dialogue agent for long-term social dialogue, where there is only a single global score given by the user at the end of the session. In this paper, we propose the usage of denser naturally-occurring multimodal communicative signals as local implicit feedback to improve the turn-level utterance generation. Therefore, our approach (dubbed GELI) learns a local, turn-level reward model by decomposing the human-provided Global Explicit (GE) session level reward, using Local Implicit (LI) multimodal reward signals to crossmodally shape the reward decomposition step. This decomposed reward model is then used as part of the RLHF pipeline to improve an LLM-based dialog agent. We run quantitative and qualitative human studies on two large-scale datasets to evaluate the performance of our GELI approach, and find that it shows consistent improvements across various conversational metrics compared to baseline methods.

14:15 - 14:30 - Brickell

Mitigating Matthew Effect: Multi-Hypergraph Boosted Multi-Interest Self-Supervised Learning for Conversational Recommendation

Yongsen Zheng, Ruilin Xu, Guohua Wang, Liang Lin

The Matthew effect is a big challenge in Recommender Systems (RSs), where popular items tend to receive increasing attention, while less popular ones are often overlooked, perpetuating existing disparities. Although many existing methods attempt to mitigate Matthew effect in the static or quasi-static recommendation scenarios, such issue will be more pronounced as users engage with the system over time. To this end, we propose a novel framework, Multi-Hypergraph Boosted Multi-Interest Self-Supervised Learning for Conversational Recommendation (HiCore), aiming to address Matthew effect in the Conversational Recommender System (CRS) involving the dynamic user-system feedback loop. It devotes to learn multi-level user interests by building a set of hypergraphs (i.e., item-, entity-, word-oriented multiple-channel hypergraphs) to alleviate the Matthew effect. Extensive experiments on four CRS-based datasets showcase that HiCore attains a new state-of-the-art performance, underscoring its superiority in mitigating the Matthew effect effectively. Our code is available at <https://github.com/zysense/HiCore>.

14:30 - 14:45 - Brickell

PANDA: Persona Attributes Navigation for Detecting and Alleviating Overuse Problem in Large Language Models

Jinsung Kim, Seonmin Koo, Heusik Lim

In the persona-grounded dialogue (PGD) task, it is required not only to respond fluently, but also to ground the attributes according to the current conversation topic properly. However, due to their tendency to overly ground given attributes, LLMs often generate unnatural responses provoked by using attributes that deviate from the flow of the conversation or by exploiting too many attributes at once. We term this phenomenon the *overuse* problem of LLMs. Unfortunately, research devising precise criteria and frameworks to quantitatively verify LLMs' *overuse* problem is obviously insufficient. To address this issue, we propose ***Persona ***A***ttributes ***N***avigation for ***D***etecting and ***A***lleviating the *overuse* problem (**PANDA***) framework. **PANDA*** is the first study to quantify the persona *overuse* problem of LLMs by establishing clear standards of the problem and verifying various LLMs based on them. Moreover, this framework navigates us into understanding persona attributes by introducing diverse and detailed dialogue topics that consider practical conversation situations. We provide insights related to LLMs' persona attribute *overuse* problem through comprehensive verification and analysis with **PANDA*** in the PGD task. Our code and resources can be found at <http://github.com/jin62304/PANDA>.

14:45 - 15:00 - Brickell

Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models

Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi

In dialogue, the addressee may initially misunderstand the speaker and respond erroneously, often prompting the speaker to correct the mis-understanding in the next turn with a Third Position Repair (TPR). The ability to process and respond appropriately to such repair sequences is thus crucial in conversational AI systems. In this paper, we first collect, analyse, and publicly release BlockWorld-Repairs: a dataset of multi-modal TPR sequences in an instruction-following manipulation task that is, by design, rife with referential ambiguity. We employ this dataset to evaluate several state-of-the-art Vision and Language Models (VLM) across multiple settings, focusing on their capability to process and accurately respond to TPRs and thus recover from miscommunication. We find that, compared to humans, all models significantly underperform in this task. We then show that VLMs can benefit from specialised losses targeting relevant tokens during fine-tuning, achieving better performance and generalising better to new scenarios. Our results suggest that these models are not yet ready to be deployed in multi-modal collaborative settings where repairs are common, and highlight the need to design training regimes and objectives that facilitate learning from interaction. Our code and data are available at www.github.com/JChiyah/blockworld-repairs

Dialogue and Interactive Systems 2

Nov 12 (Tue) 14:00-15:30 - Room: Brickell

15:00 - 15:15 - Brickell

Do LLMs suffer from Multi-Party Hangover? A Diagnostic Approach to Addressee Recognition and Response Selection in Conversations

Nicolo Penzo, Maryam Sajedinia, Bruno Lepri, Sara Tonelli, Marco Guerini

Assessing the performance of systems to classify Multi-Party Conversations (MPC) is challenging due to the interconnection between linguistic and structural characteristics of conversations. Conventional evaluation methods often overlook variances in model behavior across different levels of structural complexity on interaction graphs. In this work, we propose a methodological pipeline to investigate model performance across specific structural attributes of conversations. As a proof of concept we focus on Response Selection and Addressee Recognition tasks, to diagnose model weaknesses. To this end, we extract representative diagnostic subdatasets with a fixed number of users and a good structural variety from a large and open corpus of online MPCs. We further frame our work in terms of data minimization, avoiding the use of original usernames to preserve privacy, and propose alternatives to using original text messages. Results show that response selection relies more on the textual content of conversations, while addressee recognition requires capturing their structural dimension. Using an LLM in a zero-shot setting, we further highlight how sensitivity to prompt variations is task-dependent.

Dialogue and Interactive Systems 1

Nov 12 (Tue) 14:00-15:30 - Room: Brickell

15:15 - 15:30 - Brickell

Unsupervised End-to-End Task-Oriented Dialogue with LLMs: The Power of the Noisy Channel

Brendan King, Jeffrey Flanigan

Training task-oriented dialogue systems typically requires turn-level annotations for interacting with their APIs: e.g. a dialogue state and the system actions taken at each step. These annotations can be costly to produce, error-prone, and require both domain and annotation expertise. With advances in LLMs, we hypothesize that unlabeled data and a schema definition are sufficient for building a working task-oriented dialogue system, completely unsupervised. We consider a novel unsupervised setting of only (1) a well-defined API schema (2) a set of unlabeled dialogues between a user and agent. We propose an innovative approach using expectation-maximization (EM) that infers turn-level annotations as latent variables using a noisy channel model to build an end-to-end dialogue agent. Evaluating our approach on the MultiWOZ benchmark, our method more than doubles the dialogue success rate of a strong GPT-3.5 baseline.

Computational Social Science and Cultural Analytics 1

Nov 12 (Tue) 14:00-15:30 - Room: Flager

14:00 - 14:15 - Flager

"We Demand Justice!": Towards Social Context Grounding of Political Texts

Rajkumar Pujari, Chengfei Wu, Dan Goldwasser

Political discourse on social media often contains similar language with opposing intended meanings. For example, the phrase thoughts and prayers, is used to express sympathy for mass shooting victims, as well as satirically criticize the lack of legislative action on gun control. Understanding such discourse fully by reading only the text is difficult. However, knowledge of the social context information makes it easier. We characterize the social context required to fully understand such ambiguous discourse, by grounding the text in real-world entities, actions, and attitudes. We propose two datasets that require understanding social context and benchmark them using large pre-trained language models and several novel structured models. We show that structured models, explicitly modeling social context, outperform larger models on both tasks, but still lag significantly behind human performance. Finally, we perform an extensive analysis to obtain further insights into the language understanding challenges posed by our social grounding tasks.

14:15 - 14:30 - Flager

Fine-Grained Detection of Solidarity for Women and Migrants in 155 Years of German Parliamentary Debates

Aida Kostikova, Dominik Beese, Benjamin Paassen, Ole Pütz, Gregor Wiedemann, Steffen Eger

Solidarity is a crucial concept to understand social relations in societies. In this study, we investigate the frequency of (anti-)solidarity towards women and migrants in German parliamentary debates between 1867 and 2022. Using 2,864 manually annotated text snippets, we evaluate large language models (LLMs) like Llama 3, GPT-3.5, and GPT-4. We find that GPT-4 outperforms other models, approaching human annotation accuracy. Using GPT-4, we automatically annotate 18,300 further instances and find that solidarity with migrants outweighs anti-solidarity but that frequencies and solidarity types shift over time. Most importantly, group-based notions of (anti-)solidarity fade in favor of compassionate solidarity, focusing on the vulnerability of migrant groups, and exchange-based anti-solidarity, focusing on the lack of (economic) contribution. This study highlights the interplay of historical events, socio-economic needs, and political ideologies in shaping

migration discourse and social cohesion.

14:30 - 14:45 - Flagler

HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs

Jocelyn J Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, Maarten Sap

Empathy serves as a cornerstone in enabling prosocial behaviors, and can be evoked through sharing of personal experiences in stories. While empathy is influenced by narrative content, intuitively, people respond to the way a story is told as well, through narrative style. Yet the relationship between empathy and narrative style is not fully understood. In this work, we empirically examine and quantify this relationship between style and empathy using LLMs and large-scale crowdsourcing studies. We introduce a novel, theory-based taxonomy, HEART (Human Empathy and Narrative Taxonomy) that delineates elements of narrative style that can lead to empathy with the narrator of a story. We establish the performance of LLMs in extracting narrative elements from HEART, showing that prompting with our taxonomy leads to reasonable, human-level annotations beyond what prior lexicon-based methods can do. To show empirical use of our taxonomy, we collect a dataset of empathy judgments of stories via a large-scale crowdsourcing study with $N = 2,624$ participants. We show that narrative elements extracted via LLMs, in particular, vividness of emotions and plot volume, can elucidate the pathways by which narrative style cultivates empathy towards personal stories. Our work suggests that such models can be used for narrative analyses that lead to human-centered social and behavioral insights.

14:45 - 15:00 - Flagler

On the Relationship between Truth and Political Bias in Language Models

Suyash Falay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, Jad Kabbara

Language model alignment research often attempts to ensure that models are not only helpful and harmless, but also truthful and unbiased. However, optimizing these objectives simultaneously can obscure how improving one aspect might impact the others. In this work, we focus on analyzing the relationship between two concepts essential in both language model alignment and political science: truthfulness and political bias. We train reward models on various popular truthfulness datasets and subsequently evaluate their political bias. Our findings reveal that optimizing reward models for truthfulness on these datasets tends to result in a left-leaning political bias. We also find that existing open-source reward models (i.e., those trained on standard human preference datasets) already show a similar bias and that the bias is larger for larger models. These results raise important questions about the datasets used to represent truthfulness, potential limitations of aligning models to be both truthful and politically unbiased, and what language models capture about the relationship between truth and politics.

15:00 - 15:15 - Flagler

Outcome-Constrained Large Language Models for Countering Hate Speech

Lingzhi Hong, Pengcheng Luo, Eduardo Blanco, Xiaoying Song

Automatic counterspeech generation methods have been developed to assist efforts in combating hate speech. Existing research focuses on generating counterspeech with linguistic attributes such as being polite, informative, and intent-driven. However, the real impact of counterspeech in online environments is seldom considered. This study aims to develop methods for generating counterspeech constrained by conversation outcomes and evaluate their effectiveness. We experiment with large language models (LLMs) to incorporate into the text generation process two desired conversation outcomes: low conversation incivility and non-hateful hate reentry. Specifically, we experiment with instruction prompts, LLM finetuning, and LLM reinforcement learning (RL). Evaluation results show that our methods effectively steer the generation of counterspeech toward the desired outcomes. Our analyses, however, show that there are differences in the quality and style depending on the model.

15:15 - 15:30 - Flagler

Statistical Uncertainty in Word Embeddings: GloVe-V

Andrea Vallebuono, Cassandra Handan-Nader, Christopher D Manning, Daniel E. Ho

Static word embeddings are ubiquitous in computational social science applications and contribute to practical decision-making in a variety of fields including law and healthcare. However, assessing the statistical uncertainty in downstream conclusions drawn from word embedding statistics has remained challenging. When using only point estimates for embeddings, researchers have no streamlined way of assessing the degree to which their model selection criteria or scientific conclusions are subject to noise due to sparsity in the underlying data used to generate the embeddings. We introduce a method to obtain approximate, easy-to-use, and scalable reconstruction error variance estimates for GloVe, one of the most widely used word embedding models, using an analytical approximation to a multivariate normal model. To demonstrate the value of embeddings with variances (GloVe-V), we illustrate how our approach enables principled hypothesis testing in core word embedding tasks, such as comparing the similarity between different word pairs in vector space, assessing the performance of different models, and analyzing the relative degree of ethnic or gender bias in a corpus using different word lists.

Special Theme: Efficiency in Model Algorithms, Training, and Inference 1

Nov 12 (Tue) 14:00-15:30 - Room: Monroe

14:00 - 14:15 - Monroe

DEM: Distribution Edited Model for Training with Mixed Data Distributions

Dhananjay Ram, Aditya Rawal, Momchil Hardalov, Nikolaos Pappas, Sheng Zha

Training with mixed data distributions is a common and important part of creating multi-task and instruction-following models. The diversity of the data distributions and cost of joint training makes the optimization procedure extremely challenging. Data mixing methods partially address this problem, albeit having a sub-optimal performance across data sources and require multiple expensive training runs. In this paper, we propose a simple and efficient alternative for better optimization of the data sources by combining models individually trained on each data source with the base model using basic element-wise vector operations. The resulting model, namely Distribution Edited Model (DEM), is cheaper than standard data mixing and outperforms strong baselines on a variety of benchmarks, yielding upto 6.2% improvement on MMLU, 11.5% on BBH, 16.1% on DROP, 6% MathQA and 9.3% on HELM with models of size 3B to 13B. Notably, DEM does not require full re-training when modifying a single data-source, thus making it very flexible and scalable for training with diverse data sources.

14:15 - 14:30 - Monroe

Explaining and Improving Contrastive Decoding by Extrapolating the Probabilities of a Huge and Hypothetical LM

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, Tagyoung Chung

Contrastive decoding (CD) (Li et al., 2022) improves the next-token distribution of a large expert language model (LM) using a small amateur LM. Although CD is applied to various LMs and domains to enhance open-ended text generation, it is still unclear why CD often works well, when it could fail, and how we can make it better. To deepen our understanding of CD, we first theoretically prove that CD could be

viewed as linearly extrapolating the next-token logits from a huge and hypothetical LM. We also highlight that the linear extrapolation could make CD unable to output the most obvious answers that have already been assigned high probabilities by the amateur LM. To overcome CDs limitation, we propose a new unsupervised decoding method called Asymptotic Probability Decoding (APD). APD explicitly extrapolates the probability curves from the LMs of different sizes to infer the asymptotic probabilities from an infinitely large LM without inducing more inference costs than CD. In FactualityPrompts, an open-ended text generation benchmark, sampling using APD significantly boosts factuality in comparison to the CD sampling and its variants, and achieves state-of-the-art results for Pythia 6.9B and OPT 6.7B. Furthermore, in five commonsense QA datasets, APD is often significantly better than CD and achieves a similar effect of using a larger LLM. For example, the perplexity of APD on top of Pythia 6.9B is even lower than the perplexity of Pythia 12B in CommonsenseQA and LAMBADA.

14:30 - 14:45 - Monroe

FINCH: Key-Value Cache Compression for Large Language Models' Semantic Memory

Giulio Corallo, Paolo Papotti

Recent large language model applications, such as Retrieval-Augmented Generation and chatbots, have led to an increased need to process longer contexts. However, this requirement is hampered by inherent limitations. Architecturally, most models are constrained by a context window defined during training. Additionally, processing extensive texts requires substantial GPU memory and increases the computational complexity required by the self-attention mechanism. We propose a novel compression approach, FLINCH, that leverages the pretrained model weights of the self-attention. Given a prompt and a long text, FLINCH iteratively identifies the most relevant Key (K) and Value (V) pairs over chunks of the text conditioned on the prompt. Only such pairs are stored in the KV cache, which, within the space constrained by the context window, ultimately contains a compressed version of the long text. Our proposal enables models to consume large inputs even with high compression (up to 70x) while preserving semantic integrity without the need for fine-tuning.

14:45 - 15:00 - Monroe

FuseGen: PLM Fusion for Data-generation based Zero-shot Learning

Tianyuan Zou, Yang Liu, Peng Li, Jiangqing Zhang, Jingjing Liu, Ya-Qin Zhang

Data-generation based zero-shot learning, although effective in training Small Task-specific Models (STMs) via synthetic datasets generated by Pre-trained Language Models (PLMs), is often limited by the low quality of such synthetic datasets. Previous solutions have primarily focused on single PLM settings, where synthetic datasets are typically restricted to specific sub-spaces and often deviate from real-world distributions, leading to severe distribution bias. To mitigate such bias, we propose FuseGen, a novel data-generation based zero-shot learning framework that introduces a new criteria for subset selection from synthetic datasets via utilizing multiple PLMs and trained STMs. The chosen subset provides in-context feedback to each PLM, enhancing dataset quality through iterative data generation. Trained STMs are then used for sample re-weighting as well, further improving data quality. Extensive experiments across diverse tasks demonstrate that FuseGen substantially outperforms existing methods, highly effective in boosting STM performance in a PLM-agnostic way. The code is available at <https://github.com/LindaLydia/FuseGen>.

15:00 - 15:15 - Monroe

RoseLoRA: Row and Column-wise Sparse Low-rank Adaptation of Pre-trained Language Model for Knowledge Editing and Fine-tuning

Haoyu Wang, Tianci Liu, Ruirui Li, Monica Xiao Cheng, Tuo Zhao, Jing Gao

Pre-trained language models, trained on large-scale corpora, demonstrate strong generalizability across various NLP tasks. Fine-tuning these models for specific tasks typically involves updating all parameters, which is resource-intensive. Parameter-efficient fine-tuning (PEFT) methods, such as the popular LORA family, introduce low-rank matrices to learn only a few parameters efficiently. However, during inference, the product of these matrices updates all pre-trained parameters, complicating tasks like knowledge editing that require selective updates. We propose a novel PEFT method, which conducts row and column-wise sparse low-rank adaptation (RoseLoRA), to address this challenge. RoseLoRA identifies and updates only the most important parameters for a specific task, maintaining efficiency while preserving other model knowledge. By adding a sparsity constraint on the product of low-rank matrices and converting it to row and column-wise sparsity, we ensure efficient and precise model updates. Our theoretical analysis guarantees the lower bound of the sparsity with respect to the matrix product. Extensive experiments on five benchmarks across twenty datasets demonstrate that RoseLoRA outperforms baselines in both general fine-tuning and knowledge editing tasks.

15:15 - 15:30 - Monroe

RevMUX: Data Multiplexing with Reversible Adapters for Efficient LLM Batch Inference

Yige Xu, Xu Guo, Zhiwei Zeng, Chunyan Miao

Large language models (LLMs) have brought a great breakthrough to the natural language processing (NLP) community, while leading the challenge of handling concurrent customer queries due to their high throughput demands. Data multiplexing addresses this by merging multiple inputs into a single composite input, allowing more efficient inference through a shared forward pass. However, as distinguishing individuals from a composite input is challenging, conventional methods typically require training the entire backbone, yet still suffer from performance degradation. In this paper, we introduce RevMUX, a parameter-efficient data multiplexing framework that incorporates a reversible design in the multiplexer, which can be reused by the demultiplexer to perform reverse operations and restore individual samples for classification. Extensive experiments on four datasets and three types of LLM backbones demonstrate the effectiveness of RevMUX for enhancing LLM inference efficiency while retaining a satisfactory classification performance.

Resources and Evaluation 1

Nov 12 (Tue) 14:00-15:30 - Room: Tuttle

14:00 - 14:15 - Tuttle

Beyond Reference: Evaluating High Quality Translations Better than Human References

Keonwoong Noh, Seokjin Oh, Woochan Jung

In Machine Translation (MT) evaluations, the conventional approach is to compare a translated sentence against its human-created reference sentence. MT metrics provide an absolute score (e.g., from 0 to 1) to a candidate sentence based on the similarity with the reference sentence. Thus, existing MT metrics give the maximum score to the reference sentence. However, this approach overlooks the potential for a candidate sentence to exceed the reference sentence in terms of quality. In particular, recent advancements in Large Language Models (LLMs) have highlighted this issue, as LLM-generated sentences often exceed the quality of human-written sentences. To address the problem, we introduce the Residual score Metric (ResuMe), which evaluates the relative quality between reference and candidate sentences. ResuMe assigns a positive score to candidate sentences that outperform their reference sentences, and a negative score when they fall short. By adding the residual scores from ResuMe to the absolute scores from MT metrics, it can be possible to allocate higher scores to candidate sentences than

what reference sentences are received from MT metrics. Experimental results demonstrate that ResuMe enhances the alignments between MT metrics and human judgments both at the segment-level and the system-level.

14:15 - 14:30 - Tutte

L2CEval: Evaluating Language-to-Code Generation Capabilities of Large Language Models

Ansong Ni, Pengcheng Yin, Yilun Zhao, Martin Riddell, Troy Feng, Rui Shen, Stephen Yin, Ye Liu, Semih Yavuz, Caiming Xiong, Shafiq Joty, Yingbo Zhou, Dragomir Radev, Arman Cohan

Recently, large language models (LLMs), especially those that are pretrained on code, have demonstrated strong capabilities in generating programs from natural language inputs. Despite promising results, there is a notable lack of a comprehensive evaluation of these models' language-to-code generation capabilities. Existing studies often focus on specific tasks, model architectures, or learning paradigms, leading to a fragmented understanding of the overall landscape. In this work, we present L2CEval, a systematic evaluation of the language-to-code generation capabilities of LLMs on 7 tasks across the domain spectrum of semantic parsing, math reasoning and Python programming, analyzing the factors that potentially affect their performance, such as model size, pretraining data, instruction tuning, and different prompting methods. In addition, we assess confidence calibration, and conduct human evaluations to identify typical failures across different tasks and models. L2CEval offers a comprehensive understanding of the capabilities and limitations of LLMs in language-to-code generation. We release the evaluation framework and all model outputs, hoping to lay the groundwork for further future research.

14:30 - 14:45 - Tutte

Large Language Models for Data Annotation: A Survey

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, huan liu

Data annotation and synthesis generally refers to the labeling or generating of raw data with relevant information, which could be used for improving the efficacy of machine learning models. The process, however, is labor-intensive and costly. The emergence of advanced Large Language Models (LLMs), exemplified by GPT-4, presents an unprecedented opportunity to automate the complicated process of data annotation and synthesis. While existing surveys have extensively covered LLM architecture, training, and general applications, we uniquely focus on their specific utility for data annotation. This survey contributes to three core aspects: LLM-Based Annotation Generation, LLM-Generated Annotations Assessment, and LLM-Generated Annotations Utilization. Furthermore, this survey includes an in-depth taxonomy of data types that LLMs can annotate, a comprehensive review of learning strategies for models utilizing LLM-generated annotations, and a detailed discussion of the primary challenges and limitations associated with using LLMs for data annotation and synthesis. Serving as a key guide, this survey aims to assist researchers and practitioners in exploring the potential of the latest LLMs for data annotation, thereby fostering future advancements in this critical field.

14:45 - 15:00 - Tutte

Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA

Minzheng Wang, Longze Chen, ChengFu, LiaoShengyi, Xinghua Zhang, Bingliwu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, Yongbin Li

Long-context modeling capabilities of Large Language Models (LLMs) have garnered widespread attention, leading to the emergence of LLMs with ultra-context windows. Meanwhile, benchmarks for evaluating long-context language models are gradually catching up. However, existing benchmarks employ irrelevant noise texts to artificially extend the length of test cases, diverging from the real-world scenarios of long-context applications. To bridge this gap, we propose a novel long-context benchmark, Loong, aligning with realistic scenarios through extended multi-document question answering (QA). Unlike typical document QA, in Loong's test cases, each document is relevant to the final answer, ignoring any document will lead to the failure of the answer. Furthermore, Loong introduces four types of tasks with a range of context lengths: Spotlight Locating, Comparison, Clustering, and Chain of Reasoning, to facilitate a more realistic and comprehensive evaluation of long-context understanding. Extensive experiments indicate that existing long-context language models still exhibit considerable potential for enhancement. Retrieval augmented generation (RAG) achieves poor performance, demonstrating that Loong can reliably assess the model's long-context modeling capabilities.

15:00 - 15:15 - Tutte

MedReadMe: A Systematic Study for Fine-grained Sentence Readability in Medical Domain

Chao Jiang, Wei Xu

Medical texts are notoriously challenging to read. Properly measuring their readability is the first step towards making them more accessible. Here, we present the first systematic study on fine-grained readability measurements in the medical domain, at both sentence-level and span-level. We first introduce a new dataset MedReadMe, which consists of manually annotated readability ratings and fine-grained complex span annotation for 4,520 sentences, featuring two novel "Google-Easy" and "Google-Hard" categories. It supports our quantitative analysis, which covers 650 linguistic features and additional complex span features, to answer why medical sentences are so hard. Enabled by our high-quality annotation, we benchmark several state-of-the-art sentence-level readability metrics, including unsupervised, supervised, and prompting-based methods using recently developed large language models (LLMs). Informed by our fine-grained complex span annotation, we find that adding a single feature, capturing the number of jargon spans, into existing readability formulas can significantly improve their correlation with human judgments, and also make them more stable. We will publicly release data and code.

15:15 - 15:30 - Tutte

MINT: A Benchmark for Evaluating Instructed Information Retrieval

Weiwei Sun, Zhengliang Shi, Wu Jiu Long, Lingyong Yan, Xinyu Ma, Yiding Liu, Min Cao, Dawei Yin, Zhaochun Ren

Recent information retrieval (IR) models are pre-trained and instruction-tuned on massive datasets and tasks, enabling them to perform well on a wide range of tasks and potentially generalize to unseen tasks with instructions. However, existing IR benchmarks focus on a limited scope of tasks, making them insufficient for evaluating the latest IR models. In this paper, we propose MAIR (Massive Instructed Retrieval Benchmark), a heterogeneous IR benchmark that includes 126 distinct IR tasks across 6 domains, collected from existing datasets. We benchmark state-of-the-art instruction-tuned text embedding models and re-ranking models. Our experiments reveal that instruction-tuned models generally achieve superior performance compared to non-instruction-tuned models on MAIR. Additionally, our results suggest that current instruction-tuned text embedding models and re-ranking models still lack effectiveness in specific long-tail tasks. MAIR is publicly available at <https://github.com/sunnweiwei/Mair>.

Session 04 - Nov 12 (Tue) 16:00-17:30

Ethics, Bias, and Fairness 2

Nov 12 (Tue) 16:00-17:30 - Room: Ashe Auditorium

16:00 - 16:15 - Ashe Auditorium

Adaptable Moral Stances of Large Language Models on Sexist Content: Implications for Society and Gender Discourse

Rongchen Guo, Isar Nejadgholi, Hillary Dawkins, Kathleen C. Fraser, Svetlana Kiritchenko

This work provides an explanatory view of how LLMs can apply moral reasoning to both criticize and defend sexist language. We assessed eight large language models, all of which demonstrated the capability to provide explanations grounded in varying moral perspectives for both critiquing and endorsing views that reflect sexist assumptions. With both human and automatic evaluation, we show that all eight models produce comprehensible and contextually relevant text, which is helpful in understanding diverse views on how sexism is perceived. Also, through analysis of moral foundations cited by LLMs in their arguments, we uncover the diverse ideological perspectives in models' outputs, with some models aligning more with progressive or conservative views on gender roles and sexism. Based on our observations, we caution against the potential misuse of LLMs to justify sexist language. We also highlight that LLMs can serve as tools for understanding the roots of sexist beliefs and designing well-informed interventions. Given this dual capacity, it is crucial to monitor LLMs and design safety mechanisms for their use in applications that involve sensitive societal topics, such as sexism.

16:15 - 16:30 - Ashe Auditorium

AgentReview: Exploring Peer Review Dynamics with LLM Agents

Yiqiao Jin, Qlinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, Jindong Wang

Peer review is fundamental to the integrity and advancement of scientific publication. Traditional methods of peer review analyses often rely on exploration and statistics of existing peer review data, which do not adequately address the multivariate nature of the process, account for the latent variables, and are further constrained by privacy concerns due to the sensitive nature of the data. We introduce AgentReview, the first large language model (LLM) based peer review simulation framework, which effectively disentangles the impacts of multiple latent factors and addresses the privacy issue. Our study reveals significant insights, including a notable 37.1% variation in paper decisions due to reviewers' biases, supported by sociological theories such as the social influence theory, altruism fatigue, and authority bias. We believe that this study could offer valuable insights to improve the design of peer review mechanisms.

16:30 - 16:45 - Ashe Auditorium

Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs

Tanise Ceron, Neele Falk, Ana Bari, Dmitry Nikolaev, Sebastian Padé

Due to the widespread use of large language models (LLMs), we need to understand whether they embed a specific worldview and what these views reflect. Recent studies report that, prompted with political questionnaires, LLMs show left-liberal leanings (Feng et al., 2023; Motoki et al., 2024). However, it is as yet unclear whether these leanings are reliable (robust to prompt variations) and whether the leaning is consistent across policies and political leaning. We propose a series of tests which assess the reliability and consistency of LLMs stances on political statements based on a dataset of voting-advice questionnaires collected from seven EU countries and annotated for policy issues. We study LLMs ranging in size from 7B to 70B parameters and find that their reliability increases with parameter count. Larger models show overall stronger alignment with left-leaning parties but differ among policy programs: They show a (left-wing) positive stance towards environment protection, social welfare state and liberal society but also (right-wing) law and order, with no consistent preferences in the areas of foreign policy and migration.

16:45 - 17:00 - Ashe Auditorium

Red Teaming Language Model Detectors with Language Models

Kai-Wei Chang, Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Cho-Jui Hsieh

The prevalence and strong capability of large language models (LLMs) present significant safety and ethical risks if exploited by malicious users. To prevent the potentially deceptive usage of LLMs, recent works have proposed algorithms to detect LLM-generated text and protect LLMs. In this paper, we investigate the robustness and reliability of these LLM detectors under adversarial attacks. We study two types of attack strategies: 1) replacing certain words in an LLM's output with their synonyms given the context; 2) automatically searching for an instructional prompt to alter the writing style of the generation. In both strategies, we leverage an auxiliary LLM to generate the word replacements or the instructional prompt. Different from previous works, we consider a challenging setting where the auxiliary LLM can also be protected by a detector. Experiments reveal that our attacks effectively compromise the performance of all detectors in the study with plausible generations, underscoring the urgent need to improve the robustness of LLM-generated text detection systems. Code is available at <https://github.com/shizhouxing/LLM-Detector-Robustness>.

17:00 - 17:15 - Ashe Auditorium

STOP! Benchmarking Large Language Models with Sensitivity Testing on Offensive Progressions

Robert Morabito, Sangmitra Madhusudan, Tyler McDonald, Ali Emami

Mitigating explicit and implicit biases in Large Language Models (LLMs) has become a critical focus in the field of natural language processing. However, many current methodologies evaluate scenarios in isolation, without considering the broader context or the spectrum of potential biases within each situation. To address this, we introduce the Sensitivity Testing on Offensive Progressions (STOP) dataset, which includes 450 offensive progressions containing 2,700 unique sentences of varying severity that progressively escalate from less to more explicitly offensive. Covering a broad spectrum of 9 demographics and 46 sub-demographics, STOP ensures inclusivity and comprehensive coverage. We evaluate several leading closed- and open-source models, including GPT-4, Mixtral, and Llama 3. Our findings reveal that even the best-performing models detect bias inconsistently, with success rates ranging from 19.3% to 69.8%. Furthermore, we demonstrate how aligning models with human judgments on STOP can improve model answer rates on sensitive tasks such as BBQ, StereoSet, and CrowsPairs by up to 191%, while maintaining or even improving performance. STOP presents a novel framework for assessing the complex nature of biases in LLMs, which will enable more effective bias mitigation strategies and facilitates the creation of fairer language models.

17:15 - 17:30 - Ashe Auditorium

Voices in a Crowd: Searching for clusters of unique perspectives

Nikolas Vitsakis, Amit Parekh, Ioannis Konstas

Language models have been shown to reproduce underlying biases existing in their training data, which is the majority perspective by default. Proposed solutions aim to capture minority perspectives by either modelling annotator disagreements or grouping annotators based on shared metadata, both of which face significant challenges. We propose a framework that trains models without encoding annotator metadata, extracts latent embeddings informed by annotator behaviour, and creates clusters of similar opinions, that we refer to as voices. Resulting clusters are validated post-hoc via internal and external quantitative metrics, as well as a qualitative analysis to identify the type of voice that each cluster represents. Our results demonstrate the strong generalisation capability of our framework, indicated by resulting clusters being adequately robust, while also capturing minority perspectives based on different demographic factors throughout two distinct datasets.

Information Retrieval and Text Mining 2

Nov 12 (Tue) 16:00-17:30 - Room: Brickell

16:00 - 16:15 - Brickell

Do You Know What You Are Talking About? Characterizing Query-Knowledge Relevance For Reliable Retrieval Augmented Generation

Zhuohang Li, Jiaxin Zhang, Chao Yan, Kamalika Das, Sricharan Kumar, Murat Kantarcioglu, Bradley A. Malin

Language models (LMs) are known to suffer from hallucinations and misinformation. Retrieval augmented generation (RAG) that retrieves verifiable information from an external knowledge corpus to complement the parametric knowledge in LMs provides a tangible solution to these problems. However, the generation quality of RAG is highly dependent on the relevance between a user's query and the retrieved documents. Inaccurate responses may be generated when the query is outside of the scope of knowledge represented in the external knowledge corpus or if the information in the corpus is out-of-date. In this work, we establish a statistical framework that assesses how well a query can be answered by an RAG system by capturing the relevance of knowledge. We introduce an online testing procedure that employs goodness-of-fit (GoF) tests to inspect the relevance of each user query to detect out-of-knowledge queries with low knowledge relevance. Additionally, we develop an offline testing framework that examines a collection of user queries, aiming to detect significant shifts in the query distribution which indicates the knowledge corpus is no longer sufficiently capable of supporting the interests of the users. We demonstrate the capabilities of these strategies through a systematic evaluation on eight question-answering (QA) datasets, the results of which indicate that the new testing framework is an efficient solution to enhance the reliability of existing RAG systems.

16:15 - 16:30 - Brickell

Improved Learned Sparse Retrieval with Entity Vocabulary

Thong Nguyen, Shubham Chatterjee, Sean MacAvaney, Iain Mackie, Jeff Dalton, Andrew Yates

Learned Sparse Retrieval (LSR) models use vocabularies from pre-trained transformers, which often split entities into nonsensical fragments. Splitting entities diminishes retrieval accuracy and limits the model's ability to incorporate up-to-date world knowledge not included in the training data. In this work, we enhance the LSR vocabulary with Wikipedia concepts and entities, enabling the model to resolve ambiguities more effectively and stay current with evolving knowledge. Central to our approach is a Dynamic Vocabulary (DyVo) head, which leverages existing entity embeddings and an entity retrieval component that identifies entities relevant to a query or document. We use the DyVo head to generate entity weights, which are then merged with word piece weights to create joint representations for efficient indexing and retrieval using an inverted index. In experiments across three entity-rich document ranking datasets, the resulting DyVo model substantially outperforms several state-of-the-art baselines.

16:30 - 16:45 - Brickell

Language Concept Erasure for Language-invariant Dense Retrieval

Zhiqi Huang, Puxuan Yu, Shauli Rayfogel, James Allan

Multilingual models aim for language-invariant representations but still prominently encode language identity. This, along with the scarcity of high-quality parallel retrieval data, limits their performance in retrieval. We introduce LANCER, a multi-task learning framework that improves language-invariant dense retrieval by reducing language-specific signals in the embedding space. Leveraging the notion of linear concept erasure, we design a loss function that penalizes cross-correlation between representations and their language labels. LANCER leverages only English retrieval data and general multilingual corpora, training models to focus on language-invariant retrieval by semantic similarity without necessitating a vast parallel corpus. Experimental results on various datasets show our method consistently improves over baselines, with extensive analyses demonstrating greater language agnosticism.

16:45 - 17:00 - Brickell

Taxonomy-guided Semantic Indexing for Academic Paper Search

SeongKu Kang, Yunyi Zhang, Pengcheng Jiang, Dongha Lee, Jiawei Han, Hwanjo Yu

Academic paper search is an essential task for efficient literature discovery and scientific advancement. While dense retrieval has advanced various ad-hoc searches, it often struggles to match the underlying academic concepts between queries and documents, which is critical for paper search. To enable effective academic concept matching for paper search, we propose Taxonomy-guided Semantic Indexing (TaxoIndex) framework. TaxoIndex extracts key concepts from papers and organizes them as a semantic index guided by an academic taxonomy, and then leverages this index as foundational knowledge to identify academic concepts and link queries and documents. As a plug-and-play framework, TaxoIndex can be flexibly employed to enhance existing dense retrievers. Extensive experiments show that TaxoIndex brings significant improvements, even with highly limited training data, and greatly enhances interpretability.

17:00 - 17:15 - Brickell

Threshold-driven Pruning with Segmented Maximum Term Weights for Approximate Cluster-based Sparse Retrieval

Yifan Qiao, Parker Carlson, Shanzix He, Yingnui Yang, Tao Yang

This paper revisits dynamic pruning through rank score thresholding in cluster-based sparse retrieval to skip the index partially at cluster and document levels during inference. It proposes a two-parameter pruning control scheme called ASC with a probabilistic guarantee on rank-safeness competitiveness. ASC uses cluster-level maximum weight segmentation to improve accuracy of rank score bound estimation and threshold-driven pruning, and is targeted for speeding up retrieval applications requiring high relevance competitiveness. The experiments with MS MARCO and BEIR show that ASC improves the accuracy and safeness of pruning for better relevance while delivering a low latency on a single-threaded CPU.

17:15 - 17:30 - Brickell

Toward Robust RALMs: Revealing the Impact of Imperfect Retrieval on Retrieval-Augmented Language Models

Jay Lee, Seongil Park

Retrieval-Augmented Language Models (RALMs) have gained significant attention for their ability to generate accurate answers and improve efficiency. However, RALMs are inherently vulnerable to imperfect information due to their reliance on the imperfect retriever or knowledge source. We identify three common scenarios—unanswerable, adversarial, conflicting—where retrieved document sets can confuse RALM with plausible real-world examples. We present the first comprehensive investigation to assess how well RALMs detect and handle such problematic scenarios. Among these scenarios, to systematically examine adversarial robustness we propose a new adversarial attack method Generative model-based ADVERSarial attack (GenADV) and a novel metric Robustness under Additional Document (RAD). Our findings reveal that RALMs often fail to identify the unanswerability or contradiction of a document set, which frequently leads to hallucinations. Moreover, we show the addition of an adversary significantly degrades RALMs' performance, with the model becoming even more vulnerable when the two scenarios overlap (adversarial+unanswerable). Our research identifies critical areas for assessing and enhancing the robustness

of RALMs, laying the foundation for the development of more robust models.

Multimodality and Language Grounding to Vision, Robotics and Beyond 2

Nov 12 (Tue) 16:00-17:30 - Room: Flagler

16:00 - 16:15 - Flagler

Altogether: Image Captioning via Re-aligning Alt-text

Hu Xu, Po-Yao Huang, Xiaoxing Tan, Ching-Feng Yeh, Jacob Kahn, Christine Jou, Gargi Ghosh, Omer Levy, Luke Zettlemoyer, Wen-tau Yih, Shang-Wen Li, Saining Xie, Christoph Feichtenhofer

This paper focuses on creating synthetic data to improve the quality of image captions. Existing works typically have two shortcomings. First, they caption images from scratch, ignoring existing alt-text metadata; and second, lack transparency if the captioners training data (e.g. GPT) is unknown. In this paper, we study a principled approach Altogether based on the key idea to edit and re-align existing alt-texts associated with the images. To generate training data, we perform human annotation where annotators start with the existing alt-text and re-align it to the image content in multiple rounds, consequently constructing captions with rich visual concepts. This differs from prior work that carries out human annotation as a one-time description task solely based on images and annotator knowledge. We train a captioner on this data that generalizes the process of re-aligning alt-texts at scale. Our results show our Altogether approach leads to richer image captions that also improve text-to-image generation and zero-shot image classification tasks.

16:15 - 16:30 - Flagler

Benchmarking Vision Language Models for Cultural Understanding

Shravan Nayak, Kanishk Jain, Rabiu Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, Aishwarya Agrawal
Foundation models and vision-language pre-training have notably advanced Vision Language Models (VLMs), enabling multimodal processing of visual and linguistic data. However, their performance has been typically assessed on general scene understanding - recognizing objects, attributes, and actions - rather than cultural comprehension. This study introduces CulturalVQA, a visual question-answering benchmark aimed at assessing VLM's geo-diverse cultural understanding. We curate a diverse collection of 2,378 image-question pairs with 1-5 answers per question representing cultures from 11 countries across 5 continents. The questions probe understanding of various facets of culture such as clothing, food, drinks, rituals, and traditions. Benchmarking VLMs on CulturalVQA, including GPT-4V and Gemini, reveals disparity in their level of cultural understanding across regions, with strong cultural understanding capabilities for North America while significantly weaker capabilities for Africa. We observe disparity in their performance across cultural facets too, with clothing, rituals, and traditions seeing higher performances than food and drink. These disparities help us identify areas where VLMs lack cultural understanding and demonstrate the potential of CulturalVQA as a comprehensive evaluation set for gauging VLM progress in understanding diverse cultures.

16:30 - 16:45 - Flagler

Do Vision and Language Models Share Concepts? A Vector Space Alignment Study

Anders Søgaard, Jiang Li Li, Yova Kementchedzhieva, Kementchedzhieva, Constanza Fierro
Large-scale pretrained language models (LMs) are said to "lack the ability to connect utterances to the world" (Bender and Koller, 2020), because they do not have "mental models of the world" (Mitchell and Krakauer, 2023). If so, one would expect LM representations to be unrelated to representations induced by vision models. We present an empirical evaluation across four families of LMs (BERT, GPT-2, OPT and LLaMA-2) and three vision model architectures (ResNet, SegFormer, and MAE). Our experiments show that LMs partially converge towards representations isomorphic to those of vision models, subject to dispersion, polysemy and frequency. This has important implications for both multi-modal processing and the LM understanding debate (Mitchell and Krakauer, 2023).

16:45 - 17:00 - Flagler

MatchTime: Towards Automatic Soccer Game Commentary Generation

Jiayuan Rao, Haoming Wu, Chang Liu, Yanfeng Wang, Weidi Xie

Soccer is a globally popular sport with a vast audience, in this paper, we consider constructing an automatic soccer game commentary model to improve the audiences' viewing experience. In general, we make the following contributions: *First*, observing the prevalent video-text misalignment in existing datasets, we manually annotate timestamps for 49 matches, establishing a more robust benchmark for soccer game commentary generation, termed as *SN-Caption-test-align*. *Second*, we propose a multi-modal temporal alignment pipeline to automatically correct and filter the existing dataset at scale, creating a higher-quality soccer game commentary dataset for training, denoted as *MatchTime*. *Third*, based on our curated dataset, we train an automatic commentary generation model, named **MatchVoice**. Extensive experiments and ablation studies have demonstrated the effectiveness of our alignment pipeline, and training model on the curated datasets achieves state-of-the-art performance for commentary generation, showcasing that better alignment can lead to significant performance improvements in downstream tasks.

17:00 - 17:15 - Flagler

Selective Vision is the Challenge for Visual Reasoning: A Benchmark for Visual Argument Understanding

Jiwon Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, Youngjae Yu

Visual arguments, often used in advertising or social causes, rely on images to persuade viewers to do or believe something. Understanding these arguments requires selective vision: only specific visual stimuli within an image are relevant to the argument, and relevance can only be understood within the context of a broader argumentative structure. While visual arguments are readily appreciated by human audiences, we ask: are today's AI capable of similar understanding? We present VisArgs, a dataset of 1,611 images annotated with 5,112 visual premises (with regions), 5,574 commonsense premises, and reasoning trees connecting them into structured arguments. We propose three tasks for evaluating visual argument understanding: premise localization, premise identification, and conclusion deduction. Experiments show that 1) machines struggle to capture visual cues: GPT-4-O achieved 78.5% accuracy, while humans reached 98.0%. Models also performed 19.5% worse when distinguishing between irrelevant objects within the image compared to external objects. 2) Providing relevant visual premises improved model performance significantly.

Multimodality and Language Grounding to Vision, Robotics and Beyond 1

Nov 12 (Tue) 16:00-17:30 - Room: Flagler

17:15 - 17:30 - Flagler

Pre-trained Language Models Do Not Help Auto-regressive Text-to-Image Generation*Yuhui Zhang, Brandon McKinzie, Zhe Gan, Vaishaal Shankar, Alexander T Toshev*

Recent advances in image tokenizers, such as VQ-VAE, have enabled text-to-image generation using auto-regressive methods, similar to language modeling. However, these methods have yet to leverage pre-trained language models, despite their adaptability to various downstream tasks. In this work, we explore this gap by adapting a pre-trained language model for auto-regressive text-to-image generation, and find that pre-trained language models offer limited help. We provide a two-fold explanation by analyzing tokens from each modality. First, we demonstrate that image tokens possess significantly different semantics compared to text tokens, rendering pre-trained language models no more effective in modeling them than randomly initialized ones. Second, the text tokens in the image-text datasets are too simple compared to normal language model pre-training data, which causes the catastrophic degradation of language models' capability.

Linguistic Theories, Cognitive Modeling and Psycholinguistics 2

Nov 12 (Tue) 16:00-17:30 - Room: Monroe

16:00 - 16:15 - Monroe

ARN: Analogical Reasoning on Narratives*Zhiwar Sourati, Filip Ilievski, Pia Sommerauer, Yifan Jiang*

As a core cognitive skill that enables the transferability of information across domains, analogical reasoning has been extensively studied for both humans and computational models. However, while cognitive theories of analogy often focus on narratives and study the distinction between surface, relational, and system similarities, existing work in natural language processing has a narrower focus as far as relational analogies between word pairs. This gap brings a natural question: can state-of-the-art large language models (LLMs) detect system analogies between narratives? To gain insight into this question and extend word-based relational analogies to relational system analogies, we devise a comprehensive computational framework that operationalizes dominant theories of analogy, using narrative elements to create surface and system mappings. Leveraging the interplay between these mappings, we create a binary task and benchmark for Analogical Reasoning on Narratives (ARN), covering four categories of far (cross-domain)/near (within-domain) analogies and disanalogies. We show that while all LLMs can largely recognize near analogies, even the largest ones struggle with far analogies in a zero-shot setting, with GPT4.0 scoring below random. Guiding the models through solved examples and Chain-of-Thought reasoning enhances their analogical reasoning ability. Yet, since even in the few-shot setting, the best model only performs halfway between random and humans, ARN opens exciting directions for computational analogical reasoners. à

16:15 - 16:30 - Monroe

Decoding the Echoes of Vision from fMRI: Memory Disentangling for Past Semantic Information*Runze Xia, Congchi Yin, Piji Li*

The human visual system is capable of processing continuous streams of visual information, but how the brain encodes and retrieves recent visual memories during continuous visual processing remains unexplored. This study investigates the capacity of working memory to retain past information under continuous visual stimuli. And then we propose a new task **Memory Disentangling**, which aims to extract and decode past information from fMRI signals. To address the issue of interference from past memory information, we design a disentangled contrastive learning method inspired by the phenomenon of proactive interference. This method separates the information between adjacent fMRI signals into current and past components and decodes them into image descriptions. Experimental results demonstrate that this method effectively disentangles the information within fMRI signals. This research could advance brain-computer interfaces and mitigate the problem of low temporal resolution in fMRI.

16:30 - 16:45 - Monroe

Do Language Models Words Refer?*Matthew Mandelkern, Tal Linzen*

What do language models (LMs) do with language? They can produce sequences of (mostly) coherent strings closely resembling English. But do those sentences mean something, or are LMs simply babbling in a convincing simulacrum of language use? We address one aspect of this broad question: whether LMs words can refer, that is, achieve word-to-world connections. There is *prima facie* reason to think they do not, since LMs do not interact with the world in the way that ordinary language users do. Drawing on the externalist tradition in philosophy of language, we argue that those appearances are misleading: Even if the inputs to LMs are simply strings of text, they are strings of text with natural histories, and that may suffice for LMs words to refer.

16:45 - 17:00 - Monroe

Fine-Grained Prediction of Reading Comprehension from Eye Movements*Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, Yevgeni Berzak*

Can human reading comprehension be assessed from eye movements in reading? In this work, we address this longstanding question using large-scale eyetracking data. We focus on a cardinal and largely unaddressed variant of this question: predicting reading comprehension of a single participant for a single question from their eye movements over a single paragraph. We tackle this task using a battery of recent models from the literature, and three new multimodal language models. We evaluate the models in two different reading regimes: ordinary reading and information seeking, and examine their generalization to new textual items, new participants, and the combination of both. The evaluations suggest that the task is highly challenging, and highlight the importance of benchmarking against a strong text-only baseline. While in some cases eye movements provide improvements over such a baseline, they tend to be small. This could be due to limitations of current modelling approaches, limitations of the data, or because eye movement behavior does not sufficiently pertain to fine-grained aspects of reading comprehension processes. Our study provides an infrastructure for making further progress on this question.

17:00 - 17:15 - Monroe

Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs*Kanishka Misra, Kyle Mahowald*

Language models learn rare syntactic phenomena, but the extent to which this is attributable to generalization vs. memorization is a major open question. To that end, we iteratively trained transformer language models on systematically manipulated corpora which were human-scale in size, and then evaluated their learning of a rare grammatical phenomenon: the English Article+Adjective+Numerical+Noun (AANN) construction ("a beautiful five days"). We compared how well this construction was learned on the default corpus relative to a counterfactual corpus in which AANN sentences were removed. We found that AANNs were still learned better than systematically perturbed variants of the construction. Using additional counterfactual corpora, we suggest that this learning occurs through generalization from related constructions (e.g., "a few days"). An additional experiment showed that this learning is enhanced when there is more variability in the input. Taken

together, our results provide an existence proof that LMs can learn rare grammatical phenomena by generalization from less rare phenomena. Data and code: <https://github.com/kanishkamisra/aannalysis>.

17:15 - 17:30 - Monroe

Can Language Models Induce Grammatical Knowledge from Indirect Evidence?

Miyu Oba, Yonei Oseki, Akiyo Fukatsu, Akari Haga, Hiroki Ouchi, Taro Watanabe, Saku Sugawara

What kinds of and how much data is necessary for language models to induce grammatical knowledge to judge sentence acceptability? Recent language models still have much room for improvement in their data efficiency compared to humans. This paper investigates whether language models efficiently use indirect data (indirect evidence), from which they infer sentence acceptability. In contrast, humans use indirect evidence efficiently, which is considered one of the inductive biases contributing to efficient language acquisition. To explore this question, we introduce the Wug InDirect Evidence Test (WIDET), a dataset consisting of training instances inserted into the pre-training data and evaluation instances. We inject synthetic instances with newly coined wug words into pretraining data and explore the model's behavior on evaluation data that assesses grammatical acceptability regarding those words. We prepare the injected instances by varying their levels of indirectness and quantity. Our experiments surprisingly show that language models do not induce grammatical knowledge even after repeated exposure to instances with the same structure but differing only in lexical items from evaluation instances in certain language phenomena. Our findings suggest a potential direction for future research: developing models that use latent indirect evidence to induce grammatical knowledge.

Industry

Nov 12 (Tue) 16:00-17:30 - Room: Tuttle

16:00 - 16:10 - Tuttle

IPL: Leveraging Multimodal Large Language Models for Intelligent Product Listing

Chen Kang, Qing Heng Zhang, Chenghuolian, Yixin Ji, Xiuwei Liu, Shuguang Han, Guoqiang Wu, Fei Huang, Jufeng Chen

Unlike professional Business-to-Consumer (B2C) e-commerce platforms (e.g., Amazon), Consumer-to-Consumer (C2C) platforms (e.g., Facebook marketplace) are mainly targeting individual sellers who usually lack sufficient experience in e-commerce. Individual sellers often struggle to compose proper descriptions for selling products. With the recent advancement of Multimodal Large Language Models (MLLMs), we attempt to integrate such state-of-the-art generative AI technologies into the product listing process. To this end, we develop IPL, an Intelligent Product Listing tool tailored to generate descriptions using various product attributes such as category, brand, color, condition, etc. IPL enables users to compose product descriptions by merely uploading photos of the selling product. More importantly, it can imitate the content style of our C2C platform XXXX. This is achieved by employing domain-specific instruction tuning on MLLMs, and by adopting the multi-modal Retrieval-Augmented Generation (RAG) process. A comprehensive empirical evaluation demonstrates that the underlying model of IPL significantly outperforms the base model in domain-specific tasks while producing less hallucination. IPL has been successfully deployed in our production system, where 72% of users have their published product listings based on the generated content, and those product listings are shown to have a quality score 5.6% higher than those without AI assistance.

16:10 - 16:20 - Tuttle

Structured Object Language Modeling (SO-LM): Native Structured Objects Generation Conforming to Complex Schemas with Self-Supervised Denoising

Amir Tavanei, Kee Kiat Koo, Hayreddin Ceker, Shaobai Jiang, Qi Li, Julien Han, Karim Bouyarmene

In this paper, we study the problem of generating structured objects that conform to a complex schema, with intricate dependencies between the different components (facets) of the object. The facets of the object (attributes, fields, columns, properties) can be a mix of short, structured facts, or long natural-language descriptions. The object has to be self-consistent between the different facets in the redundant information it carries (relative consistency), while being grounded with respect to world knowledge (absolute consistency). We frame the problem as a Language Modeling problem (Structured Object Language Modeling) and train an LLM to perform the task natively, without requiring instructions or prompt-engineering. We propose a self-supervised denoising method to train the model from an existing dataset of such objects. The input query can be the existing object itself, in which case the system acts as a regenerator, completing, correcting, normalizing the input, or any unstructured blurb to be structured. We show that the self-supervised denoising training provides a strong baseline, and that additional supervised fine-tuning with small amount of human demonstrations leads to further improvement. Experimental results show that the proposed method matches or outperforms prompt-engineered general-purpose state-of-the-art LLMs (Claude 3, Mixtral-8x7B), while being order-of-magnitude more cost-efficient.

16:20 - 16:30 - Tuttle

MARS: Multilingual Aspect-centric Review Summarisation

Sandeep Sricharan Mukku, Abinesh Kanagarajan, Chetan Aggarwal, Promod Yenigalla

Summarizing customer feedback to provide actionable insights for products/services at scale is an important problem for businesses across industries. Lately, the reviews are spreading across multiple languages, the challenge of aggregating and understanding customer sentiment across multiple languages becomes increasingly vital. In this paper, we propose a novel framework involving a two-step paradigm *Extract-then-Summarise*, namely MARS to revolutionise traditions and address the domain agnostic aspect-level multilingual review summarisation. Extensive automatic and human evaluation shows that our approach brings substantial improvements over abstractive baselines and efficiency to production systems.

16:30 - 16:40 - Tuttle

Two-tiered Encoder-based Hallucination Detection for Retrieval-Augmented Generation in the Wild

Ilana Zimmerman, Jadin Tredup, Ethan Selfridge, Joseph Bradley

Detecting hallucinations, where Large Language Models (LLMs) are not factually consistent with a Knowledge Base (KB), is a challenge for Retrieval-Augmented Generation (RAG) systems. Current solutions rely on public datasets to develop prompts or fine-tune a Natural Language Inference (NLI) model. However, these approaches are not focused on developing an enterprise RAG system; they do not consider latency, train or evaluate on production data, nor do they handle non-verifiable statements such as small talk or questions. To address this, we leverage the customer service conversation data of four large brands to evaluate existing solutions and propose a set of small encoder models trained on a new dataset. We find the proposed models to outperform existing methods and highlight the value of combining a small amount of in-domain data with public datasets.

16:40 - 16:50 - Tuttle

Granite-Function Calling Model: Introducing Function Calling Abilities via Multi-task Learning of Granular Tasks

Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matthew Stallone, Rameswar Panda, Yara Rizk, G P Shrivatsa Bhargav, Maxwell Crouse, Chulaka Gunasekara, Shajith Ikbal, Sachindra Joshi, Hima Karanam, Vineet Kumar, Asim Munawar, Sumit Neelam, Dinesh Raghu, Udit Sharma, Adriana Meza Soria, Dheeraj Sreedhar, Praveen Venkateswaran, Merve Unuvar, David Daniel Cox, Salim Roukos, Luis A. Lastras, Pavan Kapanipathi

An emergent research trend explores the use of Large Language Models (LLMs) as the backbone of agentic systems (e.g., SWE-Bench, Agent-Bench). To fulfill LLMs' potential as autonomous agents, they must be able to identify, call, and interact with a variety of external tools and application program interfaces (APIs). This capability of LLMs, commonly termed function calling, leads to a myriad of advantages such as access to current and domain-specific information in databases and the outsourcing of tasks that can be reliably performed by tools. In this work, we introduce GRINDER-20B-FunctionCalling, a model trained using a multi-task training approach on seven fundamental tasks encompassed in function calling. Our comprehensive evaluation on multiple out-of-domain datasets, which compares GRINDER-20B-FunctionCalling to more than 15 other best proprietary and open models, shows that GRINDER-20B-FunctionCalling has better generalizability on multiple tasks across seven different evaluation benchmarks. Moreover, GRINDER-20B-FunctionCalling shows the best performance among all open models and ranks among the top on the Berkeley Function Calling Leaderboard (BFCL).

16:50 - 17:00 - Tuttle

Provance: A Light-weight Fact-checker for Retrieval Augmented LLM Generation Output

Mohammed Nasheed Yasin, Hithesh Sankararaman, Andreas Stöckle

We present a light-weight approach for detecting nonfactual outputs from retrieval-augmented generation (RAG). Given a context and putative output, we compute a factuality score that can be thresholded to yield a binary decision to check the results of LLM-based question-answering, summarization, or other systems. Unlike factuality checkers that themselves rely on LLMs, we use compact, open-source natural language inference (NLI) models that yield a freely accessible solution with low latency and low cost at run-time, and no need for LLM fine-tuning. The approach also enables downstream mitigation and correction of hallucinations, by tracing them back to specific context chunks. Our experiments show high ROC-AUC across a wide range of relevant open source datasets, indicating the effectiveness of our method for fact-checking RAG output.

17:00 - 17:10 - Tuttle

PEARL: Preference Extraction with Exemplar Augmentation and Retrieval with LLM Agents

Vijit Malik, Akshay Jagatap, Vinayak S Puranik, Anirban Majumder

Identifying preferences of customers in their shopping journey is a pivotal aspect in providing product recommendations. The task becomes increasingly challenging when there is a multi-turn conversation between the user and a shopping assistant chatbot. In this paper, we tackle a novel and complex problem of identifying customer preferences in the form of key-value filters on an e-commerce website in a multi-turn conversational setting. Existing systems specialize in extracting customer preferences from standalone customer queries which makes them unsuitable to multi-turn setup. We propose PEARL (Preference Extraction with ICL Augmentation and Retrieval with LLM Agents) that leverages collaborative LLM agents, generates in-context learning exemplars and dynamically retrieves relevant exemplars during inference time to extract customer preferences as a combination of key-value filters. Our experiments on proprietary and public datasets show that PEARL not only improves performance on exact match by 10% compared to competitive LLM-based baselines but additionally improves inference latency by 110%.

17:10 - 17:20 - Tuttle

RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, Cheng Niu

Retrieval-augmented generation (RAG) has emerged as a significant advancement in the field of large language models (LLMs). By integrating up-to-date information not available during their initial training, RAG greatly enhances the practical utility of LLMs in real-world applications. However, even with RAG, LLMs can still produce inaccurate outputs, such as distorting or misinterpreting source content, posing risks in high-trust scenarios. To address these issues, we introduce a novel approach called Hallucinate Aware Tuning (HAT). This method involves training hallucination detection models that generate detection labels and provide detailed descriptions of the detected hallucinations. Utilizing these detection results—particularly the hallucination descriptions—GPT-4 Turbo is employed to correct any detected hallucinations. The corrected outputs, free of hallucinations, along with the original versions, are used to create a preference dataset for Direct Preference Optimization (DPO) training. The fine-tuning through DPO leads to LLMs that exhibit a reduced rate of hallucinations and deliver improved answer quality.

17:20 - 17:30 - Tuttle

Divide-Conquer-Reasoning for Consistency Evaluation and Automatic Improvement of Large Language Models

Wendi Cui, Jiaxin Zhang, Zhiuhang Li, Damien Lopez, Kamalika Das, Bradley A. Malin, Srisharan Kumar

Evaluating the quality and consistency of text generated by Large Language Models (LLMs) poses a significant, yet unresolved challenge for industry research. We propose an automated framework for evaluating and improving the consistency of LLM-generated texts using a divide-conquer-reasoning approach. Unlike existing LLM-based evaluators operating at the paragraph level, our method employs a divide-and-conquer evaluator that breaks down the paragraph-to-paragraph comparison into sentence-to-paragraph comparisons. To facilitate this approach, we also introduce an automatic metric converter that translates the output into an interpretable numeric score. Beyond the consistency evaluation, we further present a reason-assisted improver that mitigates inconsistencies by leveraging the analytical reasons identified by divide-conquer. Through comprehensive and systematic empirical analysis, we show that our approach outperforms state-of-the-art methods by a large margin (e.g., +16.8% and +32.5% on the SummEval dataset) in consistency evaluation across multiple benchmarks. Our approach also substantially reduces nearly 90% output inconsistencies in one iteration, showing promise for effective hallucination mitigation in real-world industrial applications.

Session 06 - Nov 13 (Wed) 10:30-12:00

Multimodality and Language Grounding to Vision, Robotics and Beyond 3

Nov 13 (Wed) 10:30-12:00 - Room: Ashe Auditorium

10:30 - 10:45 - Ashe Auditorium

Do Multimodal Large Language Models and Humans Ground Language Similarly?

Cameron Jones, Benjamin Bergen, Sean Trott

Oral Presentations

Large Language Models (LLMs) have been criticized for failing to connect linguistic meaning to the world—*for failing to solve the "symbol grounding problem."* Multimodal Large Language Models (MLLMs) offer a potential solution to this challenge by combining linguistic representations and processing with other modalities. However, much is still unknown about exactly how and to what degree MLLMs integrate their distinct modalities—and whether the way they do so mirrors the mechanisms believed to underpin grounding in humans. In humans, it has been hypothesized that linguistic meaning is grounded through "embodied simulation," the activation of sensorimotor and affective representations reflecting described experiences. Across four pre-registered studies, we adapt experimental techniques originally developed to investigate embodied simulation in human comprehenders to ask whether MLLMs are sensitive to sensorimotor features that are implied but not explicit in descriptions of an event. In Experiment 1, we find sensitivity to some features (color and shape) and not others (size, orientation, and volume). In Experiment 2, we identify likely bottlenecks to explain an MLLM's lack of sensitivity. In Experiment 3, we find that despite sensitivity to implicit sensorimotor features, MLLMs cannot fully account for human behavior on the same task. Finally, in Experiment 4, we compare the psychometric predictive power of different MLLM architectures and find that ViLT, a single-stream architecture, is more predictive of human performance than CLIP, a dual-encoder architecture—*despite being trained on orders of magnitude less data.* These results reveal strengths and limitations in the ability of current MLLMs to integrate language with other modalities, and also shed light on the likely mechanisms underlying human language comprehension.

10:45 - 11:00 - Ashe Auditorium

GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha

Perceiving and understanding non-speech sounds and non-verbal speech is essential to making decisions that help us interact with our surroundings. In this paper, we propose GAMA, a novel General-purpose Large Audio-Language Model (LALM) with Advanced Audio Understanding and Complex Reasoning Abilities. We build GAMA by integrating an LLM with multiple types of audio representations, including features from a custom Audio Q-Former, a multi-layer aggregator that aggregates features from multiple layers of an audio encoder. We fine-tune GAMA on a large-scale audio-language dataset, which augments it with audio understanding capabilities. Next, we propose CompA-R (Instruction-Tuning for Complex Audio Reasoning), a synthetically generated instruction-tuning (IT) dataset with instructions that require the model to perform complex reasoning on the input audio. We instruction-tune GAMA with CompA-R to endow it with complex reasoning abilities, where we further add a soft prompt as input with high-level semantic evidence by leveraging event tags of the input audio. Finally, we also propose CompA-R-test, a human-labeled evaluation dataset for evaluating the capabilities of LALMs on open-ended audio question-answering that requires complex reasoning. Through automated and expert human evaluations, we show that GAMA outperforms all other LALMs in literature on diverse audio understanding tasks by margins of 1%-84% and demonstrates state-of-the-art performance on deductive reasoning and hallucination evaluation benchmarks. Further, GAMA IT-ed on CompA-R proves to be superior in its complex reasoning capabilities.

11:00 - 11:15 - Ashe Auditorium

Investigating and Mitigating Object Hallucinations in Pretrained Vision-Language (CLIP) Models

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, Aimin Zhou

Large Vision-Language Models (LVLMs) have achieved impressive performance, yet research has pointed out a serious issue with object hallucinations within these models. However, there is no clear conclusion as to which part of the model these hallucinations originate from. In this paper, we present an in-depth investigation into the object hallucination problem specifically within the CLIP model, which serves as the backbone for many state-of-the-art vision-language systems. We unveil that even in isolation, the CLIP model is prone to object hallucinations, suggesting that the hallucination problem is not solely due to the interaction between vision and language modalities. To address this, we propose a counterfactual data augmentation method by creating negative samples with a variety of hallucination issues. We demonstrate that our method can effectively mitigate object hallucinations for CLIP model, and we show the enhanced model can be employed as a visual encoder, effectively alleviating the object hallucination issue in LVLMs.

11:15 - 11:30 - Ashe Auditorium

Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, Yueting Zhuang

Although most current large multimodal models (LMMs) can already understand photos of natural scenes and portraits, their understanding of abstract images, e.g., charts, maps, or layouts, and visual reasoning capabilities remains quite rudimentary. They often struggle with simple daily tasks, such as reading time from a clock, understanding a flowchart, or planning a route using a road map. In light of this, we design a multi-modal self-instruct, utilizing large language models and their code capabilities to synthesize massive abstract images and visual reasoning instructions across daily scenarios. Our strategy effortlessly creates a multimodal benchmark with 11,193 instructions for eight visual scenarios: charts, tables, simulated maps, dashboards, flowcharts, relation graphs, floor plans, and visual puzzles. **This benchmark, constructed with simple lines and geometric elements, exposes the shortcomings of most advanced LMMs like GPT-4V and Llava in abstract image understanding, spatial relations reasoning, and visual element induction.** Besides, to verify the quality of our synthetic data, we fine-tune an LMM using 62,476 synthetic chart, table and road map instructions. The results demonstrate improved chart understanding and map navigation performance, and also demonstrate potential benefits for other visual reasoning tasks.

11:30 - 11:45 - Ashe Auditorium

Preserving Multi-Modal Capabilities of Pre-trained VLMs for Improving Vision-Linguistic Compositionality

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, Junmo Kim

In this paper, we propose a new method to enhance compositional understanding in pre-trained vision and language models (VLMs) without sacrificing performance in zero-shot multi-modal tasks. Traditional fine-tuning approaches often improve compositional reasoning at the cost of degrading multi-modal capabilities, primarily due to the use of global hard negative (HN) loss, which contrasts global representations of images and texts. This global HN loss pushes HN texts that are highly similar to the original ones, damaging the models multi-modal representations. To overcome this limitation, we propose Fine-grained Selective Calibrated CLIP (FSC-CLIP), which integrates local hard negative loss and selective calibrated regularization. These innovations provide fine-grained negative supervision while preserving the models representational integrity. Our extensive evaluations across diverse benchmarks for both compositionality and multi-modal tasks show that FSC-CLIP not only achieves compositionality on par with state-of-the-art models but also retains strong multi-modal capabilities. Code is available at: <https://github.com/ytaekoh/fsc-clip>.

11:45 - 12:00 - Ashe Auditorium

TopViewRS: Vision-Language Models as Top-View Spatial Reasoners

Chengzu Li, Caige Zhang, Han Zhou, Nigel Collier, Anna Korhonen, Ivan Vuli

Top-view perspective denotes a typical way in which humans read and reason over different types of maps, and it is vital for localization and navigation of humans as well as of 'non-human' agents, such as the ones backed by large Vision-Language Models (VLMs). Nonetheless,

spatial reasoning capabilities of modern VLMs in this setup remain unattested and underexplored. In this work, we study their capability to understand and reason over spatial relations from the top view. The focus on top view also enables controlled evaluations at different granularity of spatial reasoning; we clearly disentangle different abilities (e.g., recognizing particular objects versus understanding their relative positions). We introduce the TopViewRS (Top-View Reasoning in Space) dataset, consisting of 11,384 multiple-choice questions with either realistic or semantic top-view map as visual input. We then use it to study and evaluate VLMs across 4 perception and reasoning tasks with different levels of complexity. Evaluation of 10 representative open- and closed-source VLMs reveals the gap of more than 50% compared to average human performance, and it is even lower than the random baseline in some cases. Although additional experiments show that Chain-of-Thought reasoning can boost model capabilities by 5.82% on average, the overall performance of VLMs remains limited. Our findings underscore the critical need for enhanced model capability in top-view spatial reasoning and set a foundation for further research towards human-level proficiency of VLMs in real-world multimodal tasks.

Ethics, Bias, and Fairness 3

Nov 13 (Wed) 10:30-12:00 - Room: Brickell

10:30 - 10:45 - Brickell

BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability

Mamta Mamtia, Rishikanti Chirgupatti, Asif Ekbal

Toxic content detection plays a vital role in addressing the misuse of social media platforms to harm people or groups due to their race, gender or ethnicity. However, due to the nature of the datasets, systems develop unintended bias due to the over-generalization of the model to the training data. This compromises the fairness of the systems, which can impact certain groups due to their race, gender, etc. Existing methods mitigate bias using data augmentation, adversarial learning, etc., which require re-training and adding extra parameters to the model. In this work, we present a robust and generalizable technique *BiasWipe* to mitigate unintended bias in language models. *BiasWipe* utilizes model interpretability using Shapley values, which achieve fairness by pruning the neuron weights responsible for unintended bias. It first identifies the neuron weights responsible for unintended bias and then achieves fairness by pruning them without loss of original performance. It does not require re-training or adding extra parameters to the model. To show the effectiveness of our proposed technique for bias unlearning, we perform extensive experiments for Toxic content detection for BERT, RoBERTa, and GPT models.¹

10:45 - 11:00 - Brickell

Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP

Pieter Delobele, Giuseppe Attanasio, Debora Nozza, Siu Lin Blodgett, Zeerak Talat

This paper introduces the concept of actionability in the context of bias measures in natural language processing (NLP). We define actionability as the degree to which a measure's results enable informed action and propose a set of desiderata for assessing it. Building on existing frameworks such as measurement modeling, we argue that actionability is a crucial aspect of bias measures that has been largely overlooked in the literature. We conduct a comprehensive review of 146 papers proposing bias measures in NLP, examining whether and how they provide the information required for actionable results. Our findings reveal that many key elements of actionability, including a measure's intended use and reliability assessment, are often unclear or entirely absent. This study highlights a significant gap in the current approach to developing and reporting bias measures in NLP. We argue that this lack of clarity may impede the effective implementation and utilization of these measures. To address this issue, we offer recommendations for more comprehensive and actionable metric development and reporting practices in NLP bias research.

11:00 - 11:15 - Brickell

Private Language Models via Truncated Laplacian Mechanism

Tianhao Huang, Tao Yang, Ivan Habernal, Lijie Hu, Di Wang

Recently it has been shown that deep learning models for NLP tasks are prone to attacks that can even reconstruct the verbatim training texts. To prevent privacy leakage, researchers have investigated word-level perturbations, relying on the formal guarantees of differential privacy (DP) in the embedding space. However, many existing approaches either achieve unsatisfactory performance in the high privacy regime when using the Laplacian or Gaussian mechanism, or resort to weaker relaxations of DP that are inferior to the canonical DP in terms of privacy strength. This raises the question of whether a new method for private word embedding can be designed to overcome these limitations. In this paper, we propose a novel private embedding method called the high dimensional truncated Laplacian mechanism. Specifically, we introduce a non-trivial extension of the truncated Laplacian mechanism, which was previously only investigated in one-dimensional space cases. Theoretically, we show that our method has a lower variance compared to the previous private word embedding methods. To further validate its effectiveness, we conduct comprehensive experiments on private embedding and downstream tasks using three datasets. Remarkably, even in the high privacy regime, our approach only incurs a slight decrease in utility compared to the non-private scenario.

11:15 - 11:30 - Brickell

Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, Dietrich Klakow

Robust, faithful and harm-free pronoun use for individuals is an important goal for language models as their use increases, but prior work tends to study only one or two of these characteristics at a time. To measure progress towards the combined goal, we introduce the task of pronoun fidelity: given a context introducing a co-referring entity and pronoun, the task is to reuse the correct pronoun later. We present RUFF, a carefully-designed dataset of over 5 million instances to measure robust pronoun fidelity in English, and we evaluate 37 popular large language models across architectures (encoder-only, decoder-only and encoder-decoder) and scales (11M-70B parameters). When an individual is introduced with a pronoun, models can mostly faithfully reuse this pronoun in the next sentence, but they are significantly worse with she/her/her, singular they and neopronouns. Moreover, models are easily distracted by non-adversarial sentences discussing other people; even one additional sentence with a distractor pronoun causes accuracy to drop on average by 34%. Our results show that pronoun fidelity is neither robust, nor due to reasoning, in a simple, naturalistic setting where humans achieve nearly 100% accuracy. We encourage researchers to bridge the gaps we find and to carefully evaluate reasoning in settings where superficial repetition might inflate perceptions of model performance.

11:30 - 11:45 - Brickell

Social Bias Probing: Fairness Benchmarking for Language Models

¹Code is available on <https://www.iitp.ac.in/~ai-nlp-ml/resources.html> and at the GitHub repository: <https://github.com/20118/BiasWipe>

Oral Presentations

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, Isabelle Augenstein

While the impact of social biases in language models has been recognized, prior methods for bias evaluation have been limited to binary association tests on small datasets, limiting our understanding of bias complexities. This paper proposes a novel framework for probing language models for social biases by assessing disparate treatment, which involves treating individuals differently according to their affiliation with a sensitive demographic group. We curate SoFa, a large-scale benchmark designed to address the limitations of existing fairness collections. SoFa expands the analysis beyond the binary comparison of stereotypical versus anti-stereotypical identities to include a diverse range of identities and stereotypes. Comparing our methodology with existing benchmarks, we reveal that biases within language models are more nuanced than acknowledged, indicating a broader scope of encoded biases than previously recognized. Benchmarking LMs on SoFa, we expose how identities expressing different religions lead to the most pronounced disparate treatments across all models. Finally, our findings indicate that real-life adversities faced by various groups such as women and people with disabilities are mirrored in the behavior of these models.

11:45 - 12:00 - Brickell

XDetox: Text Detoxification with Token-Level Toxicity Explanations

Beomseok Lee, Hyunwoo Kim, Keon Kim, Yong Suk Choi

Methods for mitigating toxic content through masking and infilling often overlook the decision-making process, leading to either insufficient or excessive modifications of toxic tokens. To address this challenge, we propose XDetox, a novel method that integrates token-level toxicity explanations with the masking and infilling detoxification process. We utilized this approach with two strategies to enhance the performance of detoxification. First, identifying toxic tokens to improve the quality of masking. Second, selecting the regenerated sentence by re-ranking the least toxic sentence among candidates. Our experimental results show state-of-the-art performance across four datasets compared to existing detoxification methods. Furthermore, human evaluations indicate that our method outperforms baselines in both fluency and toxicity reduction. These results demonstrate the effectiveness of our method in text detoxification.

Discourse + Phonology + Syntax 2

Nov 13 (Wed) 10:30-12:00 - Room: Flagler

10:30 - 10:45 - Flagler

eRST: A Sgnaled Graph Theory of Discourse Relations and Organization

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, Luke Gessler

In this article we present Enhanced Rhetorical Structure Theory (eRST), a new theoretical framework for computational discourse analysis, based on an expansion of Rhetorical Structure Theory (RST). The framework encompasses discourse relation graphs with tree-breaking, non-projective and concurrent relations, as well as implicit and explicit signals which give explainable rationales to our analyses. We survey shortcomings of RST and other existing frameworks, such as Segmented Discourse Representation Theory (SDRT), the Penn Discourse Treebank (PDTB) and Discourse Dependencies, and address these using constructs in the proposed theory. We provide annotation, search and visualization tools for data, and present and evaluate a freely available corpus of English annotated according to our framework, encompassing 12 spoken and written genres with over 200K tokens. Finally, we discuss automatic parsing, evaluation metrics and applications for data in our framework.

10:45 - 11:00 - Flagler

Are Large Language Models Capable of Generating Human-Level Narratives?

Yufei Tian, Tenghuo Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhan Chen, Jonathan May, Nanyun Peng

As daily reliance on large language models (LLMs) grows, assessing their generation quality is crucial to understanding how they might impact on our communications. This paper investigates the capability of LLMs in storytelling, focusing on narrative development and plot progression. We introduce a novel computational framework to analyze narratives through three discourse-level aspects: i) story arcs, ii) turning points, and iii) affective dimensions, including arousal and valence. By leveraging expert and automatic annotations, we uncover significant discrepancies between the LLM- and human-written stories. While human-written stories are suspenseful, arousing, and diverse in narrative structures, LLM stories are homogeneously positive and lack tension. Next, we measure narrative reasoning skills as a precursor to generative capacities, concluding that most LLMs fall short of human abilities in discourse understanding. Finally, we show that explicit integration of aforementioned discourse features can enhance storytelling, as is demonstrated by over 40% improvement in neural storytelling in terms of diversity, suspense, and arousal. Such advances promise to facilitate greater and more natural roles LLMs in human communication.

11:00 - 11:15 - Flagler

Revisiting Supertagging for faster HPSG parsing

Olga Zamaraeva, Carlos Gómez-Rodríguez

We present new supertaggers trained on English HPSG-based treebanks and test the effects of the best tagger on parsing speed and accuracy. HPSG treebanks are produced automatically by large manually built grammars and feature high-quality annotation based on a well-developed linguistic theory. The English Resource Grammar treebanks include diverse and challenging test datasets, beyond the usual WSJ section 23 and Wikipedia data. HPSG supertagging has previously relied on MaxEnt-based models. We use SVM and neural CRF- and BERT-based methods and show that both SVM and neural supertaggers achieve considerably higher accuracy compared to the baseline and lead to an increase not only in the parsing speed but also the parser accuracy with respect to gold dependency structures. Our fine-tuned BERT-based tagger achieves 97.26% accuracy on 950 sentences from WSJ23 and 93.88% on the out-of-domain technical essay The Cathedral and the Bazaar. We present experiments with integrating the best supertagger into an HPSG parser and observe a speedup of a factor of 3 with respect to the system which uses no tagging at all, as well as large recall gains and an overall precision gain. We also compare our system to an existing integrated tagger and show that although the well-integrated tagger remains the fastest, our experimental system can be more accurate. Finally, we hope that the diverse and difficult datasets we used for evaluation will gain more popularity in the field; we show that results can differ depending on the dataset, even if it is an in-domain one. We contribute the complete datasets reformatted for Huggingface token classification.

11:15 - 11:30 - Flagler

Which questions should I answer? Salience Prediction of Inquisitive Questions

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, Junyi Jessy Li

Inquisitive questions — open-ended, curiosity-driven questions people ask as they read — are an integral part of discourse processing and comprehension. Recent work in NLP has taken advantage of question generation capabilities of LLMs to enhance a wide range of applications. But the space of inquisitive questions is vast: many questions can be evoked from a given context. So which of those should be prioritized to find answers? Linguistic theories, unfortunately, have not yet provided an answer to this question. This paper presents QSsalience, a salience

predictor of inquisitive questions. QSalience is instruction-tuned over our dataset of linguist-annotated salience scores of 1,766 (context, question) pairs. A question scores high on salience if answering it would greatly enhance the understanding of the text. We show that highly salient questions are empirically more likely to be answered in the same article, bridging potential questions with Questions Under Discussion. We further validate our findings by showing that answering salient questions is an indicator of summarization quality in news.

11:30 - 11:45 - Flagler

SSL: Contrastive Self-Supervised Learning for Dependency Parsing on Relatively Free Word Ordered and Morphologically Rich Low Resource Languages

Pretam Ray, Jivnesh Sandhan, Amrit Krishna, Pawan Goyal

Nerual dependency parsing has achieved remarkable performance for low resource morphologically rich languages. It has also been well-studied that morphologically rich languages exhibit relatively free word order. This prompts a fundamental investigation: Is there a way to enhance dependency parsing performance, making the model robust to word order variations utilizing the relatively free word order nature of morphologically rich languages? In this work, we examine the robustness of graph-based parsing architectures on 7 relatively free word order languages. We focus on scrutinizing essential modifications such as data augmentation and the removal of position encoding required to adapt these architectures accordingly. To this end, we propose a contrastive self-supervised learning method to make the model robust to word order variations. Furthermore, our proposed modification demonstrates a substantial average gain of 3.03/2.95 points in 7 relatively free word order languages, as measured by the UAS/LAS Score metric when compared to the best performing baseline.

11:45 - 12:00 - Flagler

Tokenization Is More Than Compression

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, Chris Tanner

Tokenization is a foundational step in natural language processing (NLP) tasks, bridging raw text and language models. Existing tokenization approaches like Byte-Pair Encoding (BPE) originate from the field of data compression, and it has been suggested that the effectiveness of BPE stems from its ability to condense text into a relatively small number of tokens. We test the hypothesis that fewer tokens lead to better downstream performance by introducing PathPiece, a new tokenizer that segments a document's text into the minimum number of tokens for a given vocabulary. Through extensive experimentation we find this hypothesis not to be the case, casting doubt on the understanding of the reasons for effective tokenization. To examine which other factors play a role, we evaluate design decisions across all three phases of tokenization: pre-tokenization, vocabulary construction, and segmentation, offering new insights into the design of effective tokenizers. Specifically, we illustrate the importance of pre-tokenization and the benefits of using BPE to initialize vocabulary construction. We train 64 language models with varying tokenization, ranging in size from 350M to 2.4B parameters, all of which are made publicly available.

Question Answering 2

Nov 13 (Wed) 10:30-12:00 - Room: Monroe

10:30 - 10:45 - Monroe

CLAPNQ: Cohesive Long-form Answers from Passages in Natural Questions for RAG systems

Sara Rosenthal, Avirup Sil, Radu Florian, Salim Roukos

Retrieval Augmented Generation (RAG) has become a popular application for large language models. It is preferable that successful RAG systems provide accurate answers that are supported by being grounded in a passage without any hallucinations. While considerable work is required for building a full RAG pipeline, being able to benchmark performance is also necessary. We present ClapNQ a benchmark Long-form Question Answering dataset for the full RAG pipeline. ClapNQ includes long answers with grounded gold passages from Natural Questions (NQ) and a corpus to perform either retrieval, generation, or the full RAG pipeline. The ClapNQ answers are concise, 3x smaller than the full passage, and cohesive, meaning that the answer is composed fluently, often by integrating multiple pieces of the passage that are not contiguous. RAG models must adapt to these properties to be successful at ClapNQ. We present baseline experiments and analysis for ClapNQ that highlight areas where there is still significant room for improvement in grounded RAG. ClapNQ is publicly available at <https://github.com/primeqa/clapnq>.

10:45 - 11:00 - Monroe

Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering

Sungho Ko, Hyunjin Cho, Hyungjoo Chae, Jinyoung Yeo, Dongha Lee

Recent studies have investigated utilizing Knowledge Graphs (KGs) to enhance Question Answering (QA) performance of Large Language Models (LLMs), yet structured KG verbalization remains challenging. Existing methods, like concatenation or free-form textual conversion of triples, have limitations, including duplicated entities and relations, reduced evidence density, and failure to highlight crucial evidence. To address these issues, we propose EFSum, an Evidence-focused Fact Summarization framework for enhanced QA with knowledge-augmented LLMs. We optimize an LLM as a fact summarizer through distillation and preference alignment. Our extensive experiments show that EFSum improves LLM's zero-shot QA performance with its helpful and faithful summaries, especially when noisy facts are retrieved.

Question Answering 1

Nov 13 (Wed) 10:30-12:00 - Room: Monroe

11:00 - 11:15 - Monroe

Adaptive Query Rewriting: Aligning Rewriters through Marginal Probability of Conversational Answers

Tianhua Zhang, Kun Li, Hongyin Luo, Xixin Wu, James R. Glass, Helen M. Meng

Query rewriting is a crucial technique for passage retrieval in open-domain conversational question answering (CQA). It decontextualizes conversational queries into self-contained questions suitable for off-the-shelf retrievers. Existing methods attempt to incorporate retriever's preference during the training of rewriting models. However, these approaches typically rely on extensive annotations such as in-domain rewrites and/or relevant passage labels, limiting the models' generalization and adaptation capabilities. In this paper, we introduce AdaQR (Adaptive Query Rewriting), a framework for training query rewriting models with limited rewrite annotations from seed datasets and completely no passage label. Our approach begins by fine-tuning compact large language models using only 10% of rewrite annotations from the seed dataset training split. The models are then utilized to self-sample rewrite candidates for each query instance, further eliminating the expense for human labeling or larger language model prompting often adopted in curating preference data. A novel approach is then proposed

to assess retriever's preference for these candidates with the probability of answers conditioned on the conversational query by marginalizing the Top- K passages. This serves as the reward for optimizing the rewriter further using Direct Preference Optimization (DPO), a process free of rewrite and retrieval annotations. Experimental results on four open-domain CQA datasets demonstrate that AdaQR not only enhances the in-domain capabilities of the rewriter with limited annotation requirement, but also adapts effectively to out-of-domain datasets.

Question Answering 2

Nov 13 (Wed) 10:30-12:00 - Room: Monroe

11:15 - 11:30 - Monroe

MedCoT: Medical Chain of Thought via Hierarchical Expert

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, Zuozhu Liu

Artificial intelligence has advanced in Medical Visual Question Answering (Med-VQA), but prevalent research tends to focus on the accuracy of the answers, often overlooking the reasoning paths and interpretability, which are crucial in clinical settings. Besides, current Med-VQA algorithms, typically reliant on singular models, lack the robustness needed for real-world medical diagnostics which usually require collaborative expert evaluation. To address these shortcomings, this paper presents MedCoT, a novel hierarchical expert verification reasoning chain method designed to enhance interpretability and accuracy in biomedical imaging inquiries. MedCoT is predicated on two principles: The necessity for explicit reasoning paths in Med-VQA and the requirement for multi-expert review to formulate accurate conclusions. The methodology involves an Initial Specialist proposing diagnostic rationales, followed by a Follow-up Specialist who validates these rationales, and finally, a consensus is reached through a vote among a sparse Mixture of Experts within the locally deployed Diagnostic Specialist, which then provides the definitive diagnosis. Experimental evaluations on four standard Med-VQA datasets demonstrate that MedCoT surpasses existing state-of-the-art approaches, providing significant improvements in performance and interpretability.

11:30 - 11:45 - Monroe

QUITE: Quantifying Uncertainty in Natural Language Text in Bayesian Reasoning Scenarios

Timo Pierron Schrader, Lukas Lange, Simon Razniewski, Annemarie Friedrich

Reasoning is key to many decision making processes. It requires consolidating a set of rule-like premises that are often associated with degrees of uncertainty and observations to draw conclusions. In this work, we address both the case where premises are specified as numeric probabilistic rules and situations in which humans state their estimates using words expressing degrees of certainty. Existing probabilistic reasoning datasets simplify the task, e.g., by requiring the model to only rank textual alternatives, by including only binary random variables, or by making use of a limited set of templates that result in less varied text. In this work, we present QUITE, a question answering dataset of real-world Bayesian reasoning scenarios with categorical random variables and complex relationships. QUITE provides high-quality natural language verbalizations of premises together with evidence statements and expects the answer to a question in the form of an estimated probability. We conduct an extensive set of experiments, finding that logic-based models outperform out-of-the-box large language models on all reasoning types (causal, evidential, and explaining-away). Our results provide evidence that neuro-symbolic models are a promising direction for improving complex reasoning. We release QUITE and code for training and experiments on Github.

11:45 - 12:00 - Monroe

Verifiable, Debuggable, and Repairable Commonsense Logical Reasoning via LLM-based Theory Resolution

Armin Toroghi, Willis Guo, Ali Pesarangholer, Scott Sanner

Recent advances in Large Language Models (LLM) have led to substantial interest in their application to commonsense reasoning tasks. Despite their potential, LLMs are susceptible to reasoning errors and hallucinations that may be harmful in use cases where accurate reasoning is critical. This challenge underscores the need for verifiable, debuggable, and repairable LLM reasoning. Recent works have made progress toward verifiable reasoning with LLMs by using them as either (i) a reasoner over an axiomatic knowledge base, or (ii) a semantic parser for use in existing logical inference systems. However, both settings are unable to extract commonsense axioms from the LLM that are not already formalized in the knowledge base, and also lack a reliable method to repair missed commonsense inferences. In this work, we present LLM-TRes, a logical reasoning framework based on the notion of "theory resolution" that allows for seamless integration of the commonsense knowledge from LLMs with a verifiable logical reasoning framework that mitigates hallucinations and facilitates debugging of the reasoning procedure as well as repair. We crucially prove that repaired axioms are theoretically guaranteed to be given precedence over flawed ones in our theory resolution inference process. We conclude by evaluating on three diverse language-based reasoning tasks: preference reasoning, deductive reasoning, and causal commonsense reasoning and demonstrate the superior performance of LLM-TRes vs. state-of-the-art LLM-based reasoning methods in terms of both accuracy and reasoning correctness.

Industry

Nov 13 (Wed) 10:30-12:00 - Room: Tuttle

10:30 - 10:40 - Tuttle

Greenback Bears and Fiscal Hawks: Finance is a Jungle and Text Embeddings Must Adapt

Peter Anderson, Mano Vikash Janardhanan, Jason He, Wei Cheng, Charlie Flanagan

Financial documents are filled with specialized terminology, arcane jargon, and curious acronyms that pose challenges for general-purpose text embeddings. Yet, no text embeddings specialized for finance have been reported in the literature, perhaps in part due to a lack of public datasets and benchmarks. We present BAM embeddings, a set of text embeddings finetuned on a carefully constructed dataset of 14.3M query-passage pairs including both public and proprietary financial documents. Demonstrating the benefits of domain-specific training, BAM embeddings achieve Recall@1 of 62.8% on a held-out test set, vs. only 39.2% for the best general-purpose text embedding from OpenAI. Further, BAM embeddings increase question answering accuracy by 8% on FinanceBench and show increased sensitivity to the finance-specific elements that are found in detailed, forward-looking and company and date-specific queries. To support further research we describe our approach in detail, quantify the importance of hard negative mining and dataset scale, and publicly release our embeddings.

10:40 - 10:50 - Tuttle

FastAdaSP: Multitask-Adapted Efficient Inference for Large Speech Language Model

Yichen Lu, Jiaqi Song, Chao-Han Huck Yang, Shinji Watanabe

In this study, we aim to explore Multitask Speech Language Model (SpeechLM) efficient inference via token reduction. Unlike other modal-

ties such as vision or text, speech has unique temporal dependencies, making previous efficient inference works on other modalities not directly applicable. Furthermore, methods for efficient SpeechLM inference on long sequence and sparse signals remain largely unexplored. In this work, we propose FastAdaSP, a weighted token merging framework specifically designed for various speech-related tasks to improve the trade-off between efficiency and performance. Experimental results on WavLMM and Owen-Audio show that our method achieves the state-of-the-art (SOTA) efficiency-performance trade-off compared with other baseline methods. Specifically, FastAdaSP achieved 7x memory efficiency and 1.83x decoding throughput without any degradation on tasks like Emotion Recognition (ER) and Spoken Question Answering (SQA).

10:50 - 11:00 - Tuttle

Detecting LLM-Assisted Cheating on Open-Ended Writing Tasks on Language Proficiency Tests

Chenhao Niu, Kevin P. Yancey, Ruidong Liu, Mirza Basim Baig, André Kenji Horie, James Sharpnack

The high capability of recent Large Language Models (LLMs) has led to concerns about possible misuse as cheating assistants in open-ended writing tasks in assessments. Although various detecting methods have been proposed, most of them have not been evaluated on or optimized for real-world samples from LLM-assisted cheating, where the generated text is often copy-typed imperfectly by the test-taker. In this paper, we present a framework for training LLM-generated text detectors that can effectively detect LLM-generated samples after being copy-typed. We enhance the existing transformer-based classifier training process with contrastive learning on constructed pairwise data and self-training on unlabeled data, and evaluate the improvements on real-world dataset. Our experiments demonstrate the effectiveness of the improved model over both the original transformer-based classifier and GPTZero, a commercial LLM-generated text detector.

11:00 - 11:10 - Tuttle

DiAL : Diversity Aware Listwise Ranking for Query Auto-Complete

Sonali Singh, Sachin Sudhakar Farfade, Prakash Mandayam Comar

Query Auto-Complete (QAC) is an essential search feature that suggests users with a list of potential search keyword completions as they type, enabling them to complete their queries faster. While the QAC systems in eCommerce stores generally use the Learning to Rank (LTR) approach optimized based on customer feedback, it struggles to provide diverse suggestions, leading to repetitive queries and limited navigational suggestions related to product categories, attributes, and brands. This paper proposes a novel DiAL framework that explicitly optimizes for diversity alongside customer feedback signals. It achieves this by leveraging a smooth approximation of the diversity-based metric (α NDCG) as a listwise loss function and modifying it to balance relevance and diversity. The proposed approach yielded an improvement of 8.5% in mean reciprocal rank (MRR) and 22.8% in α NDCG compared to the pairwise ranking approach on an eCommerce dataset, while meeting the ultra-low latency constraints of QAC systems. In an online experiment, the diversity-aware listwise QAC model resulted in a 0.48% lift in revenue. Furthermore, we replicated the proposed approach on a publicly available search log, demonstrating improvements in both diversity and relevance of the suggested queries.

11:10 - 11:20 - Tuttle

Neural Search Space in Gboard Decoder

Yanxiang Zhang, Yuanbo Zhang, Haicheng Sun, Yun Wang, Gary Sivek, Shumin Zhai

Gboard Decoder produces suggestions by looking for paths that best match input touch points on the context aware search space, which is backed by the language Finite State Transducers (FST). The language FST is currently an N-gram language model (LM). However, N-gram LM, limited in context length, and known to have sparsity problem under device model size constraint. In this paper, we propose Neural Search Space which substitutes the N-gram LM with a Neural Network LM (NN-LM) and dynamically constructs the search space during decoding. Specifically, we integrate the long range context awareness of NN-LM into the search space by converting its outputs given context, into the language FST at runtime. This involves language FST structure redesign, pruning strategies tuning, and data structure optimizations. Online experiments demonstrate improved quality results, reducing Words Modified Ratio by [0.26%, 1.19%] on various locales with acceptable latency increases. This work opens new avenues for further improving keyboard decoding quality by enhancing neural LM more directly.

11:20 - 11:30 - Tuttle

Prompt Leakage effect and mitigation strategies for multi-turn LLM Applications

Divyansh Agarwal, Alexander Fabri, Ben Fisher, Philippe Laban, Shafiq Joty, Chien-Sheng Wu

Prompt leakage poses a compelling security and privacy threat in LLM applications. Leakage of system prompts may compromise intellectual property, and act as adversarial reconnaissance for an attacker. A systematic evaluation of prompt leakage threats and mitigation strategies is lacking, especially for multi-turn LLM interactions. In this paper, we systematically investigate LLM vulnerabilities against prompt leakage for 10 closed- and open-source LLMs, across four domains. We design a unique threat model which leverages the LLM sycophancy effect and elevates the average attack success rate (ASR) from 17.7% to 86.2% in a multi-turn setting. Our standardized setup further allows dissecting leakage of specific prompt contents such as task instructions and knowledge documents. We measure the mitigation effect of 7 black-box defense strategies, along with finetuning an open-source model to defend against leakage attempts. We present different combination of defenses against our threat model, including a cost analysis. Our study highlights key takeaways for building secure LLM applications and provides directions for research in multi-turn LLM interactions.

11:30 - 11:40 - Tuttle

Debiasing Text Safety Classifiers through a Fairness-Aware Ensemble

Olivia Starman, Aparna R Joshi, Piyush Kumar, Bhaktipriya Radharaju, Renee Shelby

Increasing use of large language models (LLMs) demand performant guardrails to ensure the safety of inputs and outputs of LLMs. When these safeguards are trained on imbalanced data, they can learn the societal biases. We present a light-weight, post-processing method for mitigating counterfactual fairness in closed-source text safety classifiers. Our approach involves building an ensemble that not only outperforms the input classifiers and policy-aligns them, but also acts as a debiasing regularizer. We introduce two threshold-agnostic metrics to assess the counterfactual fairness of a model, and demonstrate how combining these metrics with Fair Data Reweighting (FDR) helps mitigate biases. We create an expanded Open AI dataset, and a new templated LLM-generated dataset based on user-prompts, both of which are counterfactually balanced across identity groups and cover four key areas of safety; we will work towards publicly releasing these datasets. Our results show that our approach improves counterfactual fairness with minimal impact on model performance.

11:40 - 11:50 - Tuttle

Robust ASR Error Correction with Conservative Data Filtering

Takuma Udagawa, Masayuki Suzuki, Masayasu Muraoka, Gakuto Kurata

Error correction (EC) based on large language models is an emerging technology to enhance the performance of automatic speech recognition (ASR) systems. Generally, training data for EC are collected by automatically pairing a large set of ASR hypotheses (as sources) and their gold references (as targets). However, the quality of such pairs is not guaranteed, and we observed various types of noise which can make the EC models brittle, e.g., inducing overcorrection in out-of-domain (OOD) settings. In this work, we propose two fundamental criteria that EC training data should satisfy: namely, EC targets should (1) improve linguistic acceptability over sources and (2) be inferable from the available context (e.g. source phonemes). Through these criteria, we identify low-quality EC pairs and train the models not to make

any correction in such cases, the process we refer to as conservative data filtering. In our experiments, we focus on Japanese ASR using a strong Conformer-CTC as the baseline and finetune Japanese LLMs for EC. Through our evaluation on a suite of 21 internal benchmarks, we demonstrate that our approach can significantly reduce overcorrection and improve both the accuracy and quality of ASR results in the challenging OOD settings.

11:50 - 12:00 - Tuttle

SHIELD: LLM-Driven Schema Induction for Predictive Analytics in EV Battery Supply Chain Disruptions

Zhi-Qi Cheng, Yifei Dong, Yuzhi Hu, Alike Shi, Wei Liu, Jason O'Connor, Alexander Hauptmann, Kate Whitefoot

The electric vehicle (EV) battery supply chain's vulnerability to disruptions necessitates advanced predictive analytics. We present SHIELD (Schema-based Hierarchical Induction for EV supply chain Disruption), a system integrating Large Language Models (LLMs) with domain expertise for EV battery supply chain risk assessment. SHIELD combines: (1) LLM-driven schema learning to construct a comprehensive knowledge library, (2) a disruption analysis system utilizing fine-tuned language models for event extraction, multi-dimensional similarity matching for schema matching, and Graph Convolutional Networks (GCNs) with logical constraints for prediction, and (3) an interactive interface for visualizing results and incorporating expert feedback to enhance decision-making. Evaluated on 12,070 paragraphs from 365 sources (2022-2023), SHIELD outperforms baseline GCNs and LLM+prompt methods (e.g. GPT-4o) in disruption prediction. These results demonstrate SHIELD's effectiveness in combining LLM capabilities with domain expertise for enhanced supply chain risk assessment.

Session 09 - Nov 13 (Wed) 16:00-17:30

Resources and Evaluation 4

Nov 13 (Wed) 16:00-17:30 - Room: Ashe Auditorium

16:00 - 16:15 - Ashe Auditorium

An image speaks a thousand words, but can everyone listen? On image transcreation for cultural relevance

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, Graham Neubig

Given the rise of multimodal content, human translators increasingly focus on culturally adapting not only words but also other modalities such as images to convey the same meaning. While several applications stand to benefit from this, machine translation systems remain confined to dealing with language in speech and text. In this work, we introduce a new task of translating images to make them culturally relevant. First, we build three pipelines comprising state-of-the-art generative models to do the task. Next, we build a two-part evaluation dataset – (i) concept: comprising 600 images that are cross-culturally coherent, focusing on a single concept per image; and (ii) application: comprising 100 images curated from real-world applications. We conduct a multi-faceted human evaluation of translated images to assess for cultural relevance and meaning preservation. We find that as of today, image-editing models fail at this task, but can be improved by leveraging LLMs and retrievers in the loop. Best pipelines can only translate 5% of images for some countries in the easier concept dataset and no translation is successful for some countries in the application dataset, highlighting the challenging nature of the task. Our project webpage is here: <https://machine-transcreation.github.io/image-transcreation> and our code, data and model outputs can be found here: <https://github.com/simran-khanuja/image-transcreation>.

16:15 - 16:30 - Ashe Auditorium

C3PA: An Open Dataset of Expert-Annotated and Regulation-Aware Privacy Policies to Enable Scalable Regulatory Compliance Audits

Mauz Bin Musa, Rishab Nithyanand, Padmini Srinivasan, Mihailis E. Diamantis, Steven M. Winston, Garrison Allen, Jacob Schiller, Kevin Moore, Sean Quick, Johnathan Melvin

The development of tools and techniques to analyze and extract organizations data habits from privacy policies are critical for scalable regulatory compliance audits. Unfortunately, these tools are becoming increasingly limited in their ability to identify compliance issues and fixes. After all, most were developed using regulation-agnostic datasets of annotated privacy policies obtained from a time before the introduction of landmark privacy regulations such as EU's GDPR and California's CCPA. In this paper, we describe the first open regulation-aware dataset of expert-annotated privacy policies, C3PA (CCPA Privacy Policy Provision Annotations), aimed to address this challenge. C3PA contains over 48K expert-labeled privacy policy text segments associated with responses to CCPA-specific disclosure mandates from 411 unique organizations. We demonstrate that the C3PA dataset is uniquely suited for aiding automated audits of compliance with CCPA-related disclosure mandates.

16:30 - 16:45 - Ashe Auditorium

Evaluating Synthetic Data Generation from User Generated Text

Jenny Chim, Julia Iave, Maria Liakata

User-generated content provides a rich resource to study social and behavioral phenomena. Although its application potential is currently limited by the paucity of expert labels and the privacy risks inherent in personal data, synthetic data can help mitigate this bottleneck. In this work, we introduce an evaluation framework to facilitate research on synthetic language data generation for user-generated text. We define a set of aspects for assessing data quality, namely style preservation, meaning preservation, and divergence, as a proxy for privacy. We introduce metrics corresponding to each aspect. Moreover, through a set of generation strategies and representative tasks and baselines across domains, we demonstrate the relation between the quality aspects of synthetic user generated content, generation strategies, metrics and downstream performance. To our knowledge, our work is the first unified evaluation framework for user-generated text in relation to the specified aspects, offering both intrinsic and extrinsic evaluation. We envisage it will facilitate developments towards shareable, high-quality synthetic language data.

16:45 - 17:00 - Ashe Auditorium

Finding Blind Spots in Evaluator LLMs with Interpretable Checklists

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubham Verma, Mitesh M Khapra

Large Language Models (LLMs) are increasingly relied upon to evaluate text outputs of other LLMs, thereby influencing leaderboards and development decisions. However, concerns persist over the accuracy of these assessments and the potential for misleading conclusions. In this work, we investigate the effectiveness of LLMs as evaluators for text generation tasks. We propose FBI, a novel framework designed to examine the proficiency of Evaluator LLMs in assessing four critical abilities in other LLMs: factual accuracy, instruction following, coherence in long-form writing, and reasoning proficiency. By introducing targeted perturbations in answers generated by LLMs, that clearly impact one of these key capabilities, we test whether an Evaluator LLM can detect these quality drops. By creating a total of 2400 per-

turbed answers covering 22 perturbation categories, we conduct a comprehensive study using different evaluation strategies on five prominent LLMs commonly used as evaluators in the literature. Our findings reveal significant shortcomings in current Evaluator LLMs, which failed to identify quality drops in over 50% of cases on average. Single-answer and pairwise evaluations demonstrated notable limitations, whereas reference-based evaluations showed comparatively better performance. *These results underscore the unreliable nature of current Evaluator LLMs and advocate for cautious implementation in practical applications.*

17:00 - 17:15 - Ashe Auditorium

LLM Assist NLP Researchers: Critique Paper (Meta-)Reviewing

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Bipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, Wenpeng Yin
 Claim: This work is not advocating the use of LLMs for paper (meta-)reviewing. Instead, we present a comparative analysis to identify and distinguish LLM activities from human activities. Two research goals: i) Enable better recognition of instances when someone implicitly uses LLMs for reviewing activities; ii) Increase community awareness that LLMs, and AI in general, are currently inadequate for performing tasks that require a high level of expertise and nuanced judgment. This work is motivated by two key trends. On one hand, large language models (LLMs) have shown remarkable versatility in various generative tasks such as writing, drawing, and question answering, significantly reducing the time required for many routine tasks. On the other hand, researchers, whose work is not only time-consuming but also highly expertise-demanding, face increasing challenges as they have to spend more time reading, writing, and reviewing papers. This raises the question: how can LLMs potentially assist researchers in alleviating their heavy workload? This study focuses on the topic of LLMs as NLP Researchers, particularly examining the effectiveness of LLMs in assisting paper (meta-)reviewing and its recognizability. To address this, we constructed the ReviewCritique dataset, which includes two types of information: (i) NLP papers (initial submissions rather than camera-ready) with both human-written and LLM-generated reviews, and (ii) each review comes with "deficiency" labels and corresponding explanations for individual segments annotated by experts. Using ReviewCritique, this study explores two threads of research questions: (i) "LLMs as Reviewers", how do reviews generated by LLMs compare with those written by humans in terms of quality and distinguishability? (ii) "LLMs as Metareviewers", how effectively can LLMs identify potential issues, such as Deficient or unprofessional review segments, within individual paper reviews? To our knowledge, this is the first work to provide such a comprehensive analysis.

17:15 - 17:30 - Ashe Auditorium

CaT-Bench: Benchmarking Language Model Understanding of Causal and Temporal Dependencies in Plans

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, Ray Mooney

Understanding the abilities of LLMs to reason about natural language plans, such as instructional text and recipes, is critical to reliably using them in decision-making systems. A fundamental aspect of plans is the temporal order in which their steps need to be executed, which reflects the underlying causal dependencies between them. We introduce CaT-Bench, a benchmark of Step Order Prediction questions, which test whether a step must necessarily occur before or after another in cooking recipe plans. We use this to evaluate how well frontier LLMs understand causal and temporal dependencies. We find that SOTA LLMs are underwhelming (best zero-shot is only 0.59 in F1), and are biased towards predicting dependence more often, perhaps relying on temporal order of steps as a heuristic. While prompting for explanations and using few-shot examples improve performance, the best F1 result is only 0.73. Further, human evaluation of explanations along with answer correctness show that, on average, humans do not agree with model reasoning. Surprisingly, we also find that explaining after answering leads to better performance than normal chain-of-thought prompting, and LLM answers are not consistent across questions about the same step pairs. Overall, results show that LLMs' ability to detect dependence between steps has significant room for improvement.

Interpretability and Analysis of Models for NLP 4

Nov 13 (Wed) 16:00-17:30 - Room: Brickell

16:00 - 16:15 - Brickell

Backward Lens: Projecting Language Model Gradients into the Vocabulary Space

Shahar Katz, Yonatan Belinkov, Mor Geva, Lior Wolf

Understanding how Transformer-based Language Models (LMs) learn and recall information is a key goal of the deep learning community. Recent interpretability methods project weights and hidden states obtained from the forward pass to the models' vocabularies, helping to uncover how information flows within LMs. In this work, we extend this methodology to LMs backward pass and gradients. We first prove that a gradient matrix can be cast as a low-rank linear combination of its forward and backward passes' inputs. We then develop methods to project these gradients into vocabulary items and explore the mechanics of how new information is stored in the LMs' neurons.

16:15 - 16:30 - Brickell

Dancing in Chains: Reconciling Instruction Following and Faithfulness in Language Models

Zhenyu Wu, Yuhao Zhang, Peng Qi, Yumo Xu, Ruijin Han, Yuan Zhang, Jifan Chen, Bonan Min, zhitheng huang

Modern language models (LMs) need to follow human instructions while being faithful; yet, they often fail to achieve both. Here, we provide concrete evidence of a trade-off between instruction following (i.e., follow open-ended instructions) and faithfulness (i.e., ground responses in given context) when training LMs with these objectives. For instance, fine-tuning LLaMA-7B on instruction following datasets renders it less faithful. Conversely, instruction-tuned Vicuna-7B shows degraded performance at following instructions when further optimized on tasks that require contextual grounding. One common remedy is multi-task learning (MTL) with data mixing, yet it remains far from achieving a synergistic outcome. We propose a simple yet effective method that relies on Reject-sampling by Self-instruct with Continued Fine-tuning (ReSet), which significantly outperforms vanilla MTL. Surprisingly, we find that less is more, as training ReSet with high-quality, yet substantially smaller data (three-fold less) yields superior results. Our findings offer a better understanding of objective discrepancies in alignment training of LMs.

16:30 - 16:45 - Brickell

Dissecting Fine-Tuning Unlearning in Large Language Models

Yihua Hong, Yuelin Zou, Lijie Hu, Ziguan Zeng, Di Wang, Haiqin Yang

Fine-tuning-based unlearning methods prevail for erasing targeted harmful, sensitive, or copyrighted information within large language models while preserving overall capabilities. However, the true effectiveness of the methods is unclear. In this paper, we delve into the limitations of fine-tuning-based unlearning through activation patching and parameter restoration experiments. Our findings reveal that these methods alter the model's knowledge retrieval process, rather than genuinely erasing the problematic knowledge embedded in the model parameters. Furthermore, behavioral tests demonstrate that the unlearning mechanisms inevitably impact the global behavior of the models, affecting

unrelated knowledge or capabilities. Our work advocates the development of more resilient unlearning techniques for truly erasing knowledge.

16:45 - 17:00 - Brickell

Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis

ZEPING YU, Sophia Ananiadou

We find arithmetic ability resides within a limited number of attention heads, with each head specializing in distinct operations. To delve into the reason, we introduce the Comparative Neuron Analysis (CNA) method, which identifies an internal logic chain consisting of four distinct stages from input to prediction: feature enhancing with shallow FFN neurons, feature transferring by shallow attention layers, feature predicting by arithmetic heads, and prediction enhancing among deep FFN neurons. Moreover, we identify the human-interpretable FFN neurons within both feature-enhancing and feature-predicting stages. These findings lead us to investigate the mechanism of LoRA, revealing that it enhances prediction probabilities by amplifying the coefficient scores of FFN neurons related to predictions. Finally, we apply our method in model pruning for arithmetic tasks and model editing for reducing gender bias. Code is on <https://github.com/zepingyu0512/arithmetic-mechanism>.

17:00 - 17:15 - Brickell

Pixelogy: Probing the Linguistic and Visual Knowledge of Pixel-based Language Models

Kushal Tatariya, Vladimir Araujo, Thomas Bauwens, Miryam de Lhoneux

Pixel-based language models have emerged as a compelling alternative to subword-based language modelling, particularly because they can represent virtually any script. PIXEL, a canonical example of such a model, is a vision transformer that has been pre-trained on rendered text. While PIXEL has shown promising cross-script transfer abilities and robustness to orthographic perturbations, it falls short of outperforming monolingual subword counterparts like BERT in most other contexts. This discrepancy raises questions about the amount of linguistic knowledge learnt by these models and whether their performance in language tasks stems more from their visual capabilities than their linguistic ones. To explore this, we probe PIXEL using a variety of linguistic and visual tasks to assess its position on the vision-to-language spectrum. Our findings reveal a substantial gap between the models visual and linguistic understanding. The lower layers of PIXEL predominantly capture superficial visual features, whereas the higher layers gradually learn more syntactic and semantic abstractions. Additionally, we examine variants of PIXEL trained with different text rendering strategies, discovering that introducing certain orthographic constraints at the input level can facilitate earlier learning of surface-level features. With this study, we hope to provide insights that aid the further development of pixel-based language models.

17:15 - 17:30 - Brickell

Towards Faithful Model Explanation in NLP: A Survey

Qing Lyu, Chris Callison-Burch, Marianna Apidianaki

End-to-end neural Natural Language Processing (NLP) models are notoriously difficult to understand. This has given rise to numerous efforts towards model explainability in recent years. One desideratum of model explanation is faithfulness, i.e. an explanation should accurately represent the reasoning process behind the model's prediction. In this survey, we review over 110 model explanation methods in NLP through the lens of faithfulness. We first discuss the definition and evaluation of faithfulness, as well as its significance for explainability. We then introduce recent advances in faithful explanation, grouping existing approaches into five categories: similarity methods, analysis of model-internal structures, backpropagation-based methods, counterfactual intervention, and self-explanatory models. For each category, we synthesize its representative studies, strengths, and weaknesses. Finally, we summarize their common virtues and remaining challenges, and reflect on future work directions towards faithful explainability in NLP.

NLP Applications 3

Nov 13 (Wed) 16:00-17:30 - Room: Flagler

16:00 - 16:15 - Flagler

Qui custodiet ipsos custodes? Who will watch the watchmen? On Detecting AI-generated peer-reviews

Sandeep Kumar, Mohit Sahu, Vardhan Gucche, Tirthankar Ghosal, Asif Ekbal

The integrity of the peer-review process is vital for maintaining scientific rigor and trust within the academic community. With the steady increase in the usage of large language models (LLMs) like ChatGPT in academic writing, there is a growing concern that AI-generated texts could compromise the scientific publishing including peer-reviews. Previous works have focused on generic AI-generated text detection or have presented an approach for estimating the fraction of peer-reviews that can be AI-generated. Our focus here is to solve a real-world problem by assisting the editor or chair in determining whether a review is written by ChatGPT or not. To address this, we introduce the Term Frequency (TF) model, which posits that AI often repeats tokens, and the Review Regeneration (RR) model which is based on the idea that ChatGPT generates similar outputs upon re-prompting. We stress test these detectors against token attack and paraphrasing. Finally we propose an effective defensive strategy to reduce the effect of paraphrasing on our models. Our findings suggest both our proposed methods perform better than other AI text detectors. Our RR model is more robust, although our TF model performs better than the RR model without any attacks. We make our code, dataset, model public.

16:15 - 16:30 - Flagler

CareCorpus+: Expanding and Augmenting Caregiver Strategy Data to Support Pediatric Rehabilitation

Shahla Farzana, Ivana Lucero, Vivian Villegas, Vera C Kaelin, Mary Khetani, Natalie Parde

Caregiver strategy classification in pediatric rehabilitation contexts is strongly motivated by real-world clinical constraints but highly under-sourced and seldom studied in natural language processing settings. We introduce a large dataset of 4,037 caregiver strategies in this setting, a five-fold increase over the nearest contemporary dataset. These strategies are manually categorized into clinically established constructs with high agreement ($\kappa=0.68-0.89$). We also propose two techniques to further address identified data constraints. First, we manually supplement target task data with publicly relevant data from online child health forums. Next, we propose a novel data augmentation technique to generate synthetic caregiver strategies with high downstream task utility. Extensive experiments showcase the quality of our dataset. They also establish evidence that both the publicly available data and the synthetic strategies result in large performance gains, with relative F₁ increases of 22.6% and 50.9%, respectively.

16:30 - 16:45 - Flagler

Conformal Prediction for Natural Language Processing: A Survey

Margarida M. Campos, António Farinha, Chrysoula Zerva, Mário A. T. Figueiredo, André F. T. Martins

The rapid proliferation of large language models and natural language processing (NLP) applications creates a crucial need for uncertainty quantification to mitigate risks such as hallucinations and to enhance decision-making reliability in critical applications. Conformal pre-

diction is emerging as a theoretically sound and practically useful framework, combining flexibility with strong statistical guarantees. Its model-agnostic and distribution-free nature makes it particularly promising to address the current shortcomings of NLP systems that stem from the absence of uncertainty quantification. This paper provides a comprehensive survey of conformal prediction techniques, their guarantees, and existing applications in NLP, pointing to directions for future research and open challenges.

16:45 - 17:00 - Flagler

Consistent Autoformalization for Constructing Mathematical Libraries

Lan Zhang, XIN QUAN, Andre Freitas

Autoformalization is the task of automatically translating mathematical content written in natural language to a formal language expression. The growing language interpretation capabilities of Large Language Models (LLMs), including in formal languages, are lowering the barriers for autoformalization. However, LLMs alone are not capable of consistently and reliably delivering autoformalization, in particular as the complexity and specialization of the target domain grows. As the field evolves into the direction of systematically applying autoformalization towards large mathematical libraries, the need to improve syntactic, terminological and semantic control increases. This paper proposes the coordinated use of three mechanisms, most-similar retrieval augmented generation (MS-RAG), denoising steps, and auto-correction with syntax error feedback (Auto-SEF) to improve autoformalization quality. The empirical analysis, across different models, demonstrates that these mechanisms can deliver autoformalization results which are syntactically, terminologically and semantically more consistent. These mechanisms can be applied across different LLMs and have shown to deliver improved results across different model types.

17:00 - 17:15 - Flagler

LMs learn governing principles of dynamical systems, revealing an in-context neural scaling law

Toni J.B. Liu, Nicolas Boulle, Raphaël Sarfati, Christopher Earls

We study LLMs' ability to extrapolate the behavior of various dynamical systems, including stochastic, chaotic, continuous, and discrete systems, whose evolution is governed by principles of physical interest. Our results show that LLaMA-2, a language model trained on text, achieves accurate predictions of dynamical system time series without fine-tuning or prompt engineering. Moreover, the accuracy of the learned physical rules increases with the length of the input context window, revealing an in-context version of a neural scaling law. Along the way, we present a flexible and efficient algorithm for extracting probability density functions of multi-digit numbers directly from LLMs.

17:15 - 17:30 - Flagler

Text-Tuple-Table: Towards Information Integration in Text-to-Table Generation via Global Tuple Extraction

Zheyi Deng, Chunkai Chan, Yuxin Wang, Wei Fan, Tianshi Zheng, Yaowai Yim, Yangqiu Song

The task of condensing large chunks of textual information into concise and structured tables has gained attention recently due to the emergence of Large Language Models (LLMs) and their potential benefit for downstream tasks, such as text summarization and text mining. Previous approaches often generate tables that directly replicate information from the text, limiting their applicability in broader contexts, as text-to-table generation in real-life scenarios necessitates information extraction, reasoning, and integration. However, there is a lack of both datasets and methodologies towards this task. In this paper, we introduce LiveSum, a new benchmark dataset created for generating summary tables of competitions based on real-time commentary texts. We evaluate the performances of state-of-the-art LLMs on this task in both fine-tuning and zero-shot settings, and additionally propose a novel pipeline called T^3 (Text-Tuple-Table) to improve their performances. Extensive experimental results demonstrate that LLMs still struggle with this task even after fine-tuning, while our approach can offer substantial performance gains without explicit training. Further analyses demonstrate that our method exhibits strong generalization abilities, surpassing previous approaches on several other text-to-table datasets. Our code and data can be found at <https://github.com/HKUST-KnowComp/LiveSum>.

Information Extraction 1

Nov 13 (Wed) 16:00-17:30 - Room: Monroe

16:00 - 16:15 - Monroe

ADELIE: Aligning Large Language Models on Information Extraction

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li

Large language models (LLMs) usually fall short on information extraction (IE) tasks and struggle to follow the complex instructions of IE tasks. This primarily arises from LLMs not being aligned with humans, as mainstream alignment datasets typically do not include IE data. In this paper, we introduce **ADELIE** (**A**ligning large language mo**DEL**s on **I**nformation **E**xtraction), an aligned LLM that effectively solves various IE tasks, including closed IE, open IE, and on-demand IE. We first collect and construct a high-quality alignment corpus IEInstruct for IE. Then we train *ADELIE_SFT* using instruction tuning on IEInstruct. We further train *ADELIE_SFT* with direct preference optimization (DPO) objective, resulting in *ADELIE_DPO*. Extensive experiments on various held-out IE datasets demonstrate that our models (*ADELIE_SFT* and *ADELIE_DPO*) achieve state-of-the-art (SoTA) performance among open-source models. We further explore the general capabilities of ADELIE, and experimental results reveal that their general capabilities do not exhibit a noticeable decline. We have released the code, data, and models to facilitate further research.

16:15 - 16:30 - Monroe

Embedded Named Entity Recognition using Probing Classifiers

Nicholas Popovic, Michael Färber

Streaming text generation, has become a common way of increasing the responsiveness of language model powered applications such as chat assistants. At the same time, extracting semantic information from generated text is a useful tool for applications such as automated fact checking or retrieval augmented generation. Currently, this requires either separate models during inference, which increases computational cost, or destructive fine-tuning of the language model. Instead, we propose an approach called EMBER which enables streaming named entity recognition in decoder-only language models without fine-tuning them and while incurring minimal additional computational cost at inference time. Specifically, our experiments show that EMBER maintains high token generation rates, with only a negligible decrease in speed of around 1% compared to a 43.64% slowdown measured for a baseline. We make our code and data available online, including a toolkit for training, testing, and deploying efficient token classification models optimized for streaming text generation.

16:30 - 16:45 - Monroe

Explicit, Implicit, and Scattered: Revisiting Event Extraction to Capture Complex Arguments

Omar Sharif, Joseph Gatto, MADHUSUDAN BASAK, Sarah Masud Preum

Prior works formulate the extraction of event-specific arguments as a span extraction problem, where event arguments are explicit — i.e. assumed to be contiguous spans of text in a document. In this study, we revisit this definition of Event Extraction (EE) by introducing two key

argument types that cannot be modeled by existing EE frameworks. First, implicit arguments are event arguments which are not explicitly mentioned in the text, but can be inferred through context. Second, scattered arguments are event arguments that are composed of information scattered throughout the text. These two argument types are crucial to elicit the full breadth of information required for proper event modeling. To support the extraction of explicit, implicit, and scattered arguments, we develop a novel dataset, DiscourseEE, which includes 7,464 argument annotations from online health discourse. Notably, 51.2% of the arguments are implicit, and 17.4% are scattered, making DiscourseEE a unique corpus for complex event extraction. Additionally, we formulate argument extraction as a text generation problem to facilitate the extraction of complex argument types. We provide a comprehensive evaluation of state-of-the-art models and highlight critical open challenges in generative event extraction. Our data and codebase are available at <https://omar-sharif03.github.io/DiscourseEE>.

16:45 - 17:00 - Monroe

Learning to Extract Structured Entities Using Language Models

Haojun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meijmans, Xue Liu, James Hensman, Bhaskar Mitra

Recent advances in machine learning have significantly impacted the field of information extraction, with Language Models (LMs) playing a pivotal role in extracting structured information from unstructured text. Prior works typically represent information extraction as triplet-centric and use classical metrics such as precision and recall for evaluation. We reformulate the task to be entity-centric, enabling the use of diverse metrics that can provide more insights from various perspectives. We contribute to the field by introducing Structured Entity Extraction and proposing the Approximate Entity Set Overlap^a (AESOP) metric, designed to appropriately assess model performance. Later, we introduce a new MultiStage Structured Entity Extraction (MuSEE) model that harnesses the power of LMs for enhanced effectiveness and efficiency by decomposing the extraction task into multiple stages. Quantitative and human side-by-side evaluations confirm that our model outperforms baselines, offering promising directions for future advancements in structured entity extraction. Our source code is available at <https://github.com/microsoft/Structured-Entity-Extraction>.

17:00 - 17:15 - Monroe

OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting

Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, Enhong Chen

Entity Linking (EL) is the process of associating ambiguous textual mentions to specific entities in a knowledge base. Traditional EL methods heavily rely on large datasets to enhance their performance, a dependency that becomes problematic in the context of few-shot entity linking, where only a limited number of examples are available for training. To address this challenge, we present OneNet, an innovative framework that utilizes the few-shot learning capabilities of Large Language Models (LLMs) without the need for fine-tuning. To the best of our knowledge, this marks a pioneering approach to applying LLMs to few-shot entity linking tasks. OneNet is structured around three key components prompted by LLMs: (1) an entity reduction processor that simplifies inputs by summarizing and filtering out irrelevant entities, (2) a dual-perspective entity linker that combines contextual cues and prior knowledge for precise entity linking, and (3) an entity consensus judge that employs a unique consistency algorithm to alleviate the hallucination in the entity linking reasoning. Comprehensive evaluations across seven benchmark datasets reveal that OneNet outperforms current state-of-the-art entity linking methods.

17:15 - 17:30 - Monroe

TKGT: Redefinition and A New Way of Text-to-Table Tasks Based on Real World Demands and Knowledge Graphs Augmented LLMs

Peiwen Jiang, Zibo Liu, Xinbo Lin, Ruhui Ma, Yvonne Jie Chen, Jinhua Cheng

The task of text-to-table receives widespread attention, yet its importance and difficulty are underestimated. Existing works use simple datasets similar to table-to-text tasks and employ methods that ignore domain structures. As a bridge between raw text and statistical analysis, the text-to-table task often deals with complex semi-structured texts that refer to specific domain topics in the real world with entities and events, especially from those of social sciences. In this paper, we analyze the limitations of benchmark datasets and methods used in the text-to-table literature and redefine the text-to-table task to improve its compatibility with long text-processing tasks. Based on this redefinition, we propose a new dataset called CPL (Chinese Private Lending), which consists of judgments from China and is derived from a real-world legal academic project. We further propose TKGT (Text-KG-Table), a two stages domain-aware pipeline, which firstly generates domain knowledge graphs (KGs) classes semi-automatically from raw text with the mixed information extraction (Mixed-IE) method, then adopts the hybrid retrieval augmented generation (Hybird-RAG) method to transform it to tables for downstream needs under the guidance of KGs classes. Experiment results show that TKGT achieves state-of-the-art (SOTA) performance on both traditional datasets and the CPL. Our data and main code are available at <https://github.com/jiangpw41/TKGT>.

Machine Learning for NLP 2

Nov 13 (Wed) 16:00-17:30 - Room: Tuttle

16:00 - 16:15 - Tuttle

ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings

Hao Wang, Hao Li, Minlie Huang, Lei Sha

The safety defense methods of Large language models (LLMs) stays limited because the dangerous prompts are manually curated to just few known attack types, which fails to keep pace with emerging varieties. Recent studies found that attaching suffixes to harmful instructions can hack the defense of LLMs and lead to dangerous outputs. However, similar to traditional text adversarial attacks, this approach, while effective, is limited by the challenge of the discrete tokens. This gradient based discrete optimization attack requires over 100,000 LLM calls, and due to the unreadable of adversarial suffixes, it can be relatively easily penetrated by common defense methods such as perplexity filters. To cope with this challenge, in this paper, we propose an Adversarial Suffix Embedding Translation Framework (ASETF), aimed at transforming continuous adversarial suffix embeddings into coherent and understandable text. This method greatly reduces the computational overhead during the attack process and helps to automatically generate multiple adversarial samples, which can be used as data to strengthen LLM's security defense. Experimental evaluations were conducted on Llama2, Vicuna, and other prominent LLMs, employing harmful directives sourced from the Advbench dataset. The results indicate that our method significantly reduces the computation time of adversarial suffixes and achieves a much better attack success rate than existing techniques, while significantly enhancing the textual fluency of the prompts. In addition, our approach can be generalized into a broader method for generating transferable adversarial suffixes that can successfully attack multiple LLMs, even black-box LLMs, such as ChatGPT and Gemini.

16:15 - 16:30 - Tuttle

CoBar: Convergence Balancer for Multitask Finetuning of Large Language Models

Zi Gong, Hang Yu, Cong Liao, Bingchang Liu, Chaoyu Chen, Jianguo Li

Multi-task learning (MTL) benefits the fine-tuning of large language models (LLMs) by providing a single model with improved performance

and generalization ability across tasks, presenting a resource-efficient alternative to developing separate models for each task. Yet, existing MTL strategies for LLMs often fall short by either being computationally intensive or failing to ensure simultaneous task convergence. This paper presents CoBa, a new MTL approach designed to effectively manage task convergence balance with minimal computational overhead. Utilizing Relative Convergence Scores (RCS), Absolute Convergence Scores (ACS), and a Divergence Factor (DF), CoBa dynamically adjusts task weights during the training process, ensuring that the validation loss of all tasks progress towards convergence at an even pace while mitigating the issue of individual task divergence. The results of our experiments involving three disparate datasets underscore that this approach not only fosters equilibrium in task improvement but enhances the LLMs' performance by up to 13% relative to the second-best baselines. Code is open-sourced at <https://github.com/codefuse-ai/MFTCoder>.

16:30 - 16:45 - Tuttle

Efficient Sequential Decision Making with Large Language Models

Dingyang Chen, Qi Zhang, Yinglun Zhu

This paper focuses on extending the success of large language models (LLMs) to sequential decision making. Existing efforts either (i) re-train or finetune LLMs for decision making, or (ii) design prompts for pretrained LLMs. The former approach suffers from the computational burden of gradient updates, and the latter approach does not show promising results. In this paper, we propose a new approach that leverages online model selection algorithms to efficiently incorporate LLMs agents into sequential decision making. Statistically, our approach significantly outperforms both traditional decision making algorithms and vanilla LLM agents. Computationally, our approach avoids the need for expensive gradient updates of LLMs, and throughout the decision making process, it requires only a small number of LLM calls. We conduct extensive experiments to verify the effectiveness of our proposed approach. As an example, on a large-scale Amazon dataset, our approach achieves more than a 6x performance gain over baselines while calling LLMs in only 1.5% of the time steps.

16:45 - 17:00 - Tuttle

Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, James R. Glass

When asked to summarize articles or answer questions given a passage, large language models (LLMs) can hallucinate details and respond with unsubstantiated answers that are inaccurate with respect to the input context. This paper describes a simple approach for detecting such **contextual hallucinations**. We hypothesize that contextual hallucinations are related to the extent to which an LLM attends to information in the provided context versus its own generations. Based on this intuition, we propose a simple hallucination detection model whose input features are given by the ratio of attention weights on the context versus newly generated tokens (for each attention head). We find that a linear classifier based on these _lookback ratio_ features is as effective as a richer detector that utilizes the entire hidden states of an LLM or a text-based entailment model. The lookback ratio-based detector**Lookback Lens**is found to transfer across tasks and even models, allowing a detector that is trained on a 7B model to be applied (without retraining) to a larger 13B model. We further apply this detector to mitigate contextual hallucinations, and find that a simple classifier-guided decoding approach is able to reduce the amount of hallucination, for example by 9.6% in the XSum summarization task.

17:00 - 17:15 - Tuttle

Revisiting Who's Harry Potter: Towards Targeted Unlearning from a Causal Intervention Perspective

Yujian Liu, Yang Zhang, Tommi Jaakkola, Shiyu Chang

This paper investigates Who's Harry Potter (WHP), a pioneering yet insufficiently understood method for LLM unlearning. We explore it in two steps. First, we introduce a new task of LLM targeted unlearning, where given an unlearning target (e.g., a person) and some unlearning documents, we aim to unlearn only the information about the target, rather than everything in the unlearning documents. We further argue that a successful unlearning should satisfy criteria such as not outputting gibberish, not fabricating facts about the unlearning target, and not releasing factual information under jailbreak attacks. Second, we construct a causal intervention framework for targeted unlearning, where the knowledge of the unlearning target is modeled as a confounder between LLM input and output, and the unlearning process as a decomposing process. This framework justifies and extends WHP, deriving a simple unlearning algorithm that includes WHP as a special case. Experiments on existing and new datasets show that our approach, without explicitly optimizing for the aforementioned criteria, achieves competitive performance in all of them.

17:15 - 17:30 - Tuttle

Where is the signal in tokenization space?

Renato Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, Guy Van den Broeck

Large Language Models (LLMs) are typically shipped with tokenizers that *deterministically* encode text into so-called *canonical* token sequences, to which the LLMs assign probability values. One common assumption is that the probability of a piece of text is the probability of its canonical token sequence. However, the tokenization of a string is not unique: e.g., the Llama2 tokenizer encodes 'Tokens' as '[Tok.ens]', but '[Tok.en.s]' also represents the same text. In this paper, we study non-canonical tokenizations. We prove that, given a string, it is computationally hard to find the most likely tokenization for an autoregressive LLM, as well as to compute the marginal probability over all possible tokenizations. We then show how the marginal is, in most cases, indistinguishable from the canonical probability. Surprisingly, we then empirically demonstrate the existence of a significant amount of signal hidden within tokenization space. Notably, by simply aggregating the probabilities of non-canonical tokenizations, we achieve improvements across a range of LLM evaluation benchmarks for a variety of architectures, including transformers and state space models.

Session 11 - Nov 14 (Thu) 10:30-12:00

NLP Applications 4

Nov 14 (Thu) 10:30-12:00 - Room: Ashe Auditorium

10:30 - 10:45 - Ashe Auditorium

Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework

gopendra Vikram singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, Asif Ekbal

Cognitive distortion research holds increasing significance as it sheds light on pervasive errors in thinking patterns, providing crucial insights into mental health challenges and fostering the development of targeted interventions and therapies. This paper delves into the complex domain of cognitive distortions which are prevalent distortions in cognitive processes often associated with mental health issues. Focusing on

patient-doctor dialogues, we introduce a pioneering method for detecting and reasoning about cognitive distortions utilizing Large Language Models (LLMs). Operating within a multimodal context encompassing audio, video, and textual data, our approach underscores the critical importance of integrating diverse modalities for a comprehensive understanding of cognitive distortions. By leveraging multimodal information, including audio, video, and textual data, our method offers a nuanced perspective that enhances the accuracy and depth of cognitive distortion detection and reasoning in a zero-shot manner. Our proposed hierarchical framework adeptly tackles both detection and reasoning tasks, showcasing significant performance enhancements compared to current methodologies. Through comprehensive analysis, we elucidate the efficacy of our approach, offering promising insights into the diagnosis and understanding of cognitive distortions in multimodal settings. The code and dataset can be found here: <https://github.com/clang1234/ZS-CoDR.git>

10:45 - 11:00 - Ashe Auditorium

DOLOMITES: Domain-Specific Long-Form Methodical Tasks

Chaitanya Malaviya, Priyanka Agrawal, Kuzman Ganchev, Pranesh Srinivasan, Fantine Huot, Jonathan Berant, Mark Yatskar, Dipanjan Das, Mirella Lapata, Chris Alberti

Experts in various fields routinely perform methodical writing tasks to plan, organize, and report their work. From a clinician writing a differential diagnosis for a patient, to a teacher writing a lesson plan for students, these tasks are pervasive, requiring to methodically generate structured long-form output for a given input. We develop a typology of methodical tasks structured in the form of a task objective, procedure, input, and output, and introduce DoLoMiTe, a novel benchmark with specifications for 519 such tasks elicited from hundreds of experts from across 25 fields. Our benchmark further contains specific instantiations of methodical tasks with concrete input and output examples (1,857 in total) which we obtain by collecting expert revisions of up to 10 model-generated examples of each task. We use these examples to evaluate contemporary language models highlighting that automating methodical tasks is a challenging long-form generation problem, as it requires performing complex inferences, while drawing upon the given context as well as domain knowledge.

11:00 - 11:15 - Ashe Auditorium

GoldCoin: Grounding Large Language Models in Privacy Laws via Contextual Integrity Theory

Wei Fan, Haoran Li, Zheyi Deng, Weiqi Wang, Yangqiu Song

Privacy issues arise prominently during the inappropriate transmission of information between entities. Existing research primarily studies privacy by exploring various privacy attacks, defenses, and evaluations within narrowly predefined patterns, while neglecting that privacy is not an isolated, context-free concept limited to traditionally sensitive data (e.g., social security numbers), but intertwined with intricate social contexts that complicate the identification and analysis of potential privacy violations. The advent of Large Language Models (LLMs) offers unprecedented opportunities for incorporating the nuanced scenarios outlined in privacy laws to tackle these complex privacy issues. However, the scarcity of open-source relevant case studies restricts the efficiency of LLMs in aligning with specific legal statutes. To address this challenge, we introduce a novel framework, GoldCoin, designed to efficiently ground LLMs in privacy laws for judicial assessing privacy violations. Our framework leverages the theory of contextual integrity as a bridge, creating numerous synthetic scenarios grounded in relevant privacy statutes (e.g., HIPAA), to assist LLMs in comprehending the complex contexts for identifying privacy risks in the real world. Extensive experimental results demonstrate that GoldCoin markedly enhances LLMs' capabilities in recognizing privacy risks across real court cases, surpassing the baselines on different judicial tasks.

11:15 - 11:30 - Ashe Auditorium

Impeding LLM-assisted Cheating in Introductory Programming Assignments via Adversarial Perturbation

Saiful Islam Salim, Rubin Yuchan Yang, Alexander Cooper, Surayshree Ray, Saumya Debray, Sazzadur Rahaman

While Large language model (LLM)-based programming assistants such as CoPilot and ChatGPT can help improve the productivity of professional software developers, they can also facilitate cheating in introductory computer programming courses. Assuming instructors have limited control over the industrial-strength models, this paper investigates the baseline performance of 5 widely used LLMs on a collection of introductory programming problems, examines adversarial perturbations to degrade their performance, and describes the results of a user study aimed at measuring the efficacy of such perturbations in hindering actual code generation for introductory programming assignments. The user study suggests that i) perturbations combinedly reduced the average correctness score by 77%, ii) the drop in correctness caused by these perturbations was affected based on their detectability.

11:30 - 11:45 - Ashe Auditorium

PREDICT: Multi-Agent-based Debate Simulation for Generalized Hate Speech Detection

Sooenne Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, Kyungsik Han

While a few public benchmarks have been proposed for training hate speech detection models, the differences in labeling criteria between these benchmarks pose challenges for generalized learning, limiting the applicability of the models. Previous research has presented methods to generalize models through data integration or augmentation, but overcoming the differences in labeling criteria between datasets remains a limitation. To address these challenges, we propose PREDICT, a novel framework that uses the notion of multi-agent for hate speech detection. PREDICT consists of two phases: (1) PRE (Perspective-based REASONing): Multiple agents are created based on the induced labeling criteria of given datasets, and each agent generates stances and reasons; (2) DICT (Debate using InCongruent references): Agents representing hate and non-hate stances conduct the debate, and a judge agent classifies hate or non-hate and provides a balanced reason. Experiments on five representative public benchmarks show that PREDICT achieves superior cross-evaluation performance compared to methods that focus on specific labeling criteria or majority voting methods. Furthermore, we validate that PREDICT effectively mediates differences between agents' opinions and appropriately incorporates minority opinions to reach a consensus. Our code is available at <https://github.com/Hanyang-HCC-Lab/PREDICT>

11:45 - 12:00 - Ashe Auditorium

Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation

Joseph Marvin Imperial, Gail Forey, Harish Tayyar Madabushi

Domain experts across engineering, healthcare, and education follow strict standards for producing quality content such as technical manuals, medication instructions, and children's reading materials. However, current works in controllable text generation have yet to explore using these standards as references for control. Towards this end, we introduce Standardize, a retrieval-style in-context learning-based framework to guide large language models to align with expert-defined standards. Focusing on English language standards in the education domain as a use case, we consider the Common European Framework of Reference for Languages (CEFR) and Common Core Standards (CCS) for the task of open-ended content generation. Our findings show that models can gain 45% to 100% increase in precise accuracy across open and commercial LLMs evaluated, demonstrating that the use of knowledge artifacts extracted from standards and integrating them in the generation process can effectively guide models to produce better standard-aligned content.

Computational Social Science and Cultural Analytics 3

Nov 14 (Thu) 10:30-12:00 - Room: Brickell

10:30 - 10:45 - Brickell

"They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, Tanu Mitra

Large language models (LLMs) have emerged as an integral part of modern societies, powering user-facing applications such as personal assistants and enterprise applications like recruitment tools. Despite their utility, research indicates that LLMs perpetuate systemic biases. Yet, prior works on LLM harms predominantly focus on Western concepts like race and gender, often overlooking cultural concepts from other parts of the world. Additionally, these studies typically investigate "harm" as a singular dimension, ignoring the various and subtle forms in which harms manifest. To address this gap, we introduce the Covert Harms and Social Threats (CHAST), a set of seven metrics grounded in social science literature. We utilize evaluation models aligned with human assessments to examine the presence of covert harms in LLM-generated conversations, particularly in the context of recruitment. Our experiments reveal that seven out of the eight LLMs included in this study generated conversations riddled with CHAST, characterized by malign views expressed in seemingly neutral language unlikely to be detected by existing methods. Notably, these LLMs manifested more extreme views and opinions when dealing with non-Western concepts like caste, compared to Western ones such as race.

10:45 - 11:00 - Brickell

Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection

Camilla Casula, Sebastiano Vecellio Salto, Alan Ramponi, Sara Tonelli

The use of synthetic data for training models for a variety of NLP tasks is now widespread. However, previous work reports mixed results with regards to its effectiveness on highly subjective tasks such as hate speech detection. In this paper, we present an in-depth qualitative analysis of the potential and specific pitfalls of synthetic data for hate speech detection in English, with 3,500 manually annotated examples. We show that, across different models, synthetic data created through paraphrasing gold texts can improve out-of-distribution robustness from a computational standpoint. However, this comes at a cost: synthetic data fails to reliably reflect the characteristics of real-world data on a number of linguistic dimensions, it results in drastically different class distributions, and it heavily reduces the representation of both specific identity groups and intersectional hate.

11:00 - 11:15 - Brickell

MASIVE: Open-Ended Affective State Identification in English and Spanish

Nicholas Deas, Elsbeth Turcan, Ivan Ernesto Perez Mejia, Kathleen McKeown

In the field of emotion analysis much NLP research focuses on identifying a limited number of discrete emotion categories, often applied across languages. These basic sets, however, are rarely designed with textual data in mind, and culture, language, and dialect can influence how particular emotions are interpreted. In this work, we broaden our scope to a practically unbounded set of affective states, which includes any terms that humans use to describe their experiences of feeling. We collect and publish MASIVE, a dataset of Reddit posts in English and Spanish containing over 1,000 unique affective states each. We then define the new problem of affective state identification for language generation models framed as a masked span prediction task. On this task, we find that smaller finetuned multilingual models outperform much larger LLMs, even on region-specific Spanish affective states. Additionally, we show that pretraining on MASIVE improves model performance on existing emotion benchmarks. Finally, through machine translation experiments, we find that native speaker-written data is vital to good performance on this task.

11:15 - 11:30 - Brickell

OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants

Jaspreet Ranjit, Brihi Joshi, Rebecca Dorn, Laura Petry, Olga Koumoundouros, Jayne Bottarini, Peichen Liu, Eric Rice, Swabha Swayamdipta
Warning: Contents of this paper may be upsetting. Public attitudes towards key societal issues, expressed on online media, are of immense value in policy and reform efforts, yet challenging to understand at scale. We study one such social issue: homelessness in the U.S., by leveraging the remarkable capabilities of large language models to assist social work experts in analyzing millions of posts from Twitter. We introduce a framing typology: Online Attitudes Towards Homelessness (OATH) Frames: nine hierarchical frames capturing critiques, responses and perceptions. We release annotations with varying degrees of assistance from language models, with immense benefits in scaling: 6.5x speedup in annotation time while only incurring a 3 point F1 reduction in performance with respect to the domain experts. Our experiments demonstrate the value of modeling OATH-Frames over existing sentiment and toxicity classifiers. Our large-scale analysis with predicted OATH-Frames on 2.4M posts on homelessness reveal key trends in attitudes across states, time periods and vulnerable populations, enabling new insights on the issue. Our work provides a general framework to understand nuanced public attitudes at scale, on issues beyond homelessness.

11:30 - 11:45 - Brickell

TempoFormer: A Transformer for Temporally-aware Representations in Change Detection

Talia Tseriotou, Adam Tsakalidis, Maria Liakata

Dynamic representation learning plays a pivotal role in understanding the evolution of linguistic content over time. On this front both context and time dynamics as well as their interplay are of prime importance. Current approaches model context via pre-trained representations, which are typically temporally agnostic. Previous work on modelling context and temporal dynamics has used recurrent methods, which are slow and prone to overfitting. Here we introduce TempoFormer, the first task-agnostic transformer-based and temporally-aware model for dynamic representation learning. Our approach is jointly trained on inter and intra context dynamics and introduces a novel temporal variation of rotary positional embeddings. The architecture is flexible and can be used as the temporal representation foundation of other models or applied to different transformer-based architectures. We show new SOTA performance on three different real-time change detection tasks.

11:45 - 12:00 - Brickell

The Empirical Variability of Narrative Perceptions of Social Media Texts

Joel Mire, Maria Antoniak, Elliott Ash, Andrew Piper, Maarten Sap

Most NLP work on narrative detection has focused on prescriptive definitions of stories crafted by researchers, leaving open the questions: how do crowd workers perceive texts to be a story, and why? We investigate this by building StoryPerceptions, a dataset of 2,496 perceptions of storytelling in 502 social media texts from 255 crowd workers, including categorical labels along with free-text storytelling rationales, authorial intent, and more. We construct a fine-grained bottom-up taxonomy of crowd workers' varied and nuanced perceptions of storytelling by open-coding their free-text rationales. Through comparative analyses at the label and code level, we illuminate patterns of disagreement among crowd workers and across other annotation contexts, including prescriptive labeling from researchers and LLM-based predictions. Notably, plot complexity, references to generalized or abstract actions, and holistic aesthetic judgments (such as a sense of cohesion) are especially important in disagreements. Our empirical findings broaden understanding of the types, relative importance, and contentiousness of features relevant to narrative detection, highlighting opportunities for future work on reader-contextualized models of narrative reception.

Sentiment and Semantics

Nov 14 (Thu) 10:30-12:00 - Room: Flagler

10:30 - 10:45 - Flagler

Conditional and Modal Reasoning in Large Language Models

Wesley H. Holliday, Matthew Mandelkern, Cedegao E. Zhang

The reasoning abilities of large language models (LLMs) are the topic of a growing body of research in AI and cognitive science. In this paper, we probe the extent to which twenty-nine LLMs are able to distinguish logically correct inferences from logically fallacious ones. We focus on inference patterns involving conditionals (e.g., “*If* Ann has a queen, *then* Bob has a jack”) and epistemic modals (e.g., ‘Ann *might* have an ace’, ‘Bob *must* have a king’). These inferences have been of special interest to logicians, philosophers, and linguists, since they play a central role in the fundamental human ability to reason about distal possibilities. Assessing LLMs on these inferences is thus highly relevant to the question of how much the reasoning abilities of LLMs match those of humans. All the LLMs we tested make some basic mistakes with conditionals or modals, though zero-shot chain-of-thought prompting helps them make fewer mistakes. Even the best performing LLMs make basic errors in modal reasoning, display logically inconsistent judgments across inference patterns involving epistemic modals and conditionals, and give answers about complex conditional inferences that do not match reported human judgments. These results highlight gaps in basic logical reasoning in today’s LLMs.

10:45 - 11:00 - Flagler

From Form(s) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency

Xenia Ohmer, Dieuwke Hupkes, Elia Bruni

The staggering pace with which the capabilities of large language models (LLMs) are increasing, as measured by a range of commonly used natural language understanding (NLU) benchmarks, raises many questions regarding what understanding means for a language model and how it compares to human understanding. This is especially true since many LLMs are exclusively trained on text, casting doubt on whether their stellar benchmark performances are reflective of a true understanding of the problems represented by these benchmarks, or whether LLMs simply excel at uttering textual forms that correlate with what someone who understands the problem would say. In this philosophically inspired work, we aim to create some separation between form and meaning, with a series of tests that leverage the idea that world understanding should be consistent across presentational modes inspired by Fregean senses of the same meaning. Specifically, we focus on consistency across languages as well as paraphrases. Taking GPT-3.5 as our object of study, we evaluate multisense consistency across five different languages and various tasks. We start the evaluation in a controlled setting, asking the model for simple facts, and then proceed with an evaluation on four popular NLU benchmarks. We find that the models multisense consistency is lacking and run several follow-up analyses to verify that this lack of consistency is due to a sense-dependent task understanding. We conclude that, in this aspect, the understanding of LLMs is still quite far from being consistent and human-like, and deliberate on how this impacts their utility in the context of learning about human language and understanding.

11:00 - 11:15 - Flagler

Grammatical Gender's Influence on Distributional Semantics: A Causal Perspective

Karolina Ewa Stanczak, Kevin Du, Adina Williams, Isabelle Augenstein, Ryan Cotterell

How much meaning influences gender assignment across languages is an active area of research in modern linguistics and cognitive science. We can view current approaches as aiming to determine where gender assignment falls on a spectrum, from being fully arbitrarily determined to being largely semantically determined. For the latter case, there is a formulation of the neo-Whorfian hypothesis, which claims that even inanimate noun gender influences how people conceive of and talk about objects (using the choice of adjective used to modify inanimate nouns as a proxy for meaning). We offer a novel, causal graphical model that jointly represents the interactions between a noun’s grammatical gender, its meaning, and adjective choice. In accordance with past results, we find a relationship between the gender of nouns and the adjectives which modify them. However, when we control for the meaning of the noun, we find that grammatical gender has a near-zero effect on adjective choice, thereby calling the neo-Whorfian hypothesis into question.

11:15 - 11:30 - Flagler

Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving

XIN QUAN, Marco Valentino, Louise A. Dennis, Andre Freitas

Natural language explanations represent a proxy for evaluating explanation-based and multi-step Natural Language Inference (NLI) models. However, assessing the validity of explanations for NLI is challenging as it typically involves the crowd-sourcing of apposite datasets, a process that is time-consuming and prone to logical errors. To address existing limitations, this paper investigates the verification and refinement of natural language explanations through the integration of Large Language Models (LLMs) and Theorem Provers (TPs). Specifically, we present a neuro-symbolic framework, named Explanation-Refiner, that integrates TPs with LLMs to generate and formalise explanatory sentences and suggest potential inference strategies for NLI. In turn, the TP is employed to provide formal guarantees on the logical validity of the explanations and to generate feedback for subsequent improvements. We demonstrate how Explanation-Refiner can be jointly used to evaluate explanatory reasoning, autoformalisation, and error correction mechanisms of state-of-the-art LLMs as well as to automatically enhance the quality of explanations of variable complexity in different domains.

11:30 - 11:45 - Flagler

Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering

Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, Marco Guerini

The potential effectiveness of counterspeech as a hate speech mitigation strategy is attracting increasing interest in the NLG research community, particularly towards the task of automatically producing it. However, automatically generated responses often lack the argumentative richness which characterises expert-produced counterspeech. In this work, we focus on two aspects of counterspeech generation to produce more cogent responses. First, by investigating the tension between helpfulness and harmlessness of LLMs, we test whether the presence of safety guardrails hinders the quality of the generations. Secondly, we assess whether attacking a specific component of the hate speech results in a more effective argumentative strategy to fight online hate. By conducting an extensive human and automatic evaluation, we show how the presence of safety guardrails can be detrimental also to a task that inherently aims at fostering positive social interactions. Moreover, our results show that attacking a specific component of the hate speech, and in particular its implicit negative stereotype and its hateful parts, leads to higher-quality generations.

11:45 - 12:00 - Flagler

PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling

Huachuan Qiu, Lizhi Ma, Zhenzhong Lan

As awareness of mental health issues grows, online counseling support services are becoming increasingly prevalent worldwide. Detecting whether users express suicidal ideation in text-based counseling services is crucial for identifying and prioritizing at-risk individuals. However, the lack of domain-specific systems to facilitate fine-grained suicide detection and corresponding risk assessment in online counseling poses a significant challenge for automated crisis intervention aimed at suicide prevention. In this paper, we propose PsyGUARD, an automated system for detecting suicide ideation and assessing risk in psychological counseling. To achieve this, we first develop a detailed taxonomy for detecting suicide ideation based on foundational theories. We then curate a large-scale, high-quality dataset called PsySUICIDE for suicide detection. To evaluate the capabilities of automated systems in fine-grained suicide detection, we establish a range of baselines. Subsequently, to assist automated services in providing safe, helpful, and tailored responses for further assessment, we propose to build a suite of risk assessment frameworks. Our study not only provides an insightful analysis of the effectiveness of automated risk assessment systems based on fine-grained suicide detection but also highlights their potential to improve mental health services on online counseling platforms. Code, data, and models are available at <https://github.com/qiuahuachuan/PsyGUARD>.

Language Modeling 4

Nov 14 (Thu) 10:30-12:00 - Room: Monroe

10:30 - 10:45 - Monroe

A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, Jiawei Han

In many scientific fields, large language models (LLMs) have revolutionized the way text and other modalities of data (e.g., molecules and proteins) are handled, achieving superior performance in various applications and augmenting the scientific discovery process. Nevertheless, previous surveys on scientific LLMs often concentrate on one or two fields or a single modality. In this paper, we aim to provide a more holistic view of the research landscape by unveiling cross-field and cross-modal connections between scientific LLMs regarding their architectures and pre-training techniques. To this end, we comprehensively survey over 260 scientific LLMs, discuss their commonalities and differences, as well as summarize pre-training datasets and evaluation tasks for each field and modality. Moreover, we investigate how LLMs have been deployed to benefit scientific discovery. Resources related to this survey are available at <https://github.com/yuzhimanhua/Awesome-Scientific-Language-Models>.

10:45 - 11:00 - Monroe

A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Irsrat Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, Jimmy Huang

Large Language Models (LLMs) have recently gained significant attention due to their remarkable capabilities in performing diverse tasks across various domains. However, a thorough evaluation of these models is crucial before deploying them in real-world applications to ensure they produce reliable performance. Despite the well-established importance of evaluating LLMs in the community, the complexity of the evaluation process has led to varied evaluation setups, causing inconsistencies in findings and interpretations. To address this, we systematically review the primary challenges and limitations causing these inconsistencies and unreliable evaluations in various steps of LLM evaluation. Based on our critical review, we present our perspectives and recommendations to ensure LLM evaluations are reproducible, reliable, and robust.

11:00 - 11:15 - Monroe

Consistent Bidirectional Language Modelling: Expressive Power and Representational Conciseness

Georgi Shopov, Stefan Gerdtjikov

The inability to utilise future contexts and the pre-determined left-to-right generation order are major limitations of unidirectional language models. Bidirectionality has been introduced to address those deficiencies. However, a crucial shortcoming of bidirectional language models is the potential inconsistency of their conditional distributions. This fundamental flaw greatly diminishes their applicability and hinders their capability of tractable sampling and likelihood computation. In this work, we introduce a class of bidirectional language models, called latent language models, that are consistent by definition and can be efficiently used both for generation and scoring of sequences. We define latent language models based on the well-understood formalism of bisequential decompositions from automata theory. This formal correspondence allows us to precisely characterise the abilities and limitations of a subclass of latent language models, called rational language models. As a result, we obtain that latent language models are exponentially more concise and significantly more expressive than unidirectional language models.

11:15 - 11:30 - Monroe

Data Everywhere: A Guide for Pretraining Dataset Construction

Jupinder Parmar, Shrimai Prabhunoye, Joseph Jennings, Bo Liu, Aastha Jhunjhunwala, Zhilin Wang, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro

The impressive capabilities of recent language models can be largely attributed to the multi-trillion token pretraining datasets that they are trained on. However, model developers fail to disclose their construction methodology which has lead to a lack of open information on how to develop effective pretraining sets. To address this issue, we perform the first systematic study across the entire pipeline of pretraining set construction. First, we run ablations on existing techniques for pretraining set development to identify which methods translate to the largest gains in model accuracy on downstream evaluations. Then, we categorize the most widely used data source, web crawl snapshots, across the attributes of toxicity, quality, type of speech, and domain. Finally, we show how such attribute information can be used to further refine and improve the quality of a pretraining set. These findings constitute an actionable set of steps that practitioners can use to develop high quality pretraining sets.

Language Modeling 2

Nov 14 (Thu) 10:30-12:00 - Room: Monroe

11:30 - 11:45 - Monroe

Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs

Oded Ovadia, Menachem Brief, Moshiik Mishaeli, Oren Elisha

Large language models (LLMs) encapsulate a vast amount of factual information within their pre-trained weights, as evidenced by their ability to answer diverse questions across different domains. However, this knowledge is inherently limited, relying heavily on the characteristics of the training data. Consequently, using external datasets to incorporate new information or refine the capabilities of LLMs on previously seen information poses a significant challenge. In this study, we compare two common approaches: unsupervised fine-tuning and retrieval-augmented generation (RAG). We evaluate both approaches on a variety of knowledge-intensive tasks across different topics. Our findings reveal that while unsupervised fine-tuning offers some improvement, RAG consistently outperforms it, both for existing knowledge encountered during training and entirely new knowledge. Moreover, we find that LLMs struggle to learn new factual information through unsupervised fine-tuning, and that exposing them to numerous variations of the same fact during training could alleviate this problem.

Language Modeling 4

Nov 14 (Thu) 10:30-12:00 - Room: Monroe

11:45 - 12:00 - Monroe

User Inference Attacks on Large Language Models

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, Zheng Xu

Text written by humans makes up the vast majority of the data used to pre-train and fine-tune large language models (LLMs). Many sources of this data—like code, forum posts, personal websites, and books—are easily attributable to one or a few “users”. In this paper, we ask if it is possible to infer if any of a user’s data was used to train an LLM. Not only would this constitute a breach of privacy, but it would also enable users to detect when their data was used for training. We develop the first effective attacks for user inference—at times, with near-perfect success—against LLMs. Our attacks are easy to employ, requiring only black-box access to an LLM and a few samples from the user, which need not be the ones that were trained on. We find, both theoretically and empirically, that certain properties make users more susceptible to user inference: being an outlier, having highly correlated examples, and contributing a larger fraction of data. Based on these findings, we identify several methods for mitigating user inference including training with example-level differential privacy, removing within-user duplicate examples, and reducing a user’s contribution to the training data. Though these provide partial mitigation, our work highlights the need to develop methods to fully protect LLMs from user inference.

Multilinguality and Language Diversity 2

Nov 14 (Thu) 10:30-12:00 - Room: Tuttle

10:30 - 10:45 - Tuttle

Context-aware Transliteration of Romanized South Asian Languages

While most transliteration research is focused on single tokens such as named entities for example, transliteration of from the Gujarati script to the Latin script Ahmedabad the informal romanization prevalent in South Asia and elsewhere often requires transliteration of full sentences. The lack of large parallel text collections of full sentence (as opposed to single word) transliterations necessitates incorporation of contextual information into transliteration via non-parallel resources, such as via mono-script text collections. In this article, we present a number of methods for improving transliteration in context for such a use scenario. Some of these methods in fact improve performance without making use of sentential context, allowing for better quantification of the degree to which contextual information in particular is responsible for system improvements. Our final systems, which ultimately rely upon ensembles including large pretrained language models fine-tuned on simulated parallel data, yield substantial improvements over the best previously reported results for full sentence transliteration from Latin to native script on all 12 languages in the Dakshina dataset (Roark et al. 2020), with an overall 3.3% absolute (18.6% relative) mean word-error rate reduction.

10:45 - 11:00 - Tuttle

Dotless Arabic text for Natural Language Processing

Irfan Ahmad, Maged S. Al-Shabani

This paper introduces a novel representation of Arabic text as an alternative approach for Arabic NLP, inspired by the dotless script of ancient Arabic. We explored this representation through extensive analysis on various text corpora, differing in size and domain, and tokenized using multiple tokenization techniques. Furthermore, we examined the information density of this representation and compared it with the standard dotted Arabic text using text entropy analysis. Utilizing parallel corpora, we also drew comparisons between Arabic and English text analysis to gain additional insights. Our investigation extended to various upstream and downstream NLP tasks, including language modeling, text classification, sequence labeling, and machine translation, examining the implications of both the representations. Specifically, we performed seven different downstream tasks using various tokenization schemes comparing the standard dotted text with dotless Arabic text representations. The performances using both the representations were comparable across different tokenizations. However, dotless representation achieves these results with significant reduction in vocabulary sizes, and in some scenarios showing reduction of up to 50%. Additionally, we present a system that restores dots to the dotless Arabic text. This system is useful for tasks that require Arabic texts as output.

11:00 - 11:15 - Tuttle

ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu

We present a comprehensive evaluation of large language models for multilingual readability assessment. Existing evaluation resources lack domain and language diversity, limiting the ability for cross-domain and cross-lingual analyses. This paper introduces ReadMe++, a multi-lingual multi-domain dataset with human annotations of 9757 sentences in Arabic, English, French, Hindi, and Russian, collected from 112 different data sources. This benchmark will encourage research on developing robust multilingual readability assessment methods. Using ReadMe++, we benchmark multilingual and monolingual language models in the supervised, unsupervised, and few-shot prompting settings. The domain and language diversity in ReadMe++ enables us to test more effective few-shot prompting, and identify shortcomings in state-of-the-art unsupervised methods. Our experiments also reveal exciting results of superior domain generalization and enhanced cross-lingual transfer capabilities by models trained on ReadMe++. We will make our data publicly available and release a python package tool for multilingual sentence readability prediction using our trained models at: <https://github.com/tareknaous/readme>

11:15 - 11:30 - Tuttle

RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs*John Deng, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, Sara Hooker*

Preference optimization techniques have become a standard final stage for training state-of-art large language models (LLMs). However, despite widespread adoption, the vast majority of work to-date has focused on a small set of high-resource languages like English and Chinese. This captures a small fraction of the languages in the world, but also makes it unclear which aspects of current state-of-the-art research transfer to a multilingual setting. In this work, we perform an exhaustive study to achieve a new state of the art in aligning multilingual LLMs. We introduce a novel, scalable method for generating high-quality multilingual feedback data to balance data coverage. We establish the benefits of cross-lingual transfer and increased dataset size in preference training. Our preference-trained model achieves a 54.4% win-rate against Aya 23 SB, the current state-of-the-art multilingual LLM in its parameter class, and a 69.5% win-rate or higher against widely used models like Gemma, Mistral and Llama 3. As a result of our efforts, we expand the frontier of alignment techniques to 23 languages, covering approximately half of the world's population.

Multilinguality and Language Diversity 1

Nov 14 (Thu) 10:30-12:00 - Room: Tuttle

11:30 - 11:45 - Tuttle

Getting More from Less: Large Language Models are Good Spontaneous Multilingual Learners*Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, Shuijian Huang*

Recently, Large Language Models (LLMs) have shown impressive language capabilities, while most of them have very unbalanced performance across different languages. Multilingual alignment based on the translation parallel data is an effective method to enhance LLMs' multilingual capabilities. In this work, we first discover and comprehensively investigate the spontaneous multilingual alignment of LLMs. Firstly, we find that LLMs instruction-tuned on the question translation data (i.e. without annotated answers) are able to encourage the alignment between English and a wide range of languages, even including those unseen during instruction-tuning. Additionally, we utilize different settings and mechanistic interpretability methods to analyze the LLM's performance in the multilingual scenario comprehensively. Our work suggests that LLMs have enormous potential for improving multilingual alignment efficiently with great language generalization and task generalization.

Multilinguality and Language Diversity 2

Nov 14 (Thu) 10:30-12:00 - Room: Tuttle

11:45 - 12:00 - Tuttle

What is "Typological Diversity" in NLP?*Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, Johannes Bjerva*

The NLP research community has devoted increased attention to languages beyond English, resulting in considerable improvements for multilingual NLP. However, these improvements only apply to a small subset of the world's languages. An increasing number of papers aspires to enhance generalizable multilingual performance across languages. To this end, linguistic typology is commonly used to motivate language selection, on the basis that a broad typological sample ought to imply generalization across a broad range of languages. These selections are often described as being typologically diverse. In this meta-analysis, we systematically investigate NLP research that includes claims regarding typological diversity. We find there are no set definitions or criteria for such claims. We introduce metrics to approximate the diversity of resulting language samples along several axes and find that the results vary considerably across papers. Crucially, we show that skewed language selection can lead to overestimated multilingual performance. We recommend future work to include an operationalization of typological diversity that empirically justifies the diversity of language samples. To help facilitate this, we release the code for our diversity measures.

Session 12 - Nov 14 (Thu) 14:00-15:30**Interpretability and Analysis of Models for NLP 6**

Nov 14 (Thu) 14:00-15:30 - Room: Ashe Auditorium

14:00 - 14:15 - Ashe Auditorium

Can Large Language Models Learn Independent Causal Mechanisms?*Gael Gendron, Bao Trung Nguyen, Alex Yuxuan Peng, Michael Witbrock, Gillian Dobbie*

Despite impressive performance on language modelling and complex reasoning tasks, Large Language Models (LLMs) fall short on the same tasks in uncommon settings or with distribution shifts, exhibiting a lack of generalisation ability. By contrast, systems such as causal models, that learn abstract variables and causal relationships, can demonstrate increased robustness against changes in the distribution. One reason for this success is the existence and use of Independent Causal Mechanisms (ICMs) representing high-level concepts that only sparsely interact. In this work, we apply two concepts from causality to learn ICMs within LLMs. We develop a new LLM architecture composed of multiple sparsely interacting language modelling modules. We show that such causal constraints can improve out-of-distribution performance on abstract and causal reasoning tasks. We also investigate the level of independence and domain specialisation and show that LLMs rely on pre-trained partially domain-invariant mechanisms resilient to fine-tuning.

14:15 - 14:30 - Ashe Auditorium

Do Explanations Help or Hurt? Saliency Maps vs Natural Language Explanations in a Clinical Decision-Support Setting*Maxime Guillaume Kayser, Bayar Menzat, Cornelius Emde, Bogdan Alexandru Bercean, Alex Novak, Abdalá Trinidad Espinosa Morgado,*

Bartłomiej Papiez, Susanne Gaube, Thomas Lukasiewicz, Oana-Maria Camburu

The growing capabilities of AI models are leading to their wider use, including in safety-critical domains. Explainable AI (XAI) aims to make these models safer to use by making their inference process more transparent. However, current explainability methods are seldom evaluated in the way they are intended to be used: by real-world end users. To address this, we conducted a large-scale user study with 85 healthcare practitioners in the context of human-AI collaborative chest X-ray analysis. We evaluated three types of explanations: visual explanations (saliency maps), natural language explanations, and a combination of both modalities. We specifically examined how different explanation types influence users depending on whether the AI advice and explanations are factually correct. We find that text-based explanations lead to significant over-reliance, which is alleviated by combining them with saliency maps. We also observe that the quality of explanations, that is, how much factually correct information they entail, and how much this aligns with AI correctness, significantly impacts the usefulness of the different explanation types.

14:30 - 14:45 - Ashe Auditorium

Filtered Corpus Training (FiCT) Shows that Language Models can Generalize from Indirect Evidence

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang

This paper introduces Filtered Corpus Training, a method that trains language models (LMs) on corpora with certain linguistic constructions filtered out from the training data, and uses it to measure the ability of LMs to perform linguistic generalization on the basis of indirect evidence. We apply the method to both LSTM and Transformer LMs (of roughly comparable size), developing filtered corpora that target a wide range of linguistic phenomena. Our results show that while transformers are better qua LMs (as measured by perplexity), both models perform equally and surprisingly well on linguistic generalization measures, suggesting that they are capable of generalizing from indirect evidence.

14:45 - 15:00 - Ashe Auditorium

Prompts have evil twins

Rimon Melamed, Lucas Hurley McCabe, Tanay Wakhare, Yejin Kim, H. Howie Huang, Enric Boix-Adserà

We discover that many natural-language prompts can be replaced by corresponding prompts that are unintelligible to humans but that provably elicit similar behavior in language models. We call these prompts “evil twins” because they are obfuscated and uninterpretable (evil), but at the same time mimic the functionality of the original natural-language prompts (twins). Remarkably, evil twins transfer between models. We find these prompts by solving a maximum-likelihood problem which has applications of independent interest.

15:00 - 15:15 - Ashe Auditorium

Self-AMPLIFY: Improving Small Language Models with Self Post Hoc Explanations

Milan BHAN, Jean-Noël Vittaut, Nicolas CHESNEAU, Marie-Jeanne Lesot

Incorporating natural language rationales in the prompt and In-Context Learning (ICL) have led to a significant improvement of Large Language Models (LLMs) performance. However, generating high-quality rationales require human-annotation or the use of auxiliary proxy models. In this work, we propose Self-AMPLIFY to automatically generate rationales from post hoc explanation methods applied to Small Language Models (SLMs) to improve their own performance. Self-AMPLIFY is a 3-step method that targets samples, generates rationales and builds a final prompt to leverage ICL. Self-AMPLIFY performance is evaluated on four SLMs and five datasets requiring strong reasoning abilities. Self-AMPLIFY achieves good results against competitors, leading to strong accuracy improvement. Self-AMPLIFY is the first method to apply post hoc explanation methods to autoregressive language models to generate rationales to improve their own performance in a fully automated manner.

15:15 - 15:30 - Ashe Auditorium

Words Worth a Thousand Pictures: Measuring and Understanding Perceptual Variability in Text-to-Image Generation

Raphael Tang, Crystina Zhang, Lixin Xu, Yao Lu, Wenyun Li, Pontus Stenetorp, Jimmy Lin, Ferhan Ture

Diffusion models are the state of the art in text-to-image generation, but their perceptual variability remains understudied. In this paper, we examine how prompts affect image variability in black-box diffusion-based models. We propose W1KP, a human-calibrated measure of variability in a set of images, bootstrapped from existing image-pair perceptual distances. Current datasets do not cover recent diffusion models, thus we curate three test sets for evaluation. Our best perceptual distance outperforms nine baselines by up to 18 points in accuracy, and our calibration matches graded human judgements 78% of the time. Using W1KP, we study prompt reusability and show that Imagen prompts can be reused for 10-50 random seeds before new images become too similar to already generated images, while Stable Diffusion XL and DALL-E 3 can be reused 50-200 times. Lastly, we analyze 56 linguistic features of real prompts, finding that the prompt's length, CLIP embedding norm, concreteness, and word senses influence variability most. As far as we are aware, we are the first to analyze diffusion variability from a visiolinguistic perspective. Our project page is at <http://w1kp.com>.

Speech Processing and Spoken Language Understanding 2

Nov 14 (Thu) 14:00-15:30 - Room: Brickell

14:00 - 14:15 - Brickell

Advancing Test-Time Adaptation in Wild Acoustic Test Settings

Hongfu Liu, Hengguan Huang, Ye Wang

Acoustic foundation models, fine-tuned for Automatic Speech Recognition (ASR), suffer from performance degradation in wild acoustic test settings when deployed in real-world scenarios. Stabilizing online Test-Time Adaptation (TTA) under these conditions remains an open and unexplored question. Existing wild vision TTA methods often fail to handle speech data effectively due to the unique characteristics of high-entropy speech frames, which are unreliable filtered out even when containing crucial semantic content. Furthermore, unlike static vision data, speech signals follow short-term consistency, requiring specialized adaptation strategies. In this work, we propose a novel wild acoustic TTA method tailored for ASR fine-tuned acoustic foundation models. Our method, Confidence-Enhanced Adaptation, performs frame-level adaptation using a confidence-aware weight scheme to avoid filtering out essential information in high-entropy frames. Additionally, we apply consistency regularization during test-time optimization to leverage the inherent short-term consistency of speech signals. Our experiments on both synthetic and real-world datasets demonstrate that our approach outperforms existing baselines under various wild acoustic test settings, including Gaussian noise, environmental sounds, accent variations, and sung speech.

14:15 - 14:30 - Brickell

BLSP-Emo: Towards Empathetic Large Speech-Language Models

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, Jiajun Zhang

The recent release of GPT-4o showcased the potential of end-to-end multimodal models, not just in terms of low latency but also in their

ability to understand and generate expressive speech with rich emotions. While the details are unknown to the open research community, it likely involves significant amounts of curated data and compute, neither of which is readily accessible. In this paper, we present BLSP-Emo (Bootstrapped Language-Speech Pretraining with Emotion support), a novel approach to developing an end-to-end speech-language model capable of understanding both semantics and emotions in speech and generate empathetic responses. BLSP-Emo utilizes existing speech recognition (ASR) and speech emotion recognition (SER) datasets through a two-stage process. The first stage focuses on semantic alignment, following recent work on pretraining speech-language models using ASR data. The second stage performs emotion alignment with the pretrained speech-language model on an emotion-aware continuation task constructed from SER data. Our experiments demonstrate that the BLSP-Emo model excels in comprehending speech and delivering empathetic responses, both in instruction-following tasks and conversations.

14:30 - 14:45 - Brickell

Continual Test-time Adaptation for End-to-end Speech Recognition on Noisy Speech

Guan-Ting Lin, Wei Ping Huang, Hung-yi Lee

Deep Learning-based end-to-end Automatic Speech Recognition (ASR) has made significant strides but still struggles with performance on out-of-domain samples due to domain shifts in real-world scenarios. Test-Time Adaptation (TTA) methods address this issue by adapting models using test samples at inference time. However, current ASR TTA methods have largely focused on non-continual TTA, which limits cross-sample knowledge learning compared to continual TTA. In this work, we first propose a Fast-slow TTA framework for ASR that leverages the advantage of continual and non-continual TTA. Following this framework, we introduce Dynamic SUTA (DSUTA), an entropy-minimization-based continual TTA method for ASR. To enhance DSUTA's robustness for time-varying data, we design a dynamic reset strategy to automatically detect domain shifts and reset the model, making it more effective at handling multi-domain data. Our method demonstrates superior performance on various noisy ASR datasets, outperforming both non-continual and continual TTA baselines while maintaining robustness to domain changes without requiring domain boundary information.

14:45 - 15:00 - Brickell

EH-MAM: Easy-to-Hard Masked Acoustic Modeling for Self-Supervised Speech Representation Learning

Ashish Seth, Ramaneshwar S, S Sakshi, Sonal Kumar, Sreyan Ghosh, Dinesh Manocha

In this paper, we present EH-MAM (Easy-to-Hard adaptive Masked Acoustic Modeling), a novel self-supervised learning approach for speech representation learning. In contrast to the prior methods that use random masking schemes for Masked Acoustic Modeling (MAM), we introduce a novel selective and adaptive masking strategy. Specifically, during SSL training, we progressively introduce harder regions to the model for reconstruction. Our approach automatically selects hard regions and is built on the observation that the reconstruction loss of individual frames in MAM can provide natural signals to judge the difficulty of solving the MAM pre-text task for that frame. To identify these hard regions, we employ a teacher model that first predicts the frame-wise losses and then decides which frames to mask. By learning to create challenging problems, such as identifying harder frames and solving them simultaneously, the model is able to learn more effective representations and thereby acquire a more comprehensive understanding of the speech. Quantitatively, EH-MAM outperforms several state-of-the-art baselines across various low-resource speech recognition and SUPERB benchmarks by 5%-10%. Additionally, we conduct a thorough analysis to show that the regions masked by EH-MAM effectively capture useful context across speech frames.

15:00 - 15:15 - Brickell

Interventional Speech Noise Injection for ASR Generalizable Spoken Language Understanding

YeonJoon Jung, Jaeseong Lee, Seungtaek Choi, Dohyeon Lee, Minsoo Kim, seung-won hwang

Recently, pre-trained language models (PLMs) have been increasingly adopted in spoken language understanding (SLU). However, automatic speech recognition (ASR) systems frequently produce inaccurate transcriptions, leading to noisy inputs for SLU models, which can significantly degrade their performance. To address this, our objective is to train SLU models to withstand ASR errors by exposing them to noises commonly observed in ASR systems, referred to as ASR-plausible noises. Speech noise injection (SNI) methods have pursued this objective by introducing ASR-plausible noises, but we argue that these methods are inherently biased towards specific ASR systems, or ASR-specific noises. In this work, we propose a novel and less biased augmentation method of introducing the noises that are plausible to any ASR system, by cutting off the non-causal effect of noises. Experimental results and analyses demonstrate the effectiveness of our proposed methods in enhancing the robustness and generalizability of SLU models against unseen ASR systems by introducing more diverse and plausible ASR noises in advance.

15:15 - 15:30 - Brickell

SPIRIT-LM: Interleaved Spoken and Written Language Model

Tu Anh Nguyen, Benjamin Müller, Bokai Yu, Marta Costa-jussà, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoît Sagot, Emmanuel Dupoux, Christophe Ropers, Mary Williamson,

We introduce SPIRIT-LM, a foundation multimodal language model that freely mixes text and speech. Our model is based on a 7B pretrained text language model that we extend to the speech modality by continuously training it on text and speech units. Speech and text sequences are concatenated as a single stream of tokens, and trained with a word-level interleaving method using a small automatically-curated speech-text parallel corpus. SPIRIT-LM comes in two versions: a BASE version that uses speech semantic units and an EXPRESSIVE version that models expressivity using pitch and style units in addition to the semantic units. For both versions, the text is encoded with subword BPE tokens. The resulting model displays both the semantic abilities of text models and the expressive abilities of speech models. Additionally, we demonstrate that SPIRIT-LM is able to learn new tasks in a few-shot fashion across modalities (i.e. ASR, TTS, Speech Classification).

Resources and Evaluation 6

Nov 14 (Thu) 14:00-15:30 - Room: Flagler

14:00 - 14:15 - Flagler

ASL STEMpedia: Dataset and Benchmark for Interpreting STEM Articles

Kayo Yin, Chinmay Singh, Fyodor O Minakov, Vanessa Milan, Hal Daumé III, Cyril Zhang, Alex Xijie Lu, Danielle Bragg

Deaf and hard-of-hearing (DHH) students face significant barriers in accessing science, technology, engineering, and mathematics (STEM) education, notably due to the scarcity of STEM resources in signed languages. To help address this, we introduce ASL STEM Wiki: a parallel corpus of 254 Wikipedia articles on STEM topics in English, interpreted into over 300 hours of American Sign Language (ASL). ASL STEM Wiki is the first continuous signing dataset focused on STEM, facilitating the development of AI resources for STEM education in ASL. We identify several use cases of ASL STEM Wiki with human-centered applications. For example, because this dataset highlights the frequent use of fingerspelling for technical concepts, which inhibits DHH students' ability to learn, we develop models to identify fingerspelled words—which can later be used to query for appropriate ASL signs to suggest to interpreters.

14:15 - 14:30 - Flagler

De-Identification of Sensitive Personal Data in Datasets Derived from IIT-CDIP

Stefan Larson, Nicole Cornehl Lima, Santiago Pedroza Diaz, Anogha Manoj Joshi, Siddharth Betala, Jamiu Tunde Suleiman, Yash Mathur, Kaushal Kumar Prajapati, Ramla Alakraa, Junjie Shen, Temi Okotore, Kevin Leach

The IIT-CDIP document collection is the source of several widely used and publicly accessible document understanding datasets. In this paper, manual inspection of 5 datasets derived from IIT-CDIP uncovers the presence of thousands of instances of sensitive personal data, including US Social Security Numbers (SSNs), birth places and dates, and home addresses of individuals. The presence of such sensitive personal data in commonly-used and publicly available datasets is startling and has ethical and potentially legal implications; we believe such sensitive data ought to be removed from the internet. Thus, in this paper, we develop a modular data de-identification pipeline that replaces sensitive data with synthetic, but realistic, data. Via experiments, we demonstrate that this de-identification method preserves the utility of the de-identified documents so that they can continue be used in various document understanding applications. We will release redacted versions of these datasets publicly.

Resources and Evaluation 2

Nov 14 (Thu) 14:00-15:30 - Room: Flagler

14:30 - 14:45 - Flagler

CoCoLoFa: A Dataset of News Comments with Common Logical Fallacies Written by LLM-Assisted Crowds

Min-Hsuan Yeh, Ruyuan Wan, Ting-Hao Kenneth Huang

Detecting logical fallacies in texts can help users spot argument flaws, but automating this detection is not easy. Manually annotating fallacies in large-scale, real-world text data to create datasets for developing and validating detection models is costly. This paper introduces CoCoLoFa, the largest known logical fallacy dataset, containing 7,706 comments for 648 news articles, with each comment labeled for fallacy presence and type. We recruited 143 crowd workers to write comments embodying specific fallacy types (e.g., slippery slope) in response to news articles. Recognizing the complexity of this writing task, we built an LLM-powered assistant into the workers' interface to aid in drafting and refining their comments. Experts rated the writing quality and labeling validity of CoCoLoFa as high and reliable. BERT-based models fine-tuned using CoCoLoFa achieved the highest fallacy detection ($F1=0.86$) and classification ($F1=0.87$) performance on its test set, outperforming the state-of-the-art LLMs. Our work shows that combining crowdsourcing and LLMs enables us to more effectively construct datasets for complex linguistic phenomena that crowd workers find challenging to produce on their own.

Resources and Evaluation 6

Nov 14 (Thu) 14:00-15:30 - Room: Flagler

14:45 - 15:00 - Flagler

Granular Privacy Control for Geolocation with Vision Language Models

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, Alan Ritter

Vision Language Models (VLMs) are rapidly advancing in their capability to answer information-seeking questions. As these models are widely deployed in consumer applications, they could lead to new privacy risks due to emergent abilities to identify people in photos, geolocate images, etc. As we demonstrate, somewhat surprisingly, current open-source and proprietary VLMs are very capable image geolocators, making widespread geolocation with VLMs an immediate privacy risk, rather than merely a theoretical future concern. As a first step to address this challenge, we develop a new benchmark, GPTGeoChat, to test the capability of VLMs to moderate geolocation dialogues with users. We collect a set of 1,000 image geolocation conversations between in-house annotators and GPT-4v, which are annotated with the granularity of location information revealed at each turn. Using this new dataset we evaluate the ability of various VLMs to moderate GPT-4v geolocation conversations by determining when too much location information has been revealed. We find that custom fine-tuned models perform on par with prompted API-based models when identifying leaked location information at the country or city level, however fine-tuning on supervised data appears to be needed to accurately moderate finer granularities, such as the name of a restaurant or building.

15:00 - 15:15 - Flagler

Measuring Psychological Depth in Language Models

Fabrice Y Harel-Canaida, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Nanyun Peng, Amit Sahai

Evaluations of creative stories generated by large language models (LLMs) often focus on objective properties of the text, such as its style, coherence, and diversity. While these metrics are indispensable, they do not speak to a story's subjective, psychological impact from a reader's perspective. We introduce the Psychological Depth Scale (PDS), a novel framework rooted in literary theory that measures an LLM's ability to produce authentic and narratively complex stories that provoke emotion, empathy, and engagement. We empirically validate our framework by showing that humans can consistently evaluate stories based on PDS (0.72 Krippendorff's alpha). We also explore techniques for automating the PDS to easily scale future analyses. GPT-4o, combined with a novel Mixture-of-Personas (MoP) prompting strategy, achieves an average Spearman correlation of 0.51 with human judgment while Llama-3-70B with constrained decoding scores as high as 0.68 for empathy. Finally, we compared the depth of stories authored by both humans and LLMs. Surprisingly, GPT-4 stories either surpassed or were statistically indistinguishable from highly-rated human-written stories sourced from Reddit. By shifting the focus from text to reader, the Psychological Depth Scale is a validated, automated, and systematic means of measuring the capacity of LLMs to connect with humans through the stories they tell.

15:15 - 15:30 - Flagler

Step-by-Step Reasoning to Solve Grid Puzzles: Where do LLMs Falter?

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin RRV, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, Chitta Baral

Solving grid puzzles involves a significant amount of logical reasoning. Hence, it is a good domain to evaluate reasoning capability of a model which can then guide us to improve the reasoning ability of models. However, most existing works evaluate only the final predicted answer of a puzzle, without delving into an in-depth analysis of the LLMs' reasoning chains (such as where they falter) or providing any finer metrics to evaluate them. Since LLMs may rely on simple heuristics or artifacts to predict the final answer, it is crucial to evaluate the generated reasoning chain beyond overall correctness measures, for accurately evaluating the reasoning abilities of LLMs. To this end, we first develop GridPuzzle, an evaluation dataset comprising of 274 grid-based puzzles with different complexities. Second, we propose a new error taxonomy derived from manual analysis of reasoning chains from LLMs including GPT-4, Claude-3, Gemini, Mistral, and Llama-2. Then, we develop a LLM-based framework for large-scale subjective evaluation (i.e., identifying errors) and an objective metric, PuzzleEval,

to evaluate the correctness of reasoning chains. Evaluating reasoning chains from LLMs leads to several interesting findings. We further show that existing prompting methods used for enhancing models' reasoning abilities do not improve performance on GridPuzzle. This highlights the importance of understanding fine-grained errors and presents a challenge for future research to enhance LLMs' puzzle-solving abilities by developing methods that address these errors.

Generation 3

Nov 14 (Thu) 14:00-15:30 - Room: Monroe

14:00 - 14:15 - Monroe

Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation

Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, Aihoud Alhazmi

The distractor generation task focuses on generating incorrect but plausible options for objective questions such as fill-in-the-blank and multiple-choice questions. This task is widely utilized in educational settings across various domains and subjects. The effectiveness of these questions in assessments relies on the quality of the distractors, as they challenge examinees to select the correct answer from a set of misleading options. The evolution of artificial intelligence (AI) has transitioned the task from traditional methods to the use of neural networks and pre-trained language models. This shift has established new benchmarks and expanded the use of advanced deep learning methods in generating distractors. This survey explores distractor generation tasks, datasets, methods, and current evaluation metrics for English objective questions, covering both text-based and multi-modal domains. It also evaluates existing AI models and benchmarks and discusses potential future research directions.

14:15 - 14:30 - Monroe

ECIS-VQG: Generation of Entity-centric Information-seeking Questions from Videos

Arpan Phukan, Manish Gupta, Asif Ekbal

Previous studies on question generation from videos have mostly focused on generating questions about common objects and attributes and hence are not entity-centric. In this work, we focus on the generation of entity-centric information-seeking questions from videos. Such a system could be useful for video-based learning, recommending "People Also Ask" questions, video-based chatbots, and fact-checking. Our work addresses three key challenges: identifying question-worthy information, linking it to entities, and effectively utilizing multimodal signals. Further, to the best of our knowledge, there does not exist a large-scale dataset for this task. Most video question generation datasets are on TV shows, movies, or human activities or lack entity-centric information-seeking questions. Hence, we contribute a diverse dataset of YouTube videos, VideoQuestions, consisting of 411 videos with 2265 manually annotated questions. We further propose a model architecture combining Transformers, rich context signals (titles, transcripts, captions, embeddings), and a combination of cross-entropy and contrastive loss function to encourage entity-centric question generation. Our best method yields BLEU, ROUGE, CIDEr, and METEOR scores of 71.3, 78.6, 7.31, and 81.9, respectively, demonstrating practical usability. We make the code and dataset publicly available.

14:30 - 14:45 - Monroe

Evaluating LLMs' Capability in Satisfying Lexical Constraints

Binxuan Li, Yiyue Wang, Tao Meng, Nanyun Peng, Kai-Wei Chang

This paper investigates the capability of LLMs on controllable generation with prompt-based controlling, focusing on Lexically Constrained Generation (LCG). We systematically evaluate the performance of LLMs on satisfying lexical constraints with prompt-based controlling, as well as their efficacy in downstream applications. We identified three key reasons that highlight the limitations of LLMs in LCG, including (1) position bias, where LLMs tend to satisfy constraints that appear in specific positions within the input; (2) low responsiveness to control decoding parameters, which minimally impact the performance of LLMs; and (3) struggle with handling the inherent complexity of certain constraints (e.g. compound word). We conclude that black-box LLMs face significant challenges in consistently satisfying lexical constraints with prompt-based controlling. To address this bottleneck, we introduce the Divide and Conquer Generation strategy, effective for both white-box and black-box LLMs, to enhance LLMs performance in LCG tasks, which demonstrates over 90% improvement on success rate in the most challenging LCG task. Our analysis aims to provide valuable insights into the performance of LLMs in LCG with prompt-based controlling, and our proposed strategy offers a pathway to more sophisticated and customized text generation applications.

14:45 - 15:00 - Monroe

Improving Minimum Bayes Risk Decoding with Multi-Prompt

David Heineman, Yao Dou, Wei Xu

While instruction fine-tuned LLMs are effective text generators, sensitivity to prompt construction makes performance unstable and sub-optimal in practice. Relying on a single 'best' prompt cannot capture all differing approaches to a generation problem. Using this observation, we propose multi-prompt decoding, where many candidate generations are decoded from a prompt bank at inference-time. To ensemble candidates, we use Minimum Bayes Risk (MBR) decoding, which selects a final output using a trained value metric. We show multi-prompt improves MBR across a comprehensive set of conditional generation tasks, and show this is a result of estimating a more diverse and higher quality candidate space than that of a single prompt. Our experiments confirm multi-prompt improves generation across tasks, models and metrics.

15:00 - 15:15 - Monroe

Pron vs Prompt: Can Large Language Models already Challenge a World-Class Fiction Author at Creative Text Writing?

Guillermo Marco, Julio Gonzalo, M.Teresa Mateo-Girona, Ramón del Castillo Santos

Are LLMs ready to compete in creative writing skills with a top (rather than average) novelist? To provide an initial answer for this question, we have carried out a contest between Patricio Pron (an awarded novelist, considered one of the best of his generation) and GPT-4 (one of the top performing LLMs), in the spirit of AI-human duels such as DeepBlue vs Kasparov and AlphaGo vs Lee Sidol. We asked Pron and GPT-4 to provide thirty titles each, and then to write short stories for both their titles and their opponent's. Then, we prepared an evaluation rubric inspired by Boden's definition of creativity, and we collected several detailed expert assessments of the texts, provided by literature critics and scholars. The results of our experimentation indicate that LLMs are still far from challenging a top human creative writer. We also observed that GPT-4 writes more creatively using Pron's titles than its own titles (which is an indication of the potential for human-machine co-creation). Additionally, we found that GPT-4 has a more creative writing style in English than in Spanish.

15:15 - 15:30 - Monroe

Text2Char31: Instruction Tuning for Chart Generation with Automatic Feedback

Fatemeh Pesarman zadeh, Juyeon Kim, Jin-Hwa Kim, Gunhee Kim

Large language models (LLMs) have demonstrated strong capabilities across various language tasks, notably through instruction-tuning meth-

ods. However, LLMs face challenges in visualizing complex, real-world data through charts and plots. Firstly, existing datasets rarely cover a full range of chart types, such as 3D, volumetric, and gridded charts. Secondly, supervised fine-tuning methods do not fully leverage the intricate relationships within rich datasets, including text, code, and figures. To address these challenges, we propose a hierarchical pipeline and a new dataset for chart generation. Our dataset, Text2Chart31, includes 31 unique plot types referring to the Matplotlib library, with 11.1K tuples of descriptions, code, data tables, and plots. Moreover, we introduce a reinforcement learning-based instruction tuning technique for chart generation tasks without requiring human feedback. Our experiments show that this approach significantly enhances the model performance, enabling smaller models to outperform larger open-source models and be comparable to state-of-the-art proprietary models in data visualization tasks.

Machine Learning for NLP 4

Nov 14 (Thu) 14:00-15:30 - Room: Tuttle

14:00 - 14:15 - Tuttle

Fine-Tuning and Prompt Optimization: Two Good Steps that Work Better Together

Dilara Soylu, Christopher Potts, Omar Khattab

Natural Language Processing (NLP) systems are increasingly taking the form of sophisticated modular pipelines, e.g., Retrieval Augmented Generation (RAG), where each module may involve a distinct Language Model (LM) and an associated prompt template. These compound systems often lack intermediate labels or gradient flow to optimize each module, making their end-to-end optimization challenging. Here we seek strategies to optimize both the module-level LM weights and the associated prompt templates of such systems to maximize a downstream task metric. We propose for the first time combining the weight and prompt optimization strategies to optimize a modular LM pipeline by alternating between the two to get the same LM to teach itself. In experiments with multi-hop QA, mathematical reasoning, and feature-based classification using mistral-7b, llama-2-7b, and llama-3-8b, these BetterTogether strategies optimizing the weights and prompts of a pipeline together outperform directly optimizing weights alone and prompts alone by up to 60% and 6%, respectively, on average across LMs and tasks. Our BetterTogether optimizer is released in DSPY at <http://dspy.ai>.

14:15 - 14:30 - Tuttle

Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models

Sander Land, Max Bartolo

The disconnect between tokenizer creation and model training in language models allows for specific inputs, such as the infamous SolidGold-Magikarp token, to induce unwanted model behaviour. Although such ‘glitch tokens’, tokens present in the tokenizer vocabulary but that are nearly or entirely absent during model training, have been observed across various models, a reliable method to identify and address them has been missing. We present a comprehensive analysis of Large Language Model tokenizers, specifically targeting this issue of detecting under-trained tokens. Through a combination of tokenizer analysis, model weight-based indicators, and prompting techniques, we develop novel and effective methods for automatically detecting these problematic tokens. Our findings demonstrate the prevalence of such tokens across a diverse set of models and provide insights into improving the efficiency and safety of language models.

14:30 - 14:45 - Tuttle

Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, Michael Oberst

Several recent works seek to develop foundation models specifically for medical applications, adapting general-purpose large language models (LLMs) and vision-language models (VLMs) via continued pretraining on publicly available biomedical corpora. These works typically claim that such domain-adaptive pretraining (DAPT) improves performance on downstream medical tasks, such as answering medical licensing exam questions. In this paper, we compare seven public “medical” LLMs and two VLMs against their corresponding base models, arriving at a different conclusion: all medical VLMs and nearly all medical LLMs fail to consistently improve over their base models in the zero-/few-shot prompting regime for medical question-answering (QA) tasks. For instance, across the tasks and model pairs we consider in the 3-shot setting, medical LLMs only outperform their base models in 12.1% of cases, reach a (statistical) tie in 49.8% of cases, and are significantly worse than their base models in the remaining 38.2% of cases. Our conclusions are based on (i) comparing each medical model head-to-head, directly against the corresponding base model; (ii) optimizing the prompts for each model separately; and (iii) accounting for statistical uncertainty in comparisons. While these basic practices are not consistently adopted in the literature, our ablations show that they substantially impact conclusions. Our findings suggest that state-of-the-art general-domain models may already exhibit strong medical knowledge and reasoning capabilities, and offer recommendations to strengthen the conclusions of future studies.

14:45 - 15:00 - Tuttle

Not Eliminate but Aggregate: Post-Hoc Control over Mixture-of-Experts to Address Shortcut Shifts in Natural Language Understanding

Ukyo Honda, Tatsushi Oka, Peinan Zhang, Masato Mita

Recent models for natural language understanding are inclined to exploit simple patterns in datasets, commonly known as shortcuts. These shortcuts hinge on spurious correlations between labels and latent features existing in the training data. At inference time, shortcut-dependent models are likely to generate erroneous predictions under distribution shifts, particularly when some latent features are no longer correlated with the labels. To avoid this, previous studies have trained models to eliminate the reliance on shortcuts. In this study, we explore a different direction: pessimistically aggregating the predictions of a mixture-of-experts, assuming each expert captures relatively different latent features. The experimental results demonstrate that our post-hoc control over the experts significantly enhances the model’s robustness to the distribution shift in shortcuts. Besides, we show that our approach has some practical advantages. We also analyze our model and provide results to support the assumption.

15:00 - 15:15 - Tuttle

Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs

Krista Opsahl-Ong, Michael J Ryan, Josh Purcell, David Bromman, Christopher Potts, Matei Zaharia, Omar Khattab

Language Model Programs, i.e. sophisticated pipelines of modular language model (LM) calls, are increasingly advancing NLP tasks, but they require crafting prompts that are jointly effective for all modules. We study prompt optimization for LM programs, i.e. how to update these prompts to maximize a downstream metric without access to module-level labels or gradients. To make this tractable, we factorize our problem into optimizing the free-form instructions and few-shot demonstrations of every module and introduce several strategies to craft task-grounded instructions and navigate credit assignment across modules. Our strategies include (i) program- and data-aware techniques for proposing effective instructions, (ii) a stochastic mini-batch evaluation function for learning a surrogate model of our objective, and (iii) a meta-optimization procedure in which we refine how LMs construct proposals over time. Using these insights we develop MIPRO, a novel

algorithm for optimizing LM programs. MIPRO outperforms baseline optimizers on five of seven diverse multi-stage LM programs using a best-in-class open-source model (Llama-3-8B), by as high as 13% accuracy. We have released our new optimizers and benchmark in DSPy at <http://dspy.ai>.

15:15 - 15:30 - Tuttle

Temporally Consistent Factuality Probing for Large Language Models

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, Tannoy Chakraborty

The prolific use of Large Language Models (LLMs) as an alternate knowledge base requires them to be factually consistent, necessitating both correctness and consistency traits for paraphrased queries. Recently, significant attempts have been made to benchmark datasets and metrics to evaluate LLMs for these traits. However, structural simplicity (subject-relation-object) and contemporary association in their query formulation limit the broader definition of factuality and consistency. In this study, we introduce TeCFaP, a novel Temporally Consistent Factuality Probe task to expand the consistent factuality probe in the temporal dimension. To this end, we propose TEMP-COFAC, a high-quality dataset of prefix-style English query paraphrases. Subsequently, we extend the definitions of existing metrics to represent consistent factuality across temporal dimension. We experiment with a diverse set of LLMs and find most of them performing poorly on TeCFaP. Next, we propose a novel solution CoTSeLF (Consistent-Time-Sensitive Learning Framework) combining multi-task instruction tuning (MT-IT) with consistent-time-sensitive reinforcement learning (CTSRL) to improve temporally consistent factuality in LLMs. Our experiments demonstrate the efficacy of CoTSeLF over several baselines.

11

Posters and Demos

Session 02 - Nov 12 (Tue) 11:00-12:30

Demo

Nov 12 (Tue) 11:00-12:30 - Room: Riverfront Hall

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

RETAIN: Interactive Tool for Regression Testing Guided LLM Migration

Akash V Maharaaj, Anirudh Sureshan, Daniel Lee, Sai Sree Harsha, Sally Fang, Tanay Dixit, Yunyao Li

Large Language Models (LLMs) are increasingly integrated into diverse applications. The rapid evolution of LLMs presents opportunities for developers to enhance applications continuously. However, this constant adaptation can also lead to performance regressions during model migrations. While several interactive tools have been proposed to streamline the complexity of prompt engineering, few address the specific requirements of regression testing for LLM Migrations. To bridge this gap, we introduce RETAIN (REgression Testing guided LLM migration), a tool designed explicitly for regression testing in LLM Migrations. RETAIN comprises two key components: an interactive interface tailored to regression testing needs during LLM migrations, and an error discovery module that facilitates understanding of differences in model behaviors. The error discovery module generates textual descriptions of various errors or differences between model outputs, providing actionable insights for prompt refinement. Our automatic evaluation and empirical user studies demonstrate that RETAIN, when compared to manual evaluation, enabled participants to identify twice as many errors, facilitated experimentation with 75% more prompts, and achieves 12% higher metric scores in a given time frame.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LLM-DetectAIve: A Tool for Fine-Grained Machine-Generated Text Detection

Akim Tsvigun, Alexander Aziz, Alham Fikri Aji, Artem Shelmanov, Bimarsha Adhikari, Hasan Iqbal, Iryna Gurevych, Jiahui Geng, Jonibek Mansurov, Kareem Elzeirî, Mervat Abusy, Minh Ngoc Ta, Nizar Habash, OSAMA MOHAMMED AFZAL, Preslav Nakov, Raj Vardhan Tomar, Rui Xing, Saad El Dine Ahmed, Tarek Mahmoud, Vladislav Mikhailov, Yuxia Wang, Zain Muhammad Mujahid, Zhuohan Xie, Katya Artemova

The widespread accessibility of large language models (LLMs) to the general public has significantly amplified the dissemination of machine-generated texts (MGTs). Advancements in prompt manipulation have exacerbated the difficulty in discerning the origin of a text (human-authored vs machine-generated). This raises concerns regarding the potential misuse of MGTs, particularly within educational and academic domains. In this paper, we present **LLM-DetectAIve** – a system designed for fine-grained MGT detection. It is able to classify texts into four categories: human-written, machine-generated, machine-written machine-humanized, and human-written machine-polished. Contrary to previous MGT detectors that perform binary classification, introducing two additional categories in LLM-DetectAIve offers insights into the varying degrees of LLM intervention during the text creation. This might be useful in some domains like education, where any LLM intervention is usually prohibited. Experiments show that LLM-DetectAIve can effectively identify the authorship of textual content, proving its usefulness in enhancing integrity in education, academia, and other domains. LLM-DetectAIve is publicly accessible at <https://hugging-face.co/spaces/raj-tomar001/MGT-New>. The video describing our system is available at https://youtu.be/E8cT_bE7k8c.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation

Maosong Sun, Qinyu Luo, Shihao Liang, Xiaoyin Che, Xin Cong, Yankai Lin, Yaxi Lu, Yesai Wu, Yingli Zhang, Yujia Qin, Zhuyuan Liu, Zhong Zhang, Ye Yi ning

Generative models have demonstrated considerable potential in software engineering, particularly in tasks such as code generation and debugging. However, their utilization in the domain of code documentation generation remains underexplored. To this end, we introduce RepoAgent, a large language model powered open-source framework aimed at proactively generating, maintaining, and updating code documentation. Through both qualitative and quantitative evaluations, we have validated the effectiveness of our approach, showing that RepoAgent excels in generating high-quality repository-level documentation. The code and results are publicly accessible at <https://github.com/OpenBM-B/RepoAgent>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LM-Interview: An Easy-to-use Smart Interviewer System via Knowledge-guided Language Model Exploitation

Bin Xu, Hanming Li, Jiaxi Yuan, Jifan Yu, Juanzi Li, Ruimiao Li, Yan Xuan, Zhanxin Hao, Zhiyuan Liu

Semi-structured interviews are a crucial method of data acquisition in qualitative research. Typically controlled by the interviewer, the process progresses through a question-and-answer format, aimed at eliciting information from the interviewee. However, interviews are highly time-consuming and demand considerable experience of the interviewers, which greatly limits the efficiency and feasibility of data collection. Therefore, we introduce LM-Interview, a novel system designed to automate the process of preparing, conducting and analyzing semi-structured interviews. Experimental results demonstrate that LM-interview achieves performance comparable to that of skilled human interviewers.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

FreeEval: A Modular Framework for Trustworthy and Efficient Evaluation of Large Language Models

Chang Gao, Jindong Wang, Shikun Zhang, Wei Ye, Wenjin Yao, Yidong Wang, Yue Zhang, Zhuohao Yu, ZHENGRAN ZENG

The rapid growth of evaluation methodologies and datasets for large language models (LLMs) has created a pressing need for their unified integration. Meanwhile, concerns about data contamination and bias compromise the trustworthiness of evaluation findings, while the efficiency of evaluation processes remains a bottleneck due to the significant computational costs associated with LLM inference. In response to these challenges, we introduce FreeEval, a modular framework not only for conducting trustworthy and efficient automatic evaluations of LLMs but also serving as a platform to develop and validate new evaluation methodologies. FreeEval addresses key challenges through: (1) unified abstractions that simplify the integration of diverse evaluation methods, including dynamic evaluations requiring complex LLM interactions; (2) built-in meta-evaluation techniques such as data contamination detection and human evaluation to enhance result fairness; (3) a high-performance infrastructure with distributed computation and caching strategies for efficient large-scale evaluations; and (4) an interactive Visualizer for result analysis and interpretation to support innovation of evaluation techniques. We open-source all our code at <https://github.com/WisdomShell/FreeEval> and our demostration video, live demo, installation guides are available at: <https://freeeval.zhuohao.me/>.

Generation 1

Nov 12 (Tue) 11:00-12:30 - Room: Riverfront Hall

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A Survey on Natural Language Counterfactual Generation

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, Zhiqi Shen

Natural language counterfactual generation aims to minimally modify a given text such that the modified text will be classified into a different class. The generated counterfactuals provide insight into the reasoning behind a model's predictions by highlighting which words significantly influence the outcomes. Additionally, they can be used to detect model fairness issues and augment the training data to enhance the model's robustness. A substantial amount of research has been conducted to generate counterfactuals for various NLP tasks, employing different models and methodologies. With the rapid growth of studies in this field, a systematic review is crucial to guide future researchers and developers. To bridge this gap, this survey provides a comprehensive overview of textual counterfactual generation methods, particularly those based on Large Language Models. We propose a new taxonomy that systematically categorizes the generation methods into four groups and summarizes the metrics for evaluating the generation quality. Finally, we discuss ongoing research challenges and outline promising directions for future work.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

BlendFilter: Advancing Retrieval-Augmented Large Language Models via Query Generation Blending and Knowledge Filtering

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, chen luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, Jing Gao

Retrieval-augmented Large Language Models (LLMs) offer substantial benefits in enhancing performance across knowledge-intensive scenarios. However, these methods often struggle with complex inputs and encounter difficulties due to noisy knowledge retrieval, notably hindering model effectiveness. To address this issue, we introduce BlendFilter, a novel approach that elevates retrieval-augmented LLMs by integrating query generation blending with knowledge filtering. BlendFilter proposes the blending process through its query generation method, which integrates both external and internal knowledge augmentation with the original query, ensuring comprehensive information gathering. Additionally, our distinctive knowledge filtering module capitalizes on the intrinsic capabilities of the LLM, effectively eliminating extraneous data. We conduct extensive experiments on three open-domain question answering benchmarks, and the findings clearly indicate that our innovative BlendFilter surpasses state-of-the-art baselines significantly.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Efficient Retriever for Multi-Hop Retrieval Question Answering

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang

Retrieval-augmented generation (RAG) methods encounter difficulties when addressing complex questions like multi-hop queries. While iterative retrieval methods improve performance by gathering additional information, current approaches often rely on multiple calls of large language models (LLMs). In this paper, we introduce EfficientRAG, an efficient retriever for multi-hop question answering. EfficientRAG iteratively generates new queries without the need for LLM calls at each iteration and filters out irrelevant information. Experimental results demonstrate that EfficientRAG surpasses existing RAG methods on three open-domain multi-hop question-answering datasets. The code is available in <https://aka.ms/efficientrag> (<https://github.com/NIL-zhuang/EfficientRAG-official>).

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

StyleRemix: Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, Yejin Choi

Authorship obfuscation, rewriting a text to intentionally obscure the identity of the author, is important yet challenging. Current methods using large language models (LLMs) lack interpretability and controllability, often ignoring author-specific stylistic features, resulting in less robust performance overall. To address this, we develop StyleRemix, an adaptive and interpretable obfuscation method that perturbs specific, fine-grained style elements of the original input text. StyleRemix uses pre-trained Low Rank Adaptation (LoRA) modules to rewrite inputs along various stylistic axes (e.g., formality, length) while maintaining low computational costs. StyleRemix outperforms state-of-the-art baselines and much larger LLMs on an array of domains on both automatic and human evaluation. Additionally, we release AuthorMix, a large set of 30K high-quality, long-form texts from a diverse set of 14 authors and 4 domains, and DiSC, a parallel corpus of 1,500 texts spanning seven style axes in 16 unique directions.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Satyrn: A Platform for Analytics Augmented Generation

Marko Sterbenz, Cameron Barrie, Shubham Shahi, Abhratanu Dutta, Donna Hooshmand, Harper Pack, Kristian J Hammond

Large language models (LLMs) are capable of producing documents, and retrieval augmented generation (RAG) has shown itself to be a powerful method for improving accuracy without sacrificing fluency. However, not all information can be retrieved from text. We propose an approach that uses the analysis of structured data to generate fact sets that are used to guide generation in much the same way that retrieved documents are used in RAG. This analytics augmented generation (AAG) approach supports the ability to utilize standard analytic techniques to generate facts that are then converted to text and passed to an LLM. We present a neurosymbolic platform, Satyrn, that leverages AAG to produce accurate, fluent, and coherent reports grounded in large scale databases. In our experiments, we find that Satyrn generates reports in which over 86% of claims are accurate while maintaining high levels of fluency and coherence, even when using smaller language models such as Mistral-7B, as compared to GPT-4 Code Interpreter in which just 57% of claims are accurate.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Detection and Measurement of Syntactic Templates in Generated Text

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, Byron C Wallace

The diversity of text can be measured beyond word-level features, however existing diversity evaluation focuses primarily on word-level features. Here we propose a method for evaluating diversity over syntactic features to characterize general repetition in models, beyond frequent n -grams. Specifically, we define *syntactic templates* (e.g., strings comprising parts-of-speech) and show that models tend to produce templated text in downstream tasks at a higher rate than what is found in human-reference texts. We find that most (76%) templates in model-generated text can be found in pre-training data (compared to only 35% of human-authored text), and are not overwritten during fine-tuning or alignment processes such as RLHF. The connection between templates in generated text and the pre-training data allows us to analyze syntactic templates in models where we do not have the pre-training data. We also find that templates as features are able to differentiate between models, tasks, and domains, and are useful for qualitatively evaluating common model constructions. Finally, we demonstrate the use of templates as a useful tool for analyzing style memorization of training data in LLMs.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Automatic Instruction Evolving for Large Language Models

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, Weizhu Chen

Fine-tuning large pre-trained language models with Evol-Instruct has achieved encouraging results across a wide range of tasks. However, designing effective evolving methods for instruction evolution requires substantial human expertise. This paper proposes Auto Evol-Instruct, an end-to-end framework that evolves instruction datasets using large language models without any human effort. The framework automatically analyzes and summarizes suitable evolutionary strategies for the given instruction data and iteratively improves the evolving method based on issues exposed during the instruction evolution process. Our extensive experiments demonstrate that the best method optimized by Auto Evol-Instruct outperforms human-designed methods on various benchmarks, including MT-Bench, AlpacaEval, GSM8K, and HumanEval.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Improving Retrieval-augmented Text-to-SQL with AST-based Ranking and Schema Pruning

Zhilin Shen, Pavlos Vougiouklis, Chenxin Diao, Kaustubh Vyas, Yuanqi Ji, Jeff Z. Pan

We focus on Text-to-SQL semantic parsing from the perspective of retrieval-augmented generation. Motivated by challenges related to the size of commercial database schemata and the deployability of business intelligence solutions, we propose ASTReS that dynamically retrieves input database information and uses abstract syntax trees to select few-shot examples for in-context learning. Furthermore, we investigate the extent to which an in-parallel semantic parser can be leveraged for generating approximated versions of the expected SQL queries, to support our retrieval. We take this approach to the extreme—we adapt a model consisting of less than 500M parameters, to act as an extremely efficient approximator, enhancing it with the ability to process schemata in a parallelized manner. We apply ASTReS to monolingual and cross-lingual benchmarks for semantic parsing, showing improvements over state-of-the-art baselines. Comprehensive experiments highlight the contribution of modules involved in this retrieval-augmented generation setting, revealing interesting directions for future work.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A Thorough Examination of Decoding Methods in the Era of LLMs

Chufan Shi, HAORAN YANG, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, Wai Lam

Decoding methods play an indispensable role in converting language models from next-token predictors into practical task solvers. Prior research on decoding methods, primarily focusing on task-specific models, may not extend to the current era of general-purpose large language models (LLMs). Moreover, the recent influx of decoding strategies has further complicated this landscape. This paper provides a comprehensive and multifaceted analysis of various decoding methods within the context of LLMs, evaluating their performance, robustness to hyperparameter changes, and decoding speeds across a wide range of tasks, models, and deployment environments. Our findings reveal that decoding method performance is notably task-dependent and influenced by factors such as alignment, model size, and quantization. Intriguingly, sensitivity analysis exposes that certain methods achieve superior performance at the cost of extensive hyperparameter tuning, highlighting the trade-off between attaining optimal results and the practicality of implementation in varying contexts.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Learning to Correct for QA Reasoning with Black-box LLMs

Jaehyung Kim, Dongyoung Kim, Yiming Yang

An open challenge in recent machine learning is about how to improve the reasoning capability of large language models (LLMs) in a black-box setting, i.e., without access to detailed information such as output token probabilities. Existing approaches either rely on accessibility (which is often unrealistic) or involve significantly increased train- and inference-time costs. This paper addresses those limitations or shortcomings by proposing a novel approach, namely CoBB (Correct for improving QA reasoning of Black-Box LLMs). It uses a trained adaptation model to perform a seq2seq mapping from the often-imperfect reasonings of the original black-box LLM to the correct or improved reasonings. Specifically, the adaptation model is initialized with a relatively small open-source LLM and adapted over a collection of sub-sampled training pairs. To select the representative pairs of correct and incorrect reasonings, we formulated the dataset construction as an optimization problem that minimizes the statistical divergence between the sampled subset and the entire collection, and solved it via a genetic algorithm. We then train the adaptation model over the sampled pairs by contrasting the likelihoods of correct and incorrect reasonings. Our experimental results demonstrate that CoBB significantly improves reasoning accuracy across various QA benchmarks, compared to the best-performing adaptation baselines.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

PostMark: A Robust Blackbox Watermark for Large Language Models

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, Mohit Iyyer

The most effective techniques to detect LLM-generated text rely on inserting a detectable signature—or watermark—during the model’s decoding process. Most existing watermarking methods require access to the underlying LLM’s logits, which LLM API providers are loath to share due to fears of model distillation. As such, these watermarks must be implemented independently by each LLM provider. In this paper,

we develop PostMark, a modular post-hoc watermarking procedure in which an input-dependent set of words (determined via a semantic embedding) is inserted into the text after the decoding process has completed. Critically, PostMark does not require logit access, which means it can be implemented by a third party. We also show that PostMark is more robust to paraphrasing attacks than existing watermarking methods: our experiments cover eight baseline algorithms, five base LLMs, and three datasets. Finally, we evaluate the impact of PostMark on text quality using both automated and human assessments, highlighting the trade-off between quality and robustness to paraphrasing. We release our code, outputs, and annotations at <https://github.com/lilakk/PostMark>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Assessing Implicit Retrieval Robustness of Large Language Models

Xiaoyu Shen, Rishabh Iyer, Dawei Zhu, Jiahuan Pei, Wei Zhang

Retrieval-augmented generation has gained popularity as a framework to enhance large language models with external knowledge. However, its effectiveness hinges on the retrieval robustness of the model. If the model lacks retrieval robustness, its performance is constrained by the accuracy of the retriever, resulting in significant compromises when the retrieved context is irrelevant. In this paper, we evaluate the “implicit” retrieval robustness of various large language models, instructing them to directly output the final answer without explicitly judging the relevance of the retrieved context. Our findings reveal that fine-tuning on a mix of gold and distracting context significantly enhances the model’s robustness to retrieval inaccuracies, while still maintaining its ability to extract correct answers when retrieval is accurate. This suggests that large language models can implicitly handle relevant or irrelevant retrieved context by learning solely from the supervision of the final answer in an end-to-end manner. Introducing an additional process for explicit relevance judgment can be unnecessary and disrupts the end-to-end approach.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CleanGen: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, Radha Poovendran

The remarkable performance of large language models (LLMs) in generation tasks has enabled practitioners to leverage publicly available models to power custom applications, such as chatbots and virtual assistants. However, the data used to train or fine-tune these LLMs is often undisclosed, allowing an attacker to compromise the data and inject backdoors into the models. In this paper, we develop a novel inference time defense, named CleanGen, to mitigate backdoor attacks for generation tasks in LLMs. CleanGen is a lightweight and effective decoding strategy that is compatible with the state-of-the-art (SOTA) LLMs. Our insight behind CleanGen is that compared to other LLMs, backdoored LLMs assign significantly higher probabilities to tokens representing the attacker-desired contents. These discrepancies in token probabilities enable CleanGen to identify suspicious tokens favored by the attacker and replace them with tokens generated by another LLM that is not compromised by the same attacker, thereby avoiding generation of attacker-desired content. We evaluate CleanGen against five SOTA backdoor attacks. Our results show that CleanGen achieves lower attack success rates (ASR) compared to five SOTA baseline defenses for all five backdoor attacks. Moreover, LLMs deploying CleanGen maintain helpfulness in their responses when serving benign user queries with minimal added computational overhead.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Enhancing Reinforcement Learning with Intrinsic Rewards from Language Model Critique

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichters, Yinxiao Liu, Lei Meng

Reinforcement learning (RL) can align language models with non-differentiable reward signals, such as human preferences. However, a major challenge arises from the sparsity of these reward signals – typically, there is only a single reward for an entire output. This sparsity of rewards can lead to inefficient and unstable learning. To address this challenge, our paper introduces a novel framework that utilizes the critique capability of Large Language Models (LLMs) to produce intermediate-step rewards during RL training. Our method involves coupling a policy model with a critic language model, which is responsible for providing comprehensive feedback of each part of the output. This feedback is then translated into token- or span-level rewards that can be used to guide the RL training process. We investigate this approach under two different settings: one where the policy model is smaller and is paired with a more powerful critic model, and another where a single language model fulfills both roles. We assess our approach on three text generation tasks: sentiment control, language model detoxification, and summarization. Experimental results show that incorporating artificial intrinsic rewards significantly improve both sample efficiency and the overall performance of the policy model, supported by both automatic and human evaluation.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Synchronous Faithfulness Monitoring for Trustworthy Retrieval-Augmented Generation

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, Kai-Wei Chang

Retrieval-augmented language models (RALMs) have shown strong performance and wide applicability in knowledge-intensive tasks. However, there are significant trustworthiness concerns as RALMs are prone to generating unfaithful outputs, including baseless information or contradictions with the retrieved context. This paper proposes SynCheck, a lightweight monitor that leverages fine-grained decoding dynamics including sequence likelihood, uncertainty quantification, context influence, and semantic alignment to synchronously detect unfaithful sentences. By integrating efficiently measurable and complementary signals, SynCheck enables accurate and immediate feedback and intervention. Experiments show that SynCheck significantly outperforms existing faithfulness detection baselines, achieving over 0.85 AUCROC across a suite of six long-form retrieval-augmented generation tasks. Leveraging SynCheck, we further introduce FOD, a faithfulness-oriented decoding algorithm guided by beam search for long-form retrieval-augmented generation. Empirical results demonstrate that FOD outperforms traditional strategies such as abstention, reranking, or contrastive decoding significantly in terms of faithfulness, achieving over 10% improvement across six datasets.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Enhancing Pre-Trained Generative Language Models with Question Attended Span Extraction on Machine Reading Comprehension

Lin Ai, Zheng Hui, Zichao Liu, Julia Hirschberg

Machine Reading Comprehension (MRC) poses a significant challenge in the field of Natural Language Processing (NLP). While mainstream MRC methods predominantly leverage extractive strategies using encoder-only models such as BERT, generative approaches face the issue of *out-of-control generation* – a critical problem where answers generated are often incorrect, irrelevant, or unfaithful to the source text. To address these limitations in generative models for extractive MRC, we introduce the Question-Attended Span Extraction (*QASE*) module. Integrated during the fine-tuning phase of pre-trained generative language models (PLMs), *QASE* significantly enhances their performance, allowing them to surpass the extractive capabilities of advanced Large Language Models (LLMs) such as GPT-4 in few-shot settings. Notably, these gains in performance do not come with an increase in computational demands. The efficacy of the *QASE* module has been rigorously tested across various datasets, consistently achieving or even surpassing state-of-the-art (SOTA) results, thereby bridging the gap between generative and extractive models in extractive MRC tasks. Our code is available at this GitHub repository: <https://github.com/lynneai/QASE.git>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

ATM: Adversarial Tuning Multi-agent System Makes a Robust Retrieval-Augmented Generator

Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, Lei Sha

Large language models (LLMs) are proven to benefit a lot from retrieval-augmented generation (RAG) in alleviating hallucinations confronted with knowledge-intensive questions. RAG adopts information retrieval techniques to inject external knowledge from semantic-relevant documents as input contexts. However, due to today's Internet being flooded with numerous noisy and fabricating content, it is inevitable that RAG systems are vulnerable to these noises and prone to respond incorrectly. To this end, we propose to optimize the retrieval-augmented Generator with a Adversarial Tuning Multi-agent system **(ATM)**. The ATM steers the Generator to have a robust perspective of useful documents for question answering with the help of an auxiliary Attacker agent. The Generator and the Attacker are tuned adversarially for several iterations. After rounds of multi-agent iterative tuning, the Generator can eventually better discriminate useful documents amongst fabrications. The experimental results verify the effectiveness of ATM and we also observe that the Generator can achieve better performance compared to state-of-the-art baselines.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Puzzle Solving using Reasoning of Large Language Models: A Survey

Panagiotis Giadikoglou, Maria Lymperiou, Giorgos Filandrianos, Giorgos Stamou

Exploring the capabilities of Large Language Models (LLMs) in puzzle solving unveils critical insights into their potential and challenges in AI, marking a significant step towards understanding their applicability in complex reasoning tasks. This survey leverages a unique taxonomy/dividing puzzles into rule-based and rule-less categories to critically assess LLMs through various methodologies, including prompting techniques, neuro-symbolic approaches, and fine-tuning. Through a critical review of relevant datasets and benchmarks, we assess LLMs' performance, identifying significant challenges in complex puzzle scenarios. Our findings highlight the disparity between LLM capabilities and human-like reasoning, particularly in those requiring advanced logical inference. The survey underscores the necessity for novel strategies and richer datasets to advance LLMs' puzzle-solving proficiency and contribute to AI's logical reasoning and creative problem-solving advancements.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Atomic Self-Consistency for Better Long Form Generations

Raghavveer Thirukovalluru, Yukun Huang, Bhuvan Dhingra

Recent work has aimed to improve LLM generations by filtering out hallucinations, thereby improving the precision of the information in responses. Correctness of a long-form response, however, also depends on the recall of multiple pieces of information relevant to the question. In this paper, we introduce Atomic Self-Consistency (ASC), a technique for improving the recall of relevant information in an LLM response. ASC follows recent work, Universal Self-Consistency (USC) in using multiple stochastic samples from an LLM to improve the long-form response. Unlike USC which only focuses on selecting the best single generation, ASC picks authentic subparts from the samples and merges them into a superior composite answer. Through extensive experiments and ablations, we show that merging relevant subparts of multiple samples performs significantly better than picking a single sample. ASC demonstrates significant gains over USC on multiple factoids and open-ended QA datasets - ASQA, QAMPARI, QUEST, ELI5 with ChatGPT and Llama3. Our analysis also reveals untapped potential for enhancing long-form generations using the approach of merging multiple samples.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Analysis of Plan-based Retrieval for Grounded Text Generation

Amyea Godbole, Nicholas Monath, Seungyeon Kim, Ankit Singh Rawat, Andrew McCallum, Manzil Zaheer

In text generation, hallucinations refer to the generation of seemingly coherent text that contradicts established knowledge. One compelling hypothesis is that hallucinations occur when a language model is given a generation task outside its parametric knowledge (due to rarity, recency, domain, etc.). A common strategy to address this limitation is to infuse the language models with retrieval mechanisms, providing the model with relevant knowledge for the task. In this paper, we leverage the planning capabilities of instruction-tuned LLMs and analyze how planning can be used to guide retrieval to further reduce the frequency of hallucinations. We empirically evaluate several variations of our proposed approach on long-form text generation tasks. By improving the coverage of relevant facts, plan-guided retrieval and generation can produce more informative responses while providing a higher rate of attribution to source documents.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A SMART Mnemonic Sounds like "Glue Tonic": Mixing LLMs with Student Feedback to Make Mnemonic Learning Stick

Nishant Balepur, Matthew Shu, Alexander Hoyle, Alison Robey, Shi Feng, Seraphina Goldfarb-Tarrant, Jordan Lee Boyd-Graber

Keyword mnemonics are memorable explanations that link new terms to simpler keywords. Prior work generates mnemonics for students, but they do not train models using mnemonics students prefer and aid learning. We build SMART, a mnemonic generator trained on feedback from real students learning new terms. To train SMART, we first fine-tune LLaMA-2 on a curated set of user-written mnemonics. We then use LLM alignment to enhance SMART: we deploy mnemonics generated by SMART in a flashback app to find preferences on mnemonics students favor. We gather 2684 preferences from 45 students across two types: **expressed** (inferred from ratings) and **observed** (inferred from student learning), yielding three key findings. First, expressed and observed preferences disagree; what students *think* is helpful does not always capture what is *truly* helpful. Second, Bayesian models can synthesize complementary data from multiple preference types into a single effectiveness signal. SMART is tuned via Direct Preference Optimization on this signal, which resolves ties and missing labels in the typical method of pairwise comparisons, augmenting data for LLM output quality gains. Third, mnemonic experts assess SMART as matching GPT-4 at much lower deployment costs, showing the utility of capturing diverse student feedback to align LLMs in education.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector

Xiaoxue Cheng, Junyi Li, Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di ZHANG, Kun Gai, Ji-Rong Wen

Hallucination detection is a challenging task for large language models (LLMs), and existing studies heavily rely on powerful closed-source LLMs such as GPT-4. In this paper, we propose an autonomous LLM-based agent framework, called HaluAgent, which enables relatively smaller LLMs (e.g., Baichuan2-Chat 7B) to actively select suitable tools for detecting multiple hallucination types such as text, code, and mathematical expression. In HaluAgent, we integrate the LLM, multi-functional toolbox, and design a fine-grained three-stage detection framework along with memory mechanism. To facilitate the effectiveness of HaluAgent, we leverage existing Chinese and English datasets to synthesize detection trajectories for fine-tuning, which endows HaluAgent with the capability for bilingual hallucination detection. Extensive experiments demonstrate that only using 2K samples for tuning LLMs, HaluAgent can perform hallucination detection on various types of tasks and datasets, achieving performance comparable to or even higher than GPT-4 without tool enhancements on both in-domain and out-of-domain datasets.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A Fundamental Trade-off in Aligned Language Models and its Relation to Sampling Adaptors

Naaman Tan, Josef Valvoda, Tianyu Liu, Anej Svetec, Yanxia Qin, Min-Yen Kan, Ryan Cotterell

The relationship between the quality of a string, as judged by a human reader, and its probability, $p(y)$ under a language model undergirds the development of better language models. For example, many popular algorithms for sampling from a language model have been conceived with the goal of manipulating $p(y)$ to place higher probability on strings that humans deem of high quality. In this article, we examine the

probability-quality relationship in language models explicitly aligned to human preferences, e.g., through reinforcement learning through human feedback. We show that, when sampling corpora from an aligned language model, there exists a trade-off between the strings' average reward and average log-likelihood under the prior language model, i.e., the same model before alignment with human preferences. We provide a formal treatment of this phenomenon and demonstrate how a choice of sampling adaptor allows for a selection of how much likelihood we exchange for the reward.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, Xiaojun Wan

The evaluation of natural language generation (NLG) tasks is a significant and longstanding research area. With the recent emergence of powerful large language models (LLMs), some studies have turned to LLM-based automatic evaluation methods, which demonstrate great potential to become a new evaluation paradigm following traditional string-based and model-based metrics. However, despite the improved performance of existing methods, they still possess some deficiencies, such as dependency on references and limited evaluation flexibility. Therefore, in this paper, we meticulously construct a large-scale NLG evaluation corpus ^{**}NLG-Eval^{1*} with annotations from both human and GPT-4 to alleviate the lack of relevant data in this field. Furthermore, we propose ^{**}Themis^{2*}, an LLM dedicated to NLG evaluation, which has been trained with our designed multi-perspective consistency verification and rating-oriented preference alignment methods. Themis can conduct flexible and interpretable evaluations without references, and it exhibits superior evaluation performance on various NLG tasks, simultaneously generalizing well to unseen tasks and surpassing other evaluation models, including GPT-4.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CorrSynth - A Correlated Sampling Method for Diverse dataset Generation from LLMs

Abhishek Divakar, Suhas S Kowshik, Vijit Malik

Large language models (LLMs) have demonstrated remarkable performance in diverse tasks using zero-shot and few-shot prompting. Even though their capabilities of data synthesis have been studied well in recent years, the generated data suffers from a lack of diversity, less adherence to the prompt, and potential biases that creep into the data from the generator model. In this work, we tackle the challenge of generating datasets with high diversity, upon which a student model is trained for downstream tasks. Taking the route of decoding-time guidance-based approaches, we propose CorrSynth, which generates data that is more diverse and faithful to the input prompt using a correlated sampling strategy. Further, our method overcomes the complexity drawbacks of some other guidance-based techniques like classifier-based guidance. With extensive experiments, we show the effectiveness of our approach and substantiate our claims. In particular, we perform intrinsic evaluation to show the improvements in diversity. Our experiments show that CorrSynth improves both student metrics and intrinsic metrics upon competitive baselines across four datasets, showing the innate advantage of our method.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, Yun-Hsuan Sung

As large language models (LLMs) evolve, evaluating their output reliably becomes increasingly difficult due to the high cost of human evaluation. To address this, we introduce FLAME, a family of Foundational Large Autorater Models. FLAME is trained on a diverse set of over 100 quality assessment tasks, incorporating 5M+ human judgments curated from publicly released human evaluations. FLAME outperforms models like GPT-4 and Claude-3 on various held-out tasks, and serves as a powerful starting point for fine-tuning, as shown in our reward model evaluation case study (FLAME-RM). On Reward-Bench, FLAME-RM-24B achieves 87.8% accuracy, surpassing GPT-4-0125 (85.9%) and GPT-4o (84.7%). Additionally, we introduce FLAME-Opt-RM, an efficient tail-patch fine-tuning approach that offers competitive RewardBench performance using 25xE fewer training datapoints. Our FLAME variants outperform popular proprietary LLM-as-a-Judge models on 8 of 12 autorater benchmarks, covering 53 quality assessment tasks, including RewardBench and LLM-AggrFact. Finally, our analysis shows that FLAME is significantly less biased than other LLM-as-a-Judge models on the CoBBLEr autorater bias benchmark.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Don't Forget Your Reward Values: Language Model Alignment via Value-based Calibration

Xin Mao, Feng-Lin Li, Huijin Xu, Wei Zhang, WANG CHEN, Anh Tuan Luu

While Reinforcement Learning from Human Feedback (RLHF) significantly enhances the generation quality of Large Language Models (LLMs), recent studies have raised concerns regarding the complexity and instability associated with the Proximal Policy Optimization (PPO) algorithm, proposing a series of order-based alignment methods as viable alternatives. This paper delves into existing order-based methods, unifying them into one framework and examining their inefficiencies in utilizing reward values. Building upon these findings, we propose a new Value-based Calibration (VCB) method to better align LLMs with human preferences. Experimental results demonstrate that VCB surpasses existing alignment methods on AI assistant and summarization datasets, providing impressive generalizability, robustness, and diversity in different settings.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Label Confidence Weighted Learning for Target-level Sentence Simplification

Jingshen Zhang, Xin Ying Qiu

Multi-level sentence simplification generates simplified sentences with varying language proficiency levels. We propose Label Confidence Weighted Learning (LCWL), a novel approach that incorporates a label confidence weighting scheme in the training loss of the encoder-decoder model, setting it apart from existing confidence-weighting methods primarily designed for classification. Experimentation on English grade-level simplification dataset shows that LCWL outperforms state-of-the-art unsupervised baselines. Fine-tuning the LCWL model on in-domain data and combining with Symmetric Cross Entropy (SCE) consistently delivers better simplifications compared to strong supervised methods. Our results highlight the effectiveness of label confidence weighting techniques for text simplification tasks with encoder-decoder architectures.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Prove Your Point!: Bringing Proof-Enhancement Principles to Argumentative Essay Generation

Ruyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, Ting Liu

Argumentative essay generation (AEG) aims to generate complete texts on specific controversial topics or debates. Although current AEG methods can generate individual opinions, they often overlook the high-level connections between these opinions. This often leads to the generated results being mired in logical confusion, unable to proof their own arguments effectively. The generated essay may present evidence that contradicts the claims or they may fail to assemble the claims into logical flow. In this paper, we present a unified two-stage framework: Proof-Enhancement and Self-Annotation (PESA) for AEG with a focus on logical enhancement. Specifically, we first construct pseudo-labels for logical information, claims and grounds, using a large language model. We then propose a tree planning approach that introduces proof principles and ensures logical consistency. Extensive experimental results show that, benefiting from proof principle guidance, PESA generates argumentative essays with better logical validity and persuasiveness than strong baseline models.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LLM Self-Correction with DeCRIM: Decompose, Critique, and Refine for Enhanced Following of Instructions with Multiple Constraints

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, Nanyun Peng

Instruction following is a key capability for LLMs. However, recent studies have shown that LLMs often struggle with instructions containing multiple constraints (e.g., a request to create a social media post "in a funny tone" with "no hashtag"). Despite this, most evaluations focus solely on synthetic data. To address this, we introduce RealInstruct, the first benchmark designed to evaluate LLMs' ability to follow real-world multi-constrained instructions by leveraging queries real users asked AI assistants. We also investigate model-based evaluation as a cost-effective alternative to human annotation for this task. Our findings reveal that even the proprietary GPT-4 model fails to meet at least one constraint on over 21% of instructions, highlighting the limitations of state-of-the-art models. To address the performance gap between open-source and proprietary models, we propose the Decompose, Critique and Refine (DeCRIM) self-correction pipeline, which enhances LLMs' ability to follow constraints. DeCRIM works by decomposing the original instruction into a list of constraints and using a Critic model to decide when and where the LLM's response needs refinement. Our results show that DeCRIM improves Mistral's performance by 7.3% on RealInstruct and 8.0% on IFEval even with weak feedback. Moreover, we demonstrate that with strong feedback, open-source LLMs with DeCRIM can outperform GPT-4 on both benchmarks.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Compare without Despair: Reliable Preference Evaluation with Generation Separability

Sayan Ghosh, Tejas Srinivasan, Swabha Swamyandipta

Human evaluation of generated language through pairwise preference judgments is pervasive. However, under common scenarios, such as when generations from a model pair are very similar, or when stochastic decoding results in large variations in generations, it results in inconsistent preference ratings. We address these challenges by introducing a meta-evaluation measure, separability, which estimates how suitable a test instance is for pairwise preference evaluation. For a candidate test instance, separability samples multiple generations from a pair of models, and measures how distinguishable the two sets of generations are. Our experiments show that instances with high separability values yield more consistent preference ratings from both human- and auto-raters. Further, the distribution of separability allows insights into which test benchmarks are more valuable for comparing models. Finally, we incorporate separability into ELO ratings, accounting for how suitable each test instance might be for reliably ranking LLMs. Overall, separability has implications for consistent, efficient and robust preference evaluation of LLMs with both human- and auto-raters.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Reference-based Metrics Disprove Themselves in Question Generation

Bang Nguyen, Mengxia Yu, Yun Huang, Meng Jiang

Reference-based metrics such as BLEU and BERTScore are widely used to evaluate question generation (QG). In this study, on QG benchmarks such as SQuAD and HotpotQA, we find that using human-written references cannot guarantee the effectiveness of the reference-based metrics. Most QG benchmarks have only one reference; we replicate the annotation process and collect another reference. A good metric is expected to grade a human-validated question no worse than generated questions. However, the results of reference-based metrics on our newly collected reference disproved the metrics themselves. We propose a reference-free metric consisted of multi-dimensional criteria such as naturalness, answerability, and complexity, utilizing large language models. These criteria are not constrained to the syntactic or semantic of a single reference question, and the metric does not require a diverse set of references. Experiments reveal that our metric accurately distinguishes between high-quality questions and flawed ones, and achieves state-of-the-art alignment with human judgment.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LexC-Gen: Generating Data for Extremely Low-Resource Languages with Large Language Models and Bilingual Lexicons

Zheng Xin Yong, Cristina Menghini, Stephen Bach

Data scarcity in low-resource languages can be addressed with word-to-word translations from labeled task data in high-resource languages using bilingual lexicons. However, bilingual lexicons often have limited lexical overlap with task data, which results in poor translation coverage and lexicon utilization. We propose lexicon-conditioned data generation LexC-Gen, a method that generates low-resource-language classification task data at scale. Specifically, LexC-Gen first uses high-resource-language words from bilingual lexicons to generate lexicon-compatible task data, and then it translates them into low-resource languages with bilingual lexicons via word translation. Across 17 extremely low-resource languages, LexC-Gen generated data is competitive with expert-translated gold data, and yields on average 5.6 and 8.9 points improvement over existing lexicon-based word translation methods on sentiment analysis and topic classification tasks respectively. Through ablation study, we show that conditioning on bilingual lexicons is the key component of LexC-Gen. LexC-Gen serves as a potential solution to close the performance gap between open-source multilingual models, such as BLOOMZ and Aya-101, and state-of-the-art commercial models like GPT-4o on low-resource-language tasks.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

SWAG: Storytelling With Action Guidance

Jonathan Pei, Zeeshan Patel, Karim El-Refaie, Tianle Li

Automated long-form story generation typically employs long-context large language models (LLMs) for one-shot creation, which can produce cohesive but not necessarily engaging content. We introduce Storytelling With Action Guidance (SWAG), a novel approach to storytelling with LLMs. Our approach reduces story writing to a search problem through a two-model feedback loop: one LLM generates story content, and another auxiliary LLM is used to choose the next best "action" to steer the story's future direction. Our results show that SWAG can substantially outperform previous end-to-end story generation techniques when evaluated by GPT-4 and through human evaluation. Our SWAG pipeline using only small open-source models surpasses GPT-3.5-Turbo.

Industry

Nov 12 (Tue) 11:00-12:30 - Room: Jasmine

Nov 12 (Tue) 11:00-12:30 - Jasmine

GeoIndia: A Seq2Seq Geocoding Approach for Indian Addresses

Bhavuk Singhal, Anshu Aditya, Lokesh Todwal, Shubham Jain, Debasish Mukherjee

Geocoding, the conversion of unstructured geographic text into structured spatial data, is essential for logistics, urban planning, and location-based services. Indian addresses with their diverse languages, scripts, and formats present significant challenges that existing geocoding methods often fail to address, particularly at fine-grained resolutions. In this paper, we propose GeoIndia, a novel geocoding system designed

specifically for Indian addresses using hierarchical H3-cell prediction within a Seq2Seq framework. Our methodology includes a comprehensive analysis of Indian addressing systems, leading to the development of a data correction strategy that enhances prediction accuracy. We investigate two model architectures, Flan-T5-base (T5) and Llama-3-8b (QLF-Llama-3), due to their strong sequence generation capabilities. We trained approximately 30 models for each Indian state, and results show that our approach provides superior accuracy and reliability across multiple Indian states, outperforming the well-renowned geocoding platform Google Maps. In multiple states, we achieved more than 50% reduction in mean distance error and more than 85% reduction in 99th percentile distance error compared to Google Maps. This advancement can help in optimizing logistics in the e-commerce sector, reducing delivery failures and improving customer satisfaction.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Prompt-Tuned Multi-Task Taxonomic Transformer (PTMTTaxoFormer)

Rajashreha Vasanth, Nhan Nguyen, Yue Zhang

Hierarchical Text Classification (HTC) is a subclass of multi-label classification, it is challenging because the hierarchy typically has a large number of diverse topics. Existing methods for HTC fall within two categories, local methods (a classifier for each level, node, or parent) or global methods (a single classifier for everything). Local methods are computationally expensive, whereas global methods often require complex explicit injection of the hierarchy, verbalizers, and/or prompt engineering. In this work, we propose Prompt Tuned Multi Task Taxonomic Transformer (PTMTTaxoFormer¹), a single classifier that uses a multi-task objective to predict one or more topics. The approach is capable of understanding the hierarchy during training without explicit injection, complex heads, verbalizers, or prompt engineering. PTMTTaxoFormer is a novel model architecture and training paradigm using differentiable prompts and labels that are learnt through backpropagation. PTMTTaxoFormer achieves state of the art results on several HTC benchmarks that span a range of topics consistently. Compared to most other HTC models, it has a simpler yet effective architecture, making it more production-friendly in terms of latency requirements (a factor of 2-5 lower latency). It is also robust and label-efficient, outperforming other models with 15%-50% less training data.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Optimizing Entity Resolution in Voice Interfaces: An ASR-Aware Entity Reference Expansion Approach

Jiangning Chen, Ziyun Zhang, Qianli Hu

This paper tackles the challenges presented by Automatic Speech Recognition (ASR) errors in voice-based dialog systems, specifically, their adverse impact on Entity Resolution (ER) as a downstream task. Navigating the equilibrium between accuracy and online retrieval's speed requirement proves challenging, particularly when limited data links the failed mentions to resolved entities. In this paper, we propose a query expansion system, injecting pairs of failed mentions and resolved entity names into the knowledge graph, enhancing its awareness of unresolved mentions. To address data scarcity, we introduce a synthetic data generation approach aligned with noise patterns. This, combined with an ASR-Error-Aware Loss function, facilitates the training of a RoBERTa model, which filters failed mentions and extracts entity pairs for knowledge graph expansion. These designs confront obstacles related to ASR noise, data limitations, and online entity retrieval.

Nov 12 (Tue) 11:00-12:30 - Jasmine

BPIID: A Benchmark for Personal Identity Deduplication

Runhui Wang, Yefan Tao, Adit Krishnan, Luyang Kong, Xuanqing Liu, Yuqian Deng, Yunzhao Yang, Henrik Johnson, Andrew Borthwick, Shobhit Gupta, Aditi Sinha, Davor Golac

Data deduplication is a critical task in data management and mining, focused on consolidating duplicate records that refer to the same entity. Personally Identifiable Information (PII) is a critical class of data for deduplication across various industries. Consumer data, stored and generated through various engagement channels, is crucial for marketers, agencies, and publishers. However, a major challenge to PII data deduplication is the lack of open-source benchmark datasets due to stringent privacy concerns, which hinders the research, development, and evaluation of robust solutions. This paper addresses this critical lack of PII deduplication benchmarks by introducing the first open-source, high-quality dataset for this task. We provide two datasets: one with 1,000,000 unlabeled synthetic PII profiles and a subset of 10,000 pairs curated and labeled by trained annotators as matches or non-matches. Our datasets contain synthetic profiles built from publicly available sources that do not represent any real individuals, thus ensuring privacy and ethical compliance. We provide several challenging data variations to evaluate the effectiveness of various deduplication techniques, including traditional supervised methods, deep-learning approaches, and large language models (LLMs). Our work aims to set a new standard for PII deduplication, paving the way for more accurate and secure solutions. We share our data publicly at this link: <https://zenodo.org/records/12774140>.

Nov 12 (Tue) 11:00-12:30 - Jasmine

MERLIN: Multimodal Embedding Refinement via LLM-based Iterative Navigation for Text-Video Retrieval-Rerank Pipeline

Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, Nojun Kwak

The rapid expansion of multimedia content has made accurately retrieving relevant videos from large collections increasingly challenging. Recent advancements in text-video retrieval have focused on cross-modal interactions, large-scale foundation model training, and probabilistic modeling, yet often neglect the crucial user perspective, leading to discrepancies between user queries and the content retrieved. To address this, we introduce MERLIN (Multimodal Embedding Refinement via LLM-based Iterative Navigation), a novel, training-free pipeline that leverages Large Language Models (LLMs) for iterative feedback learning. MERLIN refines query embeddings from a user perspective, enhancing alignment between queries and video content through a dynamic question answering process. Experimental results on datasets like MSR-VTT, MSVD, and ActivityNet demonstrate that MERLIN substantially improves Recall@1, outperforming existing systems and confirming the benefits of integrating LLMs into multimodal retrieval systems for more responsive and context-aware multimedia retrieval.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Identifying High Consideration E-Commerce Search Queries

Zhiyu Chen, Jason Ingyu Choi, Besnik Fetahu, Shervin Malmasi

In e-commerce, high consideration search missions typically require careful and elaborate decision making, and involve a substantial research investment from customers. We consider the task of identifying High Consideration (HC) queries. Identifying such queries enables e-commerce sites to better serve user needs using targeted experiences such as curated QA widgets that help users reach purchase decisions. We explore the task by proposing an Engagement-based Query Ranking (EQR) approach, focusing on query ranking to indicate potential engagement levels with query-related shopping knowledge content during product search. Unlike previous studies on predicting trends, EQR prioritizes query-level features related to customer behavior, finance, and catalog information rather than popularity signals. We introduce an accurate and scalable method for EQR and present experimental results demonstrating its effectiveness. Offline experiments show strong ranking performance. Human evaluation shows a precision of 96% for HC queries identified by our model. The model was commercially deployed, and shown to outperform human-selected queries in terms of downstream customer impact, as measured through engagement.

Nov 12 (Tue) 11:00-12:30 - Jasmine

KorSmishing Explainer: A Korean-centric LLM-based Framework for Smishing Detection and Explanation Generation

¹Code repo will be made available after Amazon review.

Yunseung Lee, Daehee Han

To mitigate the annual financial losses caused by SMS phishing (smishing) in South Korea, we propose an explainable smishing detection framework that adapts to a Korean-centric large language model (LLM). Our framework not only classifies smishing attempts but also provides clear explanations, enabling users to identify and understand these threats. This end-to-end solution encompasses data collection, pseudo-label generation, and parameter-efficient task adaptation for models with fewer than five billion parameters. Our approach achieves a 15% improvement in accuracy over GPT-4 and generates high-quality explanatory text, as validated by seven automatic metrics and qualitative evaluation, including human assessments.

Nov 12 (Tue) 11:00-12:30 - Jasmine

MILD Bot: Multidisciplinary Childhood Cancer Survivor Question-Answering Bot

Mirae Kim, Kyubum Hwang, Hyoung Oh, Min Ah Kim, Chaerim Park, Yehwi Park, Chungyeon Lee

This study introduces a Multidisciplinary chILDhood cancer survivor question-answering (MILD) bot designed to support childhood cancer survivors facing diverse challenges in their survivorship journey. In South Korea, a shortage of experts equipped to address these unique concerns comprehensively leaves survivors with limited access to reliable information. To bridge this gap, our MILD bot employs a dual-component model featuring an intent classifier and a semantic textual similarity model. The intent classifier first analyzes the users query to identify the underlying intent and match it with the most suitable expert who can provide advice. Then, the semantic textual similarity model identifies questions in a predefined dataset that closely align with the users query, ensuring the delivery of relevant responses. This proposed framework shows significant promise in offering timely, accurate, and high-quality information, effectively addressing a critical need for support among childhood cancer survivors.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Assisting Breastfeeding and Maternity Experts in Responding to User Queries with an AI-in-the-loop Approach

Nadjet Bouyad-Agha, Ignasi Gomez-Sabatia, Alba Padro, Enric Pallarés, David Pelayo, Rocío Tovar

Breastfeeding and Maternity experts are a scarce resource and engaging in a conversation with mothers on such a sensitive topic is a time-consuming effort. We present our journey and rationale in assisting experts to answer queries about Breastfeeding and Maternity topics from users, mainly mothers. We started by developing a RAG approach to response generation where the generated response is made available to the expert who has the option to draft an answer using the generated text or to answer from scratch. This was the start of an ongoing effort to develop a pipeline of AI/NLP-based functionalities to help experts understand user queries and craft their responses.

Nov 12 (Tue) 11:00-12:30 - Jasmine

A new approach for fine-tuning sentence transformers for intent classification and out-of-scope detection tasks

Tianyi Zhang, Atta Norouzi, Aanchan Mohan, Frederick Ducatelle

In virtual assistant (VA) systems it is important to reject or redirect user queries that fall outside the scope of the system. One of the most accurate approaches for out-of-scope (OOS) rejection is to combine it with the task of intent classification on in-scope queries, and to use methods based on the similarity of embeddings produced by transformer-based sentence encoders. Typically, such encoders are fine-tuned for the intent-classification task, using cross-entropy loss. Recent work has shown that while this produces suitable embeddings for the intent-classification task, it also tends to disperse in-scope embeddings over the full sentence embedding space. This causes the in-scope embeddings to potentially overlap with OOS embeddings, thereby making OOS rejection difficult. This is compounded when OOS data is unknown. To mitigate this issue our work proposes to regularize the cross-entropy loss with an in-scope embedding reconstruction loss learned using an auto-encoder. Our method achieves a 1-4% improvement in the area under the precision-recall curve for rejecting out-of-sample (OOS) instances, without compromising intent classification performance.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Can Machine Unlearning Reduce Social Bias in Language Models?

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, Faiza Khan Khattak

Mitigating bias in language models (LMs) has become a critical problem due to the widespread deployment of LMs in the industry and customer-facing applications. Numerous approaches revolve around data pre-processing and subsequent fine-tuning of language models, tasks that can be both time-consuming and computationally demanding. As alternatives, machine unlearning techniques are being explored, yet there is a notable lack of comparative studies evaluating the effectiveness of these methods. In this work, we explore the effectiveness of two machine unlearning methods: Partitioned Contrastive Gradient Unlearning (PCGU) applied on decoder models, and Negation via Task Vector, and compare them with Direct Preference Optimization (DPO) to reduce social biases in open-source LMs such as LLaMA-2 and OPT. We also implement distributed PCGU for large models. It is empirically shown, through quantitative and qualitative analyses, that negation via Task Vector method outperforms PCGU and is comparable to DPO in debiasing models with minimum deterioration in model performance and perplexity. Negation via Task Vector reduces the bias score by 25.5% for LLaMA-2 and achieves bias reduction of up to 40% for OPT models. Moreover, it can be easily tuned to balance the trade-off between bias reduction and generation quality, unlike DPO.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Don't be my Doctor! Recognizing Healthcare Advice in Large Language Models

Kellen Tan Cheng, Anna Lisa Gentile, Pengyuan Li, Chad Deluca, Guang-Jie Ren

Large language models (LLMs) have seen increasing popularity in daily use, with their widespread adoption by many corporations as virtual assistants, chatbots, predictors, and many more. With the growing influence of industry corporations in this field, this raises the need for safeguards and guardrails to ensure that the outputs from LLMs do not mislead or harm users. This is especially true for highly regulated domains such as healthcare, where misleading advice may influence users to unknowingly commit malpractice. Despite this vulnerability, the majority of guardrail benchmarking datasets do not have enough focus on specifically medical advice. In this paper, we present the HEAL benchmark (HEalth Advice in LLMs), a health-advice benchmark dataset that has been manually curated and annotated to evaluate LLMs' capability in recognizing health-advice - which we use to safeguard LLMs deployed in industrial settings. We use HEAL to assess several models and report a detail analysis of the findings.

Nov 12 (Tue) 11:00-12:30 - Jasmine

OMG-QA: Building Open-Domain Multi-Modal Generative Question Answering Systems

Linyong Nan, Weineng Fang, Aylin Rasteh, Pouya Lahabi, Weijin Zou, Yilin Zhao, Arman Cohan

We introduce OMG-QA, a new resource for question answering that is designed to evaluate the effectiveness of question answering systems that perform retrieval augmented generation (RAG) in scenarios that demand reasoning on multi-modal, multi-document contexts. These systems, given a user query, must retrieve relevant contexts from the web, which may include non-textual information, and then reason and synthesize these contents to generate a detailed, coherent answer. Unlike existing open-domain QA datasets, OMG-QA requires systems to navigate and integrate diverse modalities and a broad pool of information sources, making it uniquely challenging. We conduct a thorough evaluation and analysis of a diverse set of QA systems, featuring various retrieval frameworks, document retrievers, document indexing approaches, evidence retrieval methods, and LLMs tasked with both information retrieval and generation. Our findings reveal significant limitations in existing approaches using RAG or LLM agents to address open questions that require long-form answers supported by multi-modal

evidence. We believe that OMG-QA will be a valuable resource for developing QA systems that are better equipped to handle open-domain, multi-modal information-seeking tasks.

Nov 12 (Tue) 11:00-12:30 - Jasmine

AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, Anuj Kumar

We present Any-Modality Augmented Language Model (AnyMAL), a unified model that reasons over diverse input modality signals (i.e. text, image, video, audio, IMU motion sensor), and generates textual responses. AnyMAL inherits the powerful text-based reasoning abilities of the state-of-the-art LLMs including Llama-3 (70B), and converts modality-specific signals to the joint textual space through a pre-trained aligner module. In this paper, we provide details on the optimizations implemented to efficiently scale the training pipeline, and present a comprehensive recipe for model and training configurations. We conduct comprehensive empirical analysis comprising both human and automatic evaluations, and demonstrate state-of-the-art performance on various multimodal tasks compared to industry-leading models such as Gemini-1.5 and GPT-4 – albeit with a relatively small number of trainable parameters.

Nov 12 (Tue) 11:00-12:30 - Jasmine

SLM as Guardian: Pioneering AI Safety with Small Language Model

Ohjoon Kwon, Donghyeon Jeon, Nayoung Choi, Gyu-Hwung Cho, Hwiyeol Jo, Changbong Kim, Hyunwoo Lee, Inho Kang, Sun Kim, Taiwoo Park

Most prior safety research of large language models (LLMs) has focused on enhancing the alignment of LLMs to better suit the safety requirements of their use cases. However, internalizing such safeguard features into larger models brought challenges of higher training cost and unintended degradation of helpfulness. In this paper, we leverage a smaller LLM for both harmful query detection and safeguard response generation. We introduce our safety requirements and the taxonomy of harmfulness categories, and then propose a multi-task learning mechanism fusing the two tasks into a single model. We demonstrate the effectiveness of our approach, providing on par or surpassing harmful query detection and safeguard response performance compared to the publicly available LLMs.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Hyper-QKSG: Framework for Automating Query Generation and Knowledge-Snippet Extraction from Tables and Lists

Dooyoung Kim, Yoonyjin Jang, Dongwook Shin, Chanhoon Park, Youngjoong Ko

These days, there is an increasing necessity to provide a user with a short knowledge-snippet for a query in commercial information retrieval services such as the featured snippet of Google. In this paper, we focus on how to automatically extract the candidates of query-knowledge snippet pairs from structured HTML documents by using a new Language Model (HTML-PLM). In particular, the proposed system is powerful on extracting them from Tables and Lists, and provides a new framework for automate query generation and knowledge-snippet extraction based on a QA-pair filtering procedure including the snippet refinement and verification processes, which enhance the quality of generated query-knowledge snippet pairs. As a result, 53.8% of the generated knowledge-snippets includes complex HTML structures such as tables and lists and in our experiments of a real-world environments, and 66.5% of the knowledge-snippets are evaluated as valid.

Nov 12 (Tue) 11:00-12:30 - Jasmine

A Cost-Efficient Modular Sieve for Extracting Product Information from Company Websites

Anna Häfty, Dragan Milchevski, Kersten Döring, Marko Putnikovic, Mohsen Mesgar, Filip Novović, Maximilian Braun, Karina Leoni Bornemann, Igor Stranjanac

Extracting product information is crucial for informed business decisions and strategic planning across multiple industries. However, recent methods relying only on large language models (LLMs) are resource-intensive and computationally prohibitive due to website structure differences and numerous non-product pages. To address these challenges, we propose a novel modular method that leverages low-cost classification models to filter out company web pages, significantly reducing computational costs. Our approach consists of three modules: web page crawling, product page classification using efficient machine learning models, and product information extraction using LLMs on classified product pages. We evaluate our method on a new dataset of about 7000 product and non-product web pages, achieving a 6-point improvement in F1-score, 95% reduction in computational time, and 87.5% reduction in cost compared to end-to-end LLMs. Our research demonstrates the effectiveness of our proposed low-cost classification module to identify web pages containing product information, making product information extraction more effective and cost-efficient.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Predicting Entity Salience in Extremely Short Documents

Benjamin Bulloch, Harrison Lundberg, Chen Hu, Weihang Xiao

A frequent challenge in applications that use entities extracted from text documents is selecting the most salient entities when only a small number can be used by the application (e.g., displayed to a user). Solving this challenge is particularly difficult in the setting of extremely short documents, such as the response from a digital assistant, where traditional signals of salience such as position and frequency are less likely to be useful. In this paper, we propose a lightweight and data-efficient approach for entity salience detection on short text documents. Our experiments show that our approach achieves competitive performance with respect to complex state-of-the-art models, such as GPT-4, at a significant advantage in latency and cost. In limited data settings, we show that a semi-supervised fine-tuning process can improve performance further. Furthermore, we introduce a novel human-labeled dataset for evaluating entity salience on short question-answer pair documents.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Centrality-aware Product Retrieval and Ranking

Hadeel Saadany, Swapnil Bhosale, Samarth Agrawal, Constantin Orasan, Zhe Wu, Dipesh Kanjao

This paper addresses the challenge of improving user experience on e-commerce platforms by enhancing product ranking relevant to user's search queries. Ambiguity and complexity of user queries often lead to a mismatch between user's intent and retrieved product titles or documents. Recent approaches have proposed the use of Transformer-based models which need millions of annotated query-title pairs during the pre-training stage, and this data often does not take user intent into account. To tackle this, we curate samples from existing datasets at eBay, manually annotated with buyer-centric relevance scores, and centrality scores which reflect how well the product title matches the users intent. We introduce a User-intent Centrality Optimization (UCO) approach for existing models, which optimizes for the user intent in semantic product search. To that end, we propose a dual-loss based optimization to handle hard negatives, i.e., product titles that are semantically relevant but do not reflect the user's intent. Our contributions include curating challenging evaluation sets and implementing UCO, resulting in significant improvements in product ranking efficiency, observed for different evaluation metrics. Our work aims to ensure that the most buyer-centric titles for a query are ranked higher, thereby, enhancing the user experience on e-commerce platforms.

Information Retrieval and Text Mining 1

Nov 12 (Tue) 11:00-12:30 - Room: Riverfront Hall

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

XRec: Large Language Models for Explainable Recommendation

Qiuya Ma, Xubin Ren, Chao Huang

Recommender systems help users navigate information overload by providing personalized recommendations aligned with their preferences. Collaborative Filtering (CF) is a widely adopted approach, but while advanced techniques like graph neural networks (GNNs) and self-supervised learning (SSL) have enhanced CF models for better user representations, they often lack the ability to provide explanations for the recommended items. Explainable recommendations aim to address this gap by offering transparency and insights into the recommendation decision-making process, enhancing users' understanding. This work leverages the language capabilities of Large Language Models (LLMs) to push the boundaries of explainable recommender systems. We introduce a model-agnostic framework called XRec, which enables LLMs to provide comprehensive explanations for user behaviors in recommender systems. By integrating collaborative signals and designing a lightweight collaborative adaptor, the framework empowers LLMs to understand complex patterns in user-item interactions and gain a deeper understanding of user preferences. Our extensive experiments demonstrate the effectiveness of XRec, showcasing its ability to generate comprehensive and meaningful explanations that outperform baseline approaches in explainable recommender systems.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LongEmbed: Extending Embedding Models for Long Context Retrieval

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furui Wei, Sujian Li

Embedding models play a pivotal role in modern NLP applications such as document retrieval. However, existing embedding models are limited to encoding short documents of typically 512 tokens, restrained from application scenarios requiring long inputs. This paper explores context window extension of existing embedding models, pushing their input length to a maximum of 32,768. We begin by evaluating the performance of existing embedding models using our newly constructed LongEmbed benchmark, which includes two synthetic and four real-world tasks, featuring documents of varying lengths and dispersed target information. The benchmarking results highlight huge opportunities for enhancement in current models. Via comprehensive experiments, we demonstrate that training-free context window extension strategies can effectively increase the input length of these models by several folds. Moreover, comparison of models using Absolute Position Encoding (APE) and Rotary Position Encoding (RoPE) reveals the superiority of RoPE-based embedding models in context window extension, offering empirical guidance for future models. Our benchmark, code and trained models will be released to advance the research in long context embedding models.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval

Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, Zhicheng Dou

Conversational search requires accurate interpretation of user intent from complex multi-turn contexts. This paper presents ChatRetriever, which inherits the strong generalization capability of large language models to robustly represent complex conversational sessions for dense retrieval. To achieve this, we propose a simple and effective dual-learning approach that adapts LLM for retrieval via contrastive learning while enhancing the complex session understanding through masked instruction tuning on high-quality conversational instruction tuning data. Extensive experiments on five conversational search benchmarks demonstrate that ChatRetriever significantly outperforms existing conversational dense retrievers, achieving state-of-the-art performance on par with LLM-based rewriting approaches. Furthermore, ChatRetriever exhibits superior robustness in handling diverse conversational contexts. Our work highlights the potential of adapting LLMs for retrieval with complex inputs like conversational search sessions and proposes an effective approach to advance this research direction.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Learning Interpretable Legal Case Retrieval via Knowledge-Guided Case Reformulation

Chenlong Deng, Kelong Mao, Zhicheng Dou

Legal case retrieval for sourcing similar cases is critical in upholding judicial fairness. Different from general web search, legal case retrieval involves processing lengthy, complex, and highly specialized legal documents. Existing methods in this domain often overlook the incorporation of legal expert knowledge, which is crucial for accurately understanding and modeling legal cases, leading to unsatisfactory retrieval performance. This paper introduces KELLER, a legal knowledge-guided case reformulation approach based on large language models (LLMs) for effective and interpretable legal case retrieval. By incorporating professional legal knowledge about crimes and law articles, we enable large language models to accurately reformulate the original legal case into concise sub-facts of crimes, which contain the essential information of the case. Extensive experiments on two legal case retrieval benchmarks demonstrate superior retrieval performance and robustness on complex legal case queries of KELLER over existing methods.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Large Language Models as Foundations for Next-Gen Dense Retrieval: A Comprehensive Empirical Assessment

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, Kang Liu

Pre-trained language models like BERT and T5 serve as crucial backbone encoders for dense retrieval. However, these models often exhibit limited generalization capabilities and face challenges in improving in-domain accuracy. Recent research has explored using large language models (LLMs) as retrievers, achieving state-of-the-art performance across various tasks. Despite these advancements, the specific benefits of LLMs over traditional retrievers and the impact of different LLM configuration such as parameter sizes, pre-training duration, and alignment processor retrieval tasks remain unclear. In this work, we conduct a comprehensive empirical study on a wide range of retrieval tasks, including in-domain accuracy, data efficiency, zero-shot generalization, lengthy retrieval, instruction-based retrieval, and multi-task learning. We evaluate over 15 different backbone LLMs and non-LLMs. Our findings reveal that larger models and extensive pre-training consistently enhance in-domain accuracy and data efficiency. Additionally, larger models demonstrate significant potential in zero-shot generalization, lengthy retrieval, instruction-based retrieval, and multi-task learning. These results underscore the advantages of LLMs as versatile and effective backbone encoders in dense retrieval, providing valuable insights for future research and development in this field.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, Jian-Yun Nie

In this paper, we study how open-source large language models (LLMs) can be effectively deployed for improving query rewriting in conversational search, especially for ambiguous queries. We introduce CHIQ, a two-step method that leverages the capabilities of LLMs to resolve ambiguities in the conversation history before query rewriting. This approach contrasts with prior studies that predominantly use closed-source LLMs to directly generate search queries from conversation history. We demonstrate on five well-established benchmarks that CHIQ leads to state-of-the-art results across most settings, showing highly competitive performances with systems leveraging closed-source LLMs. Our study provides a first step towards leveraging open-source LLMs in conversational search, as a competitive alternative to the

prevailing reliance on commercial LLMs. Data, models, and source code will be publicly available upon acceptance at <https://github.com/fen-granMark/CHIQ>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval
Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, Guido Zuccon

Utilizing large language models (LLMs) for zero-shot document ranking is done in one of two ways: (1) prompt-based re-ranking methods, which require no further training but are only feasible for re-ranking a handful of candidate documents due to computational costs; and (2) unsupervised contrastive trained dense retrieval methods, which can retrieve relevant documents from the entire corpus but require a large amount of paired text data for contrastive training. In this paper, we propose PromptReps, which combines the advantages of both categories: no need for training and the ability to retrieve from the whole corpus. Our method only requires prompts to guide an LLM to generate query and document representations for effective document retrieval. Specifically, we prompt the LLMs to represent a given text using a single word, and then use the last token's hidden states and the corresponding logits associated with the prediction of the next token to construct a hybrid document retrieval system. The retrieval system harnesses both dense text embedding and sparse bag-of-words representations given by the LLM. Our experimental evaluation on the MSMARCO, TREC deep learning and BEIR zero-shot document retrieval datasets illustrates that this simple prompt-based LLM retrieval method can achieve a similar or higher retrieval effectiveness than state-of-the-art LLM embedding methods that are trained with large amounts of unsupervised data, especially when using a larger LLM.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs

Zheng Wang, Zhongyang Li, Jiang Zeren, Dandan Tu, Wei Shi

In the age of mobile internet, user data, often referred to as memories, is continuously generated on personal devices. Effectively managing and utilizing this data to deliver services to users is a compelling research topic. In this paper, we introduce a novel task of crafting personalized agents powered by large language models (LLMs), which utilize a user's smartphone memories to enhance downstream applications with advanced LLM capabilities. To achieve this goal, we introduce EMG-RAG, a solution that combines Retrieval-Augmented Generation (RAG) techniques with an Editable Memory Graph (EMG). This approach is further optimized using Reinforcement Learning to address three distinct challenges: data collection, editability, and selectability. Extensive experiments on a real-world dataset validate the effectiveness of EMG-RAG, achieving an improvement of approximately 10% over the best existing approach. Additionally, the personalized agents have been transferred into a real smartphone AI assistant, which leads to enhanced usability.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Unifying Multimodal Retrieval via Document Screenshot Embedding

Xueguang Ma, Sheng-Cheieh Lin, Minghan Li, Wenhui Chen, Jimmy Lin

In the real world, documents are organized in different formats and varied modalities. Traditional retrieval pipelines require tailored document parsing techniques and content extraction modules to prepare input for indexing. This process is tedious, prone to errors, and has information loss. To this end, we propose Document Screenshot Embedding (DSE), a novel retrieval paradigm that regards document screenshots as a unified input format, which does not require any content extraction preprocess and preserves all the information in a document (e.g., text, image and layout). DSE leverages a large vision-language model to directly encode document screenshots into dense representations for retrieval. To evaluate our method, we first craft the dataset of Wiki-SS, a 1.3M Wikipedia web page screenshots as the corpus to answer the questions from the Natural Questions dataset. In such a text-intensive document retrieval setting, DSE shows competitive effectiveness compared to other text retrieval methods relying on parsing. For example, DSE outperforms BM25 by 17 points in top-1 retrieval accuracy. Additionally, in a mixed-modality task of slide retrieval, DSE significantly outperforms OCR text retrieval methods by over 15 points in nDCG@10. These experiments show that DSE is an effective document retrieval paradigm for diverse types of documents. Model checkpoints, code, and Wiki-SS collection will be released.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

GENRA: Enhancing Zero-shot Retrieval with Rank Aggregation

Georgios Katsimpras, Georgios Palioras

Large Language Models (LLMs) have been shown to effectively perform zero-shot document retrieval, a process that typically consists of two steps: i) retrieving relevant documents, and ii) re-ranking them based on their relevance to the query. This paper presents GENRA, a new approach to zero-shot document retrieval that incorporates rank aggregation to improve retrieval effectiveness. Given a query, GENRA first utilizes LLMs to generate informative passages that capture the query's intent. These passages are then employed to guide the retrieval process, selecting similar documents from the corpus. Next, we use LLMs again for a second refinement step. This step can be configured for either direct relevance assessment of each retrieved document or for re-ranking the retrieved documents. Ultimately, both approaches ensure that only the most relevant documents are kept. Upon this filtered set of documents, we perform multi-document retrieval, generating individual rankings for each document. As a final step, GENRA leverages rank aggregation, combining the individual rankings to produce a single refined ranking. Extensive experiments on benchmark datasets demonstrate that GENRA improves existing approaches, highlighting the effectiveness of the proposed methodology in zero-shot retrieval.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

FIRST: Faster Improved Listwise Reranking with Single Token Decoding

Revanth Gangi Reddy, JaeHyek Doo, Yifei Xu, Md Arfaat Sultan, Deeyava Swain, Avirup Sil, Heng Ji

Large Language Models (LLMs) have significantly advanced the field of information retrieval, particularly for reranking. Listwise LLM rerankers have showcased superior performance and generalizability compared to existing supervised approaches. However, conventional listwise LLM reranking methods lack efficiency as they provide ranking output in the form of a generated ordered sequence of candidate passage identifiers. Further, they are trained with the typical language modeling objective, which treats all ranking errors uniformly—potentially at the cost of missing highly relevant passages. Addressing these limitations, we introduce FIRST, a novel listwise LLM reranking approach leveraging the output logits of the first generated identifier to directly obtain a ranked ordering of the candidates. Further, we incorporate a learning-to-rank loss during training, prioritizing ranking accuracy for the more relevant passages. Empirical results demonstrate that FIRST accelerates inference by 50% while maintaining a robust ranking performance with gains across the BEIR benchmark. Finally, to illustrate the practical effectiveness of listwise LLM rerankers, we investigate their application in providing relevance feedback for retrievers during inference. Our results show that LLM rerankers can provide a stronger distillation signal compared to cross-encoders, yielding substantial improvements in retriever recall after relevance feedback.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Matryoshka-Adaptor: Unsupervised and Supervised Tuning for Smaller Embedding Dimensions

Jinsung Yoon, Rajarshi Sinha, Sercan O Arik, Tomas Pfister

Embeddings from Large Language Models (LLMs) have emerged as critical components in various applications, particularly for information retrieval. While high-dimensional embeddings generally demonstrate superior performance as they contain more salient information, their

practical application is frequently hindered by elevated computational latency and the associated higher cost. To address these challenges, we propose Matryoshka-Adaptor, a novel tuning framework designed for the customization of LLM embeddings. Matryoshka-Adaptor facilitates substantial dimensionality reduction while maintaining comparable performance levels, thereby achieving a significant enhancement in computational efficiency and cost-effectiveness. Our framework directly modifies the embeddings from pre-trained LLMs which is designed to be seamlessly integrated with any LLM architecture, encompassing those accessible exclusively through black-box APIs. Also, it exhibits efficacy in both unsupervised and supervised learning settings. A rigorous evaluation conducted across a diverse corpus of English, multilingual, and multimodal datasets consistently reveals substantial gains with Matryoshka-Adaptor. Notably, with Google and OpenAI Embedding APIs, Matryoshka-Adaptor achieves a reduction in dimensionality ranging from two- to twelve-fold without compromising performance across multiple BEIR datasets.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

MixGR: Enhancing Retriever Generalization for Scientific Domain through Complementary Granularity

Fengyu Cai, Xinran Zhao, Tong Chen, Siyao Chen, Hongming Zhang, Iryna Gurevych, Heinz Koepll

Recent studies show the growing significance of document retrieval in the generation of LLMs, i.e., RAG, within the scientific domain by bridging their knowledge gap. However, dense retrievers often struggle with domain-specific retrieval and complex query-document relationships, particularly when query segments correspond to various parts of a document. To alleviate such prevalent challenges, this paper introduces MixGR, which improves dense retrievers' awareness of query-document matching across various levels of granularity in queries and documents using a zero-shot approach. MixGR fuses various metrics based on these granularities to a unified score that reflects a comprehensive query-document similarity. Our experiments demonstrate that MixGR outperforms previous document retrieval by 24.7%, 9.8%, and 6.9% on nDCG@5 with unsupervised, supervised, and LLM-based retrievers, respectively, averaged on queries containing multiple sub-queries from five scientific retrieval datasets. Moreover, the efficacy of two downstream scientific question-answering tasks highlights the advantage of MixGR to boost the application of LLMs in the scientific domain.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Improve Dense Passage Retrieval with Entailment Tuning

Lu Dai, Hao Liu, Hui Xiong

Retrieval module can be plugged into many downstream NLP tasks to improve their performance, such as open-domain question answering and retrieval-augmented generation. The key to a retrieval system is to calculate relevance scores to query and passage pairs. However, the definition of relevance is often ambiguous. We observed that a major class of relevance aligns with the concept of entailment in NLI tasks. Based on this observation, we designed a method called entailment tuning to improve the embedding of dense retrievers. Specifically, we unify the form of retrieval data and NLI data using existence claim as a bridge. Then, we train retrievers to predict the claims entailed in a passage with a variant task of masked prediction. Our method can be efficiently plugged into current dense retrieval methods, and experiments show the effectiveness of our method.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LitSearch: A Retrieval Benchmark for Scientific Literature Search

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Dangi Chen, Tianyu Gao

Literature search questions, such as where can I find research on the evaluation of consistency in generated summaries? pose significant challenges for modern search engines and retrieval systems. These questions often require a deep understanding of research concepts and the ability to reason over entire articles. In this work, we introduce LitSearch, a retrieval benchmark comprising 597 realistic literature search queries about recent ML and NLP papers. LitSearch is constructed using a combination of (1) questions generated by GPT-4 based on paragraphs containing inline citations from research papers and (2) questions about recently published papers, manually written by their authors. All LitSearch questions were manually examined or edited by experts to ensure high quality. We extensively benchmark state-of-the-art retrieval models and also evaluate two LLM-based reranking pipelines. We find a significant performance gap between BM25 and state-of-the-art dense retrievers, with a 24.8% difference in absolute recall@5. The LLM-based reranking strategies further improve the best-performing dense retriever by 4.4%. Additionally, commercial search engines and research tools like Google Search perform poorly on LitSearch, lagging behind the best dense retriever by 32 points. Taken together, these results show that LitSearch is an informative new testbed for retrieval systems while catering to a real-world use case.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Dense X Retrieval: What Retrieval Granularity Should We Use?

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, Dong Yu

Dense retrieval has become a prominent method to obtain relevant context or world knowledge in open-domain NLP tasks. When we use a learned dense retriever on a retrieval corpus at inference time, an often-overlooked design choice is the retrieval unit in which the corpus is indexed, e.g. document, passage, or sentence. We discover that the retrieval unit choice significantly impacts the performance of both retrieval and downstream tasks. Distinct from the typical approach of using passages or sentences, we introduce a novel retrieval unit, proposition, for dense retrieval. Propositions are defined as atomic expressions within text, each encapsulating a distinct factoid and presented in a concise, self-contained natural language format. We conduct an empirical comparison of different retrieval granularity. Our experiments reveal that indexing a corpus by fine-grained units such as propositions significantly outperforms passage-level units in retrieval tasks. Moreover, constructing prompts with fine-grained retrieved units for retrieval-augmented language models improves the performance of downstream QA tasks given a specific computation budget.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

PairDistill: Pairwise Relevance Distillation for Dense Retrieval

Chao-Wei Huang, Yun-Nung Chen

Effective information retrieval (IR) from vast datasets relies on advanced techniques to extract relevant information in response to queries. Recent advancements in dense retrieval have showcased remarkable efficacy compared to traditional sparse retrieval methods. To further enhance retrieval performance, knowledge distillation techniques, often leveraging robust cross-encoder rerankers, have been extensively explored. However, existing approaches primarily distill knowledge from pointwise rerankers, which assign absolute relevance scores to documents, thus facing challenges related to inconsistent comparisons. This paper introduces Pairwise Relevance Distillation (PairDistill) to leverage pairwise reranking, offering fine-grained distinctions between similarly relevant documents to enrich the training of dense retrieval models. Our experiments demonstrate that PairDistill outperforms existing methods, achieving new state-of-the-art results across multiple benchmarks. This highlights the potential of PairDistill in advancing dense retrieval techniques effectively. Our source code and trained models are released at <https://github.com/MiuLab/PairDistill>

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Discovering Biases in Information Retrieval Models Using Relevance Thesaurus as Global Explanation

Youngwoo Kim, Razieh Rahimi, James Allan

Most of the efforts in interpreting neural relevance models have been on local explanations, which explain the relevance of a document to

a query. However, local explanations are not effective in predicting the model's behavior on unseen texts. We aim at explaining a neural relevance model by providing lexical explanations that can be globally generalized. Specifically, we construct a relevance thesaurus containing semantically relevant query term and document term pairs, which can augment BM25 scoring functions to better approximate the neural model's predictions. We propose a novel method to build a relevance thesaurus construction. Our method involves training a neural relevance model which can score the relevance for partial segments of query and documents. The trained model is used to identify relevant terms over the vocabulary space. The resulting thesaurus explanation is evaluated based on ranking effectiveness and fidelity to the targeted neural ranking model. Finally, our thesaurus reveals the existence of brand name bias in ranking models, which further supports the utility of our explanation method.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

SPREADSHEETLLM: Encoding Spreadsheets for Large Language Models

Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Junyu Xiong, Shiyu Xia, Mengyu Zhou, Yun Lin, José Cambroner, Yeye He, Shi Han, Dongmei Zhang

Spreadsheets are characterized by their extensive two-dimensional grids, flexible layouts, and varied formatting options, which pose significant challenges for large language models (LLMs). In response, we introduce SheetEncoder, pioneering an efficient encoding method designed to unleash and optimize LLMs' powerful understanding and reasoning capability on spreadsheets. Initially, we propose a vanilla serialization approach that incorporates cell addresses, values, and formats. However, this approach was limited by LLMs' token constraints, making it impractical for most applications. To tackle this challenge, three innovative modules are proposed to compress spreadsheets effectively: structural-anchor-based compression, inverse index translation, and data-format-aware aggregation. It significantly improves performance in spreadsheet table detection task, outperforming the vanilla approach by 25.6% in GPT4s in-context learning setting. Moreover, fine-tuned LLM with SheetEncoder has an average compression ratio of 25xE, but achieves a state-of-the-art 78.9% F1 score, surpassing the best existing models by 12.3%, demonstrating that SheetEncoder greatly boosts LLMs's performance on spreadsheet data.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Comparing Neighbors Together Makes it Easy: Jointly Comparing Multiple Candidates for Efficient and Effective Retrieval

Jonghyun Song, Cheyun Jin, Wanlong Zhao, Jay-Yoon Lee

A common retrieve-and-rerank paradigm involves retrieving relevant candidates from a broad set using a fast bi-encoder (BE), followed by applying expensive but accurate cross-encoders (CE) to a limited candidate set. However, relying on this small subset is often susceptible to error propagation from the bi-encoders, which limits the overall performance. To address these issues, we propose the Comparing Multiple Candidates (CMC) framework. CMC compares a query and multiple embeddings of similar candidates (i.e., neighbors) through shallow self-attention layers, delivering rich representations contextualized to each other. Furthermore, CMC is scalable enough to handle multiple comparisons simultaneously. For example, comparing 10K candidates with CMC takes a similar amount of time as comparing 16 candidates with CE. Experimental results on the ZeSHEL dataset demonstrate that CMC, when plugged in between bi-encoders and cross-encoders as a seamless intermediate reranker (BE-CMC-CE), can effectively improve recall@k (+6.7%-p, +3.5%-p for R@16, R@64) compared to using only bi-encoders (BE-CE), with negligible slowdown (<7%). Additionally, to verify CMC's effectiveness as the final-stage reranker in improving top-1 accuracy, we conduct experiments on downstream tasks such as entity, passage, and dialogue ranking. The results indicate that CMC is not only faster (11x) but also often more effective than CE, with improved prediction accuracy in Wikipedia entity linking (+0.7%-p) and DSTC7 dialogue ranking (+3.3%-p).

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LongForm: Effective Instruction Tuning with Reverse Instructions

Abdullahi Köksal, Timo Schick, Anna Korhonen, Hinrich Schütze

Instruction tuning enables language models to more effectively generalize and better follow user intent. However, obtaining instruction data is costly and challenging. Prior work employs methods such as expensive human annotation, crowd-sourced datasets with alignment issues, and generating noisy examples via LLMs. We introduce the LongForm-C dataset, which is created by reverse instructions. We generate instructions via LLMs for human-written corpus examples using reverse instructions. First we select a diverse set of human-written documents from corpora such as C4 and Wikipedia; then we generate instructions for these documents via LLMs. This approach provides a cheaper and cleaner instruction-tuning dataset with natural output and one suitable for long text generation. Our models outperform 10x larger language models without instruction tuning on tasks such as story/recipe generation and long-form question answering. Moreover, LongForm models outperform prior instruction-tuned models such as FLAN-T5 and Alpaca by a large margin, and improve language understanding capabilities further. We publicly release our data and models: [Anonymized-URL].

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Textual Dataset Distillation via Language Model Embedding

Yefan Tao, Luyang Kong, Andrey Kan, Laurent Calot

Dataset distillation is a process aimed at condensing datasets while preserving essential characteristics. In the text domain, prevailing methods typically generate distilled data as embedding vectors, which are not human-readable. This approach simplifies optimization but limits the transferability of distilled data across different model architectures. To address this limitation, we introduce a model-agnostic, data-efficient method that leverages Language Model (LM) embeddings. Compared to parameter-efficient methods such as LORA, our approach achieves comparable performance with significantly faster processing times. We evaluate our methodology through classification tasks on datasets like IMDB and AG-News, demonstrating performance that is on par with or exceeds previous model-dependent techniques. By utilizing LM embeddings, our method offers enhanced flexibility and improved transferability, expanding the range of potential applications.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

MATE: Meet At The Embedding - Connecting Images with Long Texts

Young Kyun Jang, Junmo Kang, Yong Jae Lee, Donghyun Kim

While advancements in Vision Language Models (VLMs) have significantly improved the alignment of visual and textual data, these models primarily focus on aligning images with short descriptive captions. This focus limits their ability to handle complex text interactions, particularly with longer texts such as lengthy captions or documents, which have not been extensively explored yet. In this paper, we introduce Meet At The Embedding (MATE), a novel approach that combines the capabilities of VLMs with Large Language Models (LLMs) to overcome this challenge without the need for additional image-long text pairs. Specifically, we replace the text encoder of the VLM with a pretrained LLM-based encoder that excels in understanding long texts. To bridge the gap between VLM and LLM, MATE incorporates a projection module that is trained in a multi-stage manner. It starts by aligning the embeddings from the VLM text encoder with those from the LLM using extensive text pairs. This module is then employed to seamlessly align image embeddings closely with LLM embeddings. We propose two new cross-modal retrieval benchmarks to assess the task of connecting images with long texts (lengthy captions / documents). Extensive experimental results demonstrate that MATE effectively connects images with long texts, uncovering diverse semantic relationships.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Meta-Prompting Efficient Task-Adaptive Query Generator for Retrieval

Yoonsang Lee, Minsoo Kim, seung-won hwang

This paper studies the problem of information retrieval, to adapt to unseen tasks. Existing work generates synthetic queries from domain-specific documents to jointly train the retriever. However, the conventional query generator assumes the query as a question, thus failing to accommodate general search intents. A more lenient approach incorporates task-adaptive elements, such as few-shot learning with an 137B LLM. In this paper, we challenge a trend equating query and question, and instead conceptualize query generation task as a “compilation” of high-level intent into task-adaptive query. Specifically, we propose EGG, a query generator that better adapts to wide search intents expressed in the BeIR benchmark. Our method outperforms baselines and existing models on four tasks with underexplored intents, while utilizing a query generator 47 times smaller than the previous state-of-the-art. Our findings reveal that instructing the LM with explicit search intent is a key aspect of modeling an effective query generator.

Linguistic Theories, Cognitive Modeling and Psycholinguistics 1

Nov 12 (Tue) 11:00-12:30 - Room: Riverfront Hall

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

In-Context Learning May Not Elicit Trustworthy Reasoning: A-Not-B Errors in Pretrained Language Models

Pengrui Han, Peiyang Song, Haofei Yu, Jiaxuan You

Recent advancements in artificial intelligence have led to the creation of highly capable large language models (LLMs) that can perform tasks in a human-like manner. However, LLMs exhibit only infant-level cognitive abilities in certain areas. One such area is the A-Not-B error, a phenomenon seen in infants where they repeat a previously rewarded behavior despite well-observed changed conditions. This highlights their lack of inhibitory control – the ability to stop a habitual or impulsive response. In our work, we design a text-based multi-choice QA scenario similar to the A-Not-B experimental settings to systematically test the inhibitory control abilities of LLMs. We found that state-of-the-art LLMs (like Llama3-8b) perform consistently well with in-context learning (ICL) but make errors and show a significant drop of as many as 83.3% in reasoning tasks when the context changes trivially. This suggests that LLMs only have inhibitory control abilities on par with human infants in this regard, often failing to suppress the previously established response pattern during ICL.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models

Tyler Lookman, YUCHENG LI, Chenghua Lin

Recently, Large Language Models (LLMs) and Vision Language Models (VLMs) have demonstrated aptitude as potential substitutes for human participants in experiments testing psycholinguistic phenomena. However, an understudied question is to what extent models that only have access to vision and text modalities are able to implicitly understand sound-based phenomena via abstract reasoning from orthography and imagery alone. To investigate this, we analyse the ability of VLMs and LLMs to demonstrate sound symbolism (i.e., to recognise a non-arbitrary link between sounds and concepts) as well as their ability to “hear” via the interplay of the language and vision modules of open and closed-source multimodal models. We perform multiple experiments, including replicating the classic Kiki-Bouba and Mi-MaL shape and magnitude symbolism tasks and comparing human judgements of linguistic iconicity with that of LLMs. Our results show that VLMs demonstrate varying levels of agreement with human labels, and more task information may be required for VLMs versus their human counterparts for *in silico* experimentation. We additionally see through higher maximum agreement levels that Magnitude Symbolism is an easier pattern for VLMs to identify than Shape Symbolism, and that an understanding of linguistic iconicity is highly dependent on model size.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

On The Role of Context in Reading Time Prediction

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, Ethan Wilcox

We present a new perspective on how readers integrate context during real-time language comprehension. Our proposals build on surprisal theory, which posits that the processing effort of a linguistic unit (e.g., a word) is an affine function of its in-context information content. We first observe that surprisal is only one out of many potential ways that a contextual predictor can be derived from a language model. Another one is the pointwise mutual information (PMI) between a unit and its context, which turns out to yield the same predictive power as surprisal when controlling for unigram frequency. Moreover, both PMI and surprisal are correlated with frequency. This means that neither PMI nor surprisal contains information about context alone. In response to this, we propose a technique where we project surprisal onto the orthogonal complement of frequency, yielding a new contextual predictor that is uncorrelated with frequency. Our experiments show that the proportion of variance in reading times explained by context is a lot smaller when context is represented by the orthogonalized predictor. From an interpretability standpoint, this indicates that previous studies may have overstated the role that context has in predicting reading times.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Leading Whitespace of Language Models' Subword Vocabulary Poses a Confound for Calculating Word Probabilities

Byoung-Doh Oh, William Schuler

Predictions of word-by-word conditional probabilities from Transformer-based language models are often evaluated to model the incremental processing difficulty of human readers. In this paper, we argue that there is a confound posed by the most common method of aggregating subword probabilities of such language models into word probabilities. This is due to the fact that tokens in the subword vocabulary of most language models have leading whitespaces and therefore do not naturally define stop probabilities of words. We first prove that this can result in distributions over word probabilities that sum to more than one, thereby violating the axiom that $P(\Omega) = 1$. This property results in a misallocation of word-by-word surprisal, where the unacceptability of the end of the current word is incorrectly carried over to the next word. Additionally, this implicit prediction of word boundaries incorrectly models psycholinguistic experiments where human subjects directly observe upcoming word boundaries. We present a simple decoding technique to recount the probability of the trailing whitespace into that of the current word, which resolves this confound. Experiments show that this correction reveals lower estimates of garden-path effects in transitive/intransitive sentences and poorer fits to naturalistic reading times.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Semantic Training Signals Promote Hierarchical Syntactic Generalization in Transformers

Aditya Yedatore, Najoung Kim

Neural networks without hierarchical biases often struggle to learn linguistic rules that come naturally to humans. However, neural networks are trained primarily on form alone, while children acquiring language additionally receive data about meaning. Would neural networks generalize more like humans when trained on both form and meaning? We investigate this by examining if Transformers—neural networks without a hierarchical bias—better achieve hierarchical generalization when trained on both form and meaning compared to when trained on form alone. Our results show that Transformers trained on form and meaning do favor the hierarchical generalization more than those trained on form alone, suggesting that statistical learners without hierarchical biases can leverage semantic training signals to bootstrap hierarchical

syntactic generalization.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Predicting Nonnative Sentence Processing with L2LMs

Tatsuya Aoyama, Nathan Schneider

We study LMs pretrained sequentially on two languages (L2LMs) for modeling nonnative sentence processing. In particular, we pretrain GPT2 on 6 different first languages (L1s), followed by English as the second language (L2). We examine the effect of the choice of pretraining L1 on the models ability to predict human reading times, evaluating on English readers from a range of L1 backgrounds. Experimental results show that, while all of the LMs word surprises improve prediction of L2 reading times, especially for human L1s distant from English, there is no reliable effect of the choice of L2LMs L1. We also evaluate the learning trajectory of a monolingual English LM: for predicting L2 as opposed to L1 reading, it peaks much earlier and immediately falls off, possibly mirroring the difference in proficiency between the native and nonnative populations. Lastly, we provide examples of L2LMs surprises, which could potentially generate hypotheses about human L2 reading.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Toward Compositional Behavior in Neural Models: A Survey of Current Views

Kate McCurdy, Paul Soulos, Paul Smolensky

Compositionality is a core property of natural language, and compositional behavior (CB) is a crucial goal for modern NLP systems. The research literature, however, includes conflicting perspectives on how CB should be defined, evaluated, and achieved. We propose a conceptual framework to address these questions and survey researchers active in this area. We find consensus on several key points. Researchers broadly accept our proposed definition of CB, agree that it is not solved by current models, and doubt that scale alone will achieve the target behavior. In other areas, we find the field is split on how to move forward, identifying diverse opportunities for future research.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Reverse-Engineering the Reader

Samuel Kiegeland, Ethan Wilcox, Afra Amini, David Robert Reich, Ryan Cotterell

Numerous previous studies have sought to determine to what extent language models, pretrained on natural language text, can serve as useful models of human cognition. In this paper, we are interested in the opposite question: whether we can directly optimize a language model to be a useful cognitive model by aligning it to human psychometric data. To achieve this, we introduce a novel alignment technique in which we fine-tune a language model to implicitly optimize the parameters of a linear regressor that directly predicts humans' reading times of in-context linguistic units, e.g., phonemes, morphemes, or words, using surprisal estimates derived from the language model. Using words as a test case, we evaluate our technique across multiple model sizes and datasets and find that it improves language models' psychometric predictive power. However, we find an inverse relationship between psychometric power and a model's performance on downstream NLP tasks as well as its perplexity on held-out test data. While this latter trend has been observed before (Oh et al., 2022; Shain et al., 2024), we are the first to induce it by manipulating a model's alignment to psychometric data.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Development of Cognitive Intelligence in Pre-trained Language Models

Raj Sanjay Shah, Khushi Bhardwaj, Sashank Varma

Recent studies show evidence for emergent cognitive abilities in Large Pre-trained Language Models (PLMs). The increasing cognitive alignment of these models has made them candidates for cognitive science theories. Prior research into the emergent cognitive abilities of PLMs has been path-independent to model training, i.e. has only looked at the final model weights and not the intermediate steps. However, building plausible models of human cognition using PLMs also requires aligning their performance during training to the developmental trajectories of children's thinking. Guided by psychometric tests of human intelligence, we choose four task categories to investigate the alignment of ten popular families of PLMs and evaluate each of their available intermediate and final training steps: Numerical ability, Linguistic abilities, Conceptual understanding, and Fluid reasoning. We find a striking regularity: regardless of model size, the developmental trajectories of PLMs consistently exhibit a window of maximal alignment to human cognitive development. Before that window, training appears to endow models with the requisite structure to be poised to rapidly learn from experience. After that window, training appears to serve the engineering goal of reducing loss but not the scientific goal of increasing alignment with human cognition.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Don't Underestimate the Octopus - Why The Symbol Grounding Problem Does Not Apply to LLMs

Reto Gubermann

Do LLMs fall prey to Harnad's symbol grounding problem (SGP), as it has recently been claimed? We argue that this is not the case. Starting out with countering the arguments of Bender and Koller (2020), we trace the origins of the SGP to the computational theory of mind (CTM), and we show that it only arises with natural language when questionable theories of meaning are presupposed. We conclude by showing that it would apply to LLMs only if they were interpreted in the manner of how the CTM conceives the mind, i.e., by postulating that LLMs rely on a version of a language of thought, or by adopting said questionable theories of meaning; since neither option is rational, we conclude that the SGP does not apply to LLMs.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Let Me Teach You: Pedagogical Foundations of Feedback for Language Models

Beatriz Borges, Niket Tandon, Tanja Käser, Antoine Bosselut

Natural Language Feedback (NLF) is an increasingly popular mechanism for aligning Large Language Models (LLMs) to human preferences. Despite the diversity of the information that it can convey, NLF methods are often hand-designed and arbitrary, with little systematic grounding. At the same time, research in learning sciences has long established several effective feedback models. In this opinion piece, we compile ideas from pedagogy to introduce FELT, a feedback framework for LLMs that outlines various characteristics of the feedback space, and a feedback content taxonomy based on these variables, providing a general mapping of the feedback space. In addition to streamlining NLF designs, FELT also brings out new, unexplored directions for research in NLF. We make our taxonomy available to the community, providing guides and examples for mapping our categorizations to future research.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Learning to Write Rationally: How Information Is Distributed in Non-native Speakers Essays

Zixin Tang, Janet van Hell

People tend to distribute information evenly in language production for better and clearer communication. In this study, we compared essays written by second language (L2) learners with various native language (L1) backgrounds to investigate how they distribute information in their non-native L2 production. Analyses of surprisal and constancy of entropy rate indicated that writers with higher L2 proficiency can reduce the expected uncertainty of language production while still conveying informative content. However, the uniformity of information distribution showed less variability among different groups of L2 speakers, suggesting that this feature may be universal in L2 essay writing

and less affected by L2 writers variability in L1 background and L2 proficiency.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences

Leonardo Bertolazzi, Albert Gatt, Raffaella Bernardi

The reasoning abilities of Large Language Models (LLMs) are becoming a central focus of study in NLP. In this paper, we consider the case of syllogistic reasoning, an area of deductive reasoning studied extensively in logic and cognitive psychology. Previous research has shown that pre-trained LLMs exhibit reasoning biases, such as context effects, avoid answering that no conclusion follows, align with human difficulties, and struggle with multi-step reasoning. We contribute to this research line by systematically investigating the effects of chain-of-thought reasoning, in-context learning (ICL), and supervised fine-tuning (SFT) on syllogistic reasoning, considering syllogisms with conclusions that support or violate world knowledge and with multiple premises. Crucially, we go beyond the standard focus on accuracy, with an in-depth analysis of the conclusions generated by the models. Our results suggest that the behavior of pre-trained LLMs can be explained by heuristics studied in cognitive science and that both ICL and SFT improve model performance on valid inferences, although only the latter can mitigate most reasoning biases while being consistent.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Defining Knowledge: Bridging Epistemology and Large Language Models

Constanza Fierro, Ruchira Dhar, Filippos Stamatou, Nicolas Garneau, Anders Søgaard

Knowledge claims are abundant in the literature on large language models (LLMs); but can we say that GPT-4 truly "knows" the Earth is round? To address this question, we review standard definitions of knowledge in epistemology and we formalize interpretations applicable to LLMs. In doing so, we identify inconsistencies and gaps in how current NLP research conceptualizes knowledge with respect to epistemological frameworks. Additionally, we conduct a survey of 100 professional philosophers and computer scientists to compare their preferences in knowledge definitions and their views on whether LLMs can really be said to know. Finally, we suggest evaluation protocols for testing knowledge in accordance to the most relevant definitions.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Towards a Semantically-aware Surprise Theory

Clara Meister, Mario Giulianelli, Tiago Pimentel

Surprise theory posits that the cognitive effort required to comprehend a word is determined by its contextual predictability, quantified as surprisal. Traditionally, surprisal theory treats words as distinct entities, overlooking any potential similarity between them. Giulianelli et al. (2023) address this limitation by introducing information value, a measure of predictability designed to account for similarities between communicative units. Our work leverages Ricotta and Szeidl (2006) diversity index to extend surprisal into a metric that we term similarity-adjusted surprisal, exposing a mathematical relationship between surprisal and information value. Similarity-adjusted surprisal aligns with information value when considering graded similarities and reduces to standard surprisal when words are treated as distinct. Experimental results with reading time data indicate that similarity-adjusted surprisal adds predictive power beyond standard surprisal for certain datasets, suggesting it serves as a complementary measure of comprehension effort.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Scaling Cognitive Limits: Identifying Working Memory Limits in LLMs

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, Soroush Vosoughi

This study explores the inherent limitations of large language models (LLMs) from a scaling perspective, focusing on the upper bounds of their cognitive capabilities. We integrate insights from cognitive science to quantitatively examine how LLMs perform on n-back tasks a benchmark used to assess working memory, which involves temporarily holding and manipulating information. Our findings reveal that despite the increased model size, LLMs still face significant challenges in holding and processing information effectively, especially under complex task conditions. We also assess various prompting strategies, revealing their diverse impacts on LLM performance. The results highlight the struggle of current LLMs to autonomously discover optimal problem-solving patterns without heavily relying on manually corrected prompts. To move beyond these constraints, fundamental improvements in the planning and search of LLMs are essential for them to reason autonomously. Improving these capabilities will reduce the reliance on external corrections and enable LLMs to become more autonomous in their problem-solving processes.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Do LLMs learn a true syntactic universal?

John T. Hale, Milo Stanojević

Do large multilingual language models learn language universals? We consider a candidate universal much-discussed in the linguistics literature, the Final-over-Final Condition (Sheehan et al., 2017b). This Condition is syntactic in the sense that it can only be stated by reference to abstract sentence properties such as nested phrases and head direction. A study of typologically diverse mixed head direction languages confirms that the Condition holds in corpora. But in a targeted syntactic evaluation, Gemini Pro only seems to respect the Condition in German, Russian, Hungarian and Serbian. These relatively high-resource languages contrast with Basque, where Gemini Pro does not seem to have learned the Condition at all. This result suggests that modern language models may need additional sources of bias in order to become truly human-like, within a developmentally-realistic budget of training data.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

How to Compute the Probability of a Word

Tiago Pimentel, Clara Meister

Language models (LMs) estimate a probability distribution over strings in a natural language; these distributions are crucial for computing perplexity and surprisal in linguistics research. While we are usually concerned with measuring these values for words, most LMs operate over subwords. Despite seemingly straightforward, accurately computing probabilities over one unit given probabilities over the other requires care. Indeed, we show here that many recent linguistic studies have been incorrectly computing these values. This paper derives the correct methods for computing word probabilities, highlighting issues when relying on language models that use beginning-of-word (bow)-marking tokenisers, e.g., the GPT family. Empirically, we show that correcting the widespread bug in probability computations affects measured outcomes in sentence comprehension and lexical optimisation analyses.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Language models and brains align due to more than next-word prediction and word-level information

Gabriele Merlin, Mariya Toneva

Pretrained language models have been shown to significantly predict brain recordings of people comprehending language. Recent work suggests that the prediction of the next word is a key mechanism that contributes to this alignment. What is not yet understood is whether prediction of the next word is necessary for this observed alignment or simply sufficient, and whether there are other shared mechanisms or information that are similarly important. In this work, we take a step towards understanding the reasons for brain alignment via two simple

perturbations in popular pretrained language models. These perturbations help us design contrasts that can control for different types of information. By contrasting the brain alignment of these differently perturbed models, we show that improvements in alignment with brain recordings are due to more than improvements in next-word prediction and word-level information.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

On the Proper Treatment of Tokenization in Psycholinguistics

Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, Ryan Cotterell, Mario Giulianelli

Language models are widely used in computational psycholinguistics to test theories that relate the negative log probability (the surprisal) of a region of interest (a substring of characters) under a language model to its cognitive cost experienced by readers, as operationalized, for example, by gaze duration on the region. However, the application of modern language models to psycholinguistic studies is complicated by the practice of using tokenization as an intermediate step in training a model. Doing so results in a language model over *token* strings rather than one over character strings. Vexingly, regions of interest are generally misaligned with these token strings. The paper argues that token-level language models should be (approximately) marginalized into character-level language models before they are used in psycholinguistic studies to compute the surprisal of a region of interest; then, the marginalized character-level language model can be used to compute the surprisal of an arbitrary character substring, which we term a focal area, that the experimenter may wish to use as a predictor. Our proposal of marginalizing a token-level model into a character-level one solves this misalignment issue independently of the tokenization scheme. Empirically, we discover various focal areas whose surprisal is a better psychometric predictor than the surprisal of the region of interest itself.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Semantics and Sentiment: Cross-lingual Variations in Emoji Use

Giulio Zhou, Sydelle de Souza, Ella Markham, Oghenetekewee Kwakpovwe, Sumin Zhao

Over the past decade, the use of emoji in social media has seen a rapid increase. Despite their popularity and image-grounded nature, previous studies have found that people interpret emoji inconsistently when presented in context and in isolation. In this work, we explore whether emoji semantics differ across languages and how semantics interacts with sentiment in emoji use across languages. To do so, we developed a corpus containing the literal meanings for a set of emojis, as defined by L1 speakers in English, Portuguese and Chinese. We then use these definitions to assess whether speakers of different languages agree on whether an emoji is being used literally or figuratively in the context where they are grounded in, as well as whether this literal and figurative use correlates with the sentiment of the context itself. We found that there were varying levels of disagreement on the definition for each emoji but that these stayed fairly consistent across languages. We also demonstrated a correlation between the sentiment of a tweet and the figurative use of an emoji, providing theoretical underpinnings for empirical results in NLP tasks, particularly offering insights that can benefit sentiment analysis models.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Surprisal Curves of Discourse

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, Alex Warstadt

The Uniform Information Density (UID) hypothesis posits that speakers tend to distribute information evenly across linguistic units to achieve efficient communication. Of course, information rate in texts and discourses is not perfectly uniform. While these fluctuations can be viewed as theoretically uninteresting noise on top of a uniform target, another explanation is that UID is not the only functional pressure regulating information content in a language. Speakers may also seek to maintain interest, adhere to writing conventions, and build compelling arguments. In this paper, we propose one such functional pressure; namely that speakers modulate information rate based on location within a hierarchically-structured model of discourse. We term this the Structured Context Hypothesis and test it by predicting the surprisal contours of naturally occurring discourses extracted from large language models using predictors derived from discourse structure. We find that hierarchical predictors are significant predictors of a discourse's information contour and that deeply nested hierarchical predictors are more predictive than shallow ones. This work takes an initial step beyond UID to propose testable hypotheses for why the information rate fluctuates in predictable ways.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Emotion Granularity from Text: An Aggregate-Level Indicator of Mental Health

Krishnapriya Vishnubhotla, Daniela Teodorescu, Mallory J Feldman, Kristen Lindquist, Saif M. Mohammad

We are united in how emotions are central to shaping our experiences; yet, individuals differ greatly in how we each identify, categorize, and express emotions. In psychology, variation in the ability of individuals to differentiate between emotion concepts is called emotion granularity (determined through self-reports of one's emotions). High emotion granularity has been linked with better mental and physical health; whereas low emotion granularity has been linked with maladaptive emotion regulation strategies and poor health outcomes. In this work, we propose computational measures of emotion granularity derived from temporally-ordered speaker utterances in social media (in lieu of self reports that suffer from various biases). We then investigate the effectiveness of such text-derived measures of emotion granularity in functioning as markers of various mental health conditions (MHCs). We establish baseline measures of emotion granularity derived from textual utterances, and show that, at an aggregate level, emotion granularities are significantly lower for people self-reporting as having an MHC than for the control population. This paves the way towards a better understanding of the MHCs, and specifically the role emotions play in our well-being.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Unveiling Multi-level and Multi-modal Semantic Representations in the Human Brain using Large Language Models

Yuko Nakagi, Takuwa Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, Yu Takagi

In recent studies, researchers have used large language models (LLMs) to explore semantic representations in the brain; however, they have typically assessed different levels of semantic content, such as speech, objects, and stories, separately. In this study, we recorded brain activity using functional magnetic resonance imaging (fMRI) while participants viewed 8.3 hours of dramas and movies. We annotated these stimuli at multiple semantic levels, which enabled us to extract latent representations of LLMs for this content. Our findings demonstrate that LLMs predict human brain activity more accurately than traditional language models, particularly for complex background stories. Furthermore, we identify distinct brain regions associated with different semantic representations, including multi-modal vision-semantic representations, which highlights the importance of modeling multi-level and multi-modal semantic representations simultaneously. We will make our fMRI dataset publicly available to facilitate further research on aligning LLMs with human brain function.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories

Tarun Tater, Sabine Schulte in Walde, Diego Frassineti

The visual representation of a concept varies significantly depending on its meaning and the context where it occurs; this poses multiple challenges both for vision and multimodal models. Our study focuses on concreteness, a well-researched lexical-semantic variable, using it as a case study to examine the variability in visual representations. We rely on images associated with approximately 1,000 abstract and concrete concepts extracted from two different datasets: Bing and YFCC. Our goals are: (i) evaluate whether visual diversity in the depiction of concepts can reliably distinguish between concrete and abstract concepts; (ii) analyze the variability of visual features across multiple images

of the same concept through a nearest neighbor analysis; and (iii) identify challenging factors contributing to this variability by categorizing and annotating images. Our findings indicate that for classifying images of abstract versus concrete concepts, a combination of basic visual features such as color and texture is more effective than features extracted by more complex models like Vision Transformer (ViT). However, ViTs show better performances in the nearest neighbor analysis, emphasizing the need for a careful selection of visual features when analyzing conceptual variables through modalities other than text.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Is Child-Directed Speech Effective Training Data for Language Models?

Steven Y. Feng, Noah Goodman, Michael Frank

While high-performing language models are typically trained on hundreds of billions of words, human children become fluent language users with a much smaller amount of data. What are the features of the data they receive, and how do these features support language modeling objectives? To investigate this question, we train GPT-2 and RoBERTa models on 29M words of English child-directed speech and a new matched, synthetic dataset (TinyDialogues), comparing to OpenSubtitles, Wikipedia, and a heterogeneous blend of datasets from the BabyLM challenge. We evaluate the syntactic and semantic knowledge of these models using developmentally-inspired evaluations. Through pretraining experiments, we test whether the global developmental ordering or the local discourse ordering of children's training data supports high performance relative to other datasets. The local properties of the data affect model results, but surprisingly, global properties do not. Further, child language input is not uniquely valuable for training language models. These findings support the hypothesis that, rather than proceeding from better data, the child's learning algorithm is substantially more data-efficient than current language modeling techniques.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Generalized Measures of Anticipation and Responsivity in Online Language Processing

Mario Giulianelli, Andrea Opedal, Ryan Cotterell

We introduce a generalization of classic information-theoretic measures of predictive uncertainty in online language processing, based on the simulation of expected continuations of incremental linguistic contexts. Our framework provides a formal definition of anticipatory and responsive measures, and it equips experimenters with the tools to define new, more expressive measures beyond standard next-symbol entropy and surprisal. While extracting these standard quantities from language models is convenient, we demonstrate that using Monte Carlo simulation to estimate alternative responsive and anticipatory measures pays off empirically: New special cases of our generalized formula exhibit enhanced predictive power compared to surprisal for human cloze completion probability as well as ELAN, LAN, and N400 amplitudes, and greater complementarity with surprisal in predicting reading times.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

EmbodyBERT: Cognitively Informed Metaphor Detection Incorporating Sensorimotor Information

Yu Xi Li, Bo Peng, Yu-Yin Hsu, Chu-Ren Huang

The identification of metaphor is a crucial prerequisite for many downstream language tasks, such as sentiment analysis, opinion mining, and textual entailment. State-of-the-art systems of metaphor detection implement heuristic principles such as Metaphor Identification Procedure (MIP) and Selection Preference Violation (SPV). We propose an innovative approach that leverages the cognitive information of embodiment that can be derived from word embeddings, and explicitly models the process of sensorimotor change that has been demonstrated as essential for human metaphor processing. We showed that this cognitively motivated module is effective and can improve metaphor detection, compared with the heuristic MIP that has been applied previously.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A Psycholinguistic Evaluation of Language Models' Sensitivity to Argument Roles

Eun-Kyung Rosa Lee, Sathvik Nair, Naomi Feldman

We present a systematic evaluation of large language models' sensitivity to argument roles, i.e., *who* did what to *whom*, by replicating psycholinguistic studies on human argument role processing. In three experiments, we find that language models are able to distinguish verbs that appear in plausible and implausible contexts, where plausibility is determined through the relation between the verb and its preceding arguments. However, none of the models capture the same selective patterns that human comprehenders exhibit during real-time verb prediction. This indicates that language models' capacity to detect verb plausibility does not arise from the same mechanism that underlies human real-time sentence processing.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding

Chunkit Chan, Cheng Jiayang, Yuhuai Wang, Zheyue Deng, Wei Fan, Haoran Li, Xin Liu, Hongning Zhang, Weiqi Wang, Yangqiu Song

Large Language Models (LLMs) have sparked substantial interest and debate concerning their potential emergence of Theory of Mind (ToM) ability. Theory of mind evaluations currently focuses on testing models using machine-generated data or game settings prone to shortcuts and spurious correlations, which lacks evaluation of machine ToM ability in real-world human interaction scenarios. This poses a pressing demand to develop new real-world scenario benchmarks. We introduce NegotiationToM, a new benchmark designed to stress-test machine ToM in real-world negotiation surrounding covered multi-dimensional mental states (i.e., desires, beliefs, and intentions). Our benchmark builds upon the Belief-Desire-Intention (BDI) agent modeling theory and conducts the necessary empirical experiments to evaluate large language models. Our findings demonstrate that NegotiationToM is challenging for state-of-the-art LLMs, as they consistently perform significantly worse than humans, even when employing the chain-of-thought (CoT) method.

Multimodality and Language Grounding to Vision, Robotics and Beyond 1

Nov 12 (Tue) 11:00-12:30 - Room: Jasmine

Nov 12 (Tue) 11:00-12:30 - Jasmine

ImageInWords: Unlocking Hyper-Detailed Image Descriptions

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, Radu Soricut

Despite the longstanding adage "an image is worth a thousand words," generating accurate hyper-detailed image descriptions remains unsolved. Trained on short web-scraped image-text, vision-language models often generate incomplete descriptions with visual inconsistencies. We address this via a novel data-centric approach with ImageInWords (IIW), a carefully designed human-in-the-loop framework for curating hyper-detailed image descriptions. Human evaluations on IIW data show major gains compared to recent datasets (+66%) and GPT-4V (+48%) across comprehensiveness, specificity, hallucinations, and more. We also show that fine-tuning with IIW data improves these metrics by +31% against models trained with prior work, even with only 9k samples. Lastly, we evaluate IIW models with text-to-image generation

and vision-language reasoning tasks. Our generated descriptions result in the highest fidelity images, and boost compositional reasoning by up to 6% on ARO, SVO-Probes, and Winoground datasets. We release the IIW-Eval benchmark with human judgement labels, object and image-level annotations from our framework, and existing image caption datasets enriched via IIW-model.

Nov 12 (Tue) 11:00-12:30 - Jasmine

UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation

Xiangyu Zhao, Yuehan Zhang, Zhangwenlong, Xiao-Ming Wu

The fashion domain encompasses a variety of real-world multimodal tasks, including multimodal retrieval and multimodal generation. The rapid advancements in artificial intelligence generated content, particularly technologies like large language models for text generation and diffusion models for visual generation, have sparked widespread research interest in applying these multimodal models in the fashion domain. However, tasks that use embeddings, such as image-to-text or text-to-image retrieval, have been largely ignored from this perspective due to the diverse nature of the multimodal fashion domain. And current research on multi-task single models lack focus on image generation. In this work, we present UniFashion, a unified framework that simultaneously tackles the challenges of multimodal generation and retrieval tasks within the fashion domain, integrating image generation with retrieval tasks and text generation tasks. UniFashion unifies embedding and generative tasks by integrating a diffusion model and LLM, enabling controllable and high-fidelity generation. Our model significantly outperforms previous single-task state-of-the-art models across diverse fashion tasks, and can be readily adapted to manage complex vision-language tasks. This work demonstrates the potential learning synergy between multimodal generation and retrieval, offering a promising direction for future research in the fashion domain.

Nov 12 (Tue) 11:00-12:30 - Jasmine

HELPD: Mitigating Hallucination of LVLMs by Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding

Fan Yuan, Chi Qin, Xiaogang Xu, Piji Li

Large Vision-Language Models (LVLMs) have shown remarkable performance on many visual-language tasks. However, these models still suffer from *multimodal hallucination*, which means the generation of objects or content that violates the images. Many existing work detects hallucination by directly judging whether an object exists in an image, overlooking the association between the object and semantics. To address this issue, we propose Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding (HELPD). This framework incorporates hallucination feedback at both object and sentence semantic levels. Remarkably, even with a marginal degree of training, this approach can alleviate over 15% of hallucination. Simultaneously, HELPD penalizes the output logits according to the image attention window to avoid being overly affected by generated text. HELPD can be seamlessly integrated with any LVLMs. Our experiments demonstrate that the proposed framework yields favorable results across multiple hallucination benchmarks. It effectively mitigates hallucination for different LVLMs and concurrently improves their text generation quality.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Tag-grounded Visual Instruction Tuning with Retrieval Augmentation

Daqing Qi, Handong Zhao, Zijun Wei, Sheng Li

Despite recent advances in the general visual instruction-following ability of Multimodal Large Language Models (MLLMs), they still struggle with critical problems when required to provide a precise and detailed response to a visual instruction: (1) failure to identify novel objects or entities, (2) mention of non-existent objects, and (3) neglect of object's attributed details. Intuitive solutions include improving the size and quality of data or using larger foundation models. They show effectiveness in mitigating these issues, but at an expensive cost of collecting a vast amount of new data and introducing a significantly larger model. Standing at the intersection of these approaches, we examine the three object-oriented problems from the perspective of the image-to-text mapping process by the multimodal connector. In this paper, we first identify the limitations of multimodal connectors stemming from insufficient training data. Driven by this, we propose to enhance the mapping with retrieval-augmented tag tokens, which contain rich object-aware information such as object names and attributes. With our Tag-grounded visual instruction tuning with retrieval Augmentation (TUNA), we outperform baselines that share the same language model and training data on 12 benchmarks. Furthermore, we show the zero-shot capability of TUNA when provided with specific datastores.

Nov 12 (Tue) 11:00-12:30 - Jasmine

VGBench: A Comprehensive Benchmark of Vector Graphics Understanding and Generation for Large Language Models

Bocheng Zou, Mu Cai, Jianrui Zhang, Yong Jae Lee

In the realm of vision models, the primary mode of representation is using pixels to rasterize the visual world. Yet this is not always the best or unique way to represent visual content, especially for designers and artists who depict the world using geometry primitives such as polygons. Vector graphics (VG), on the other hand, offer a textual representation of visual content, which can be more concise and powerful for content like cartoons, sketches and scientific figures. Recent studies have shown promising results on processing vector graphics with capable Large Language Models (LLMs). However, such works focus solely on qualitative results, understanding, or a specific type of vector graphics. We propose VGBench, a comprehensive benchmark for LLMs on handling vector graphics through diverse aspects, including (a) both visual understanding and generation, (b) evaluation of various vector graphics formats, (c) diverse question types, (d) wide range of prompting techniques, (e) under multiple LLMs and (f) comparison with VLMs on rasterized representations. Evaluating on our collected 4279 understanding and 5845 generation samples, we find that LLMs show strong capability on both aspects while exhibiting less desirable performance on low-level formats (SVG). Both data and evaluation pipeline will be open-sourced.

Nov 12 (Tue) 11:00-12:30 - Jasmine

VIMI: Grounding Video Generation through Multi-modal Instruction

Yuwei Fang, Willi Menapace, Atikasandri Starorin, Tsai-Shien Chen, Kuan-Chieh Wang, Ivan Skorokhodov, Graham Neubig, Sergey Tulyakov

Existing text-to-video diffusion models rely solely on text-only encoders for their pretraining. This limitation stems from the absence of large-scale multimodal prompt video datasets, resulting in a lack of visual grounding and restricting their versatility and application in multimodal integration. To address this, we construct a large-scale multimodal prompt dataset by employing retrieval methods to pair in-context examples with the given text prompts and then utilize a two-stage training strategy to enable diverse video generation tasks within a model. In the first stage, we propose a multimodal conditional video generation framework for pretraining on these augmented datasets, establishing a foundational model for grounded video generation. Secondly, we fine-tune the model from the first stage on various video generation tasks, incorporating multimodal instructions. This process further refines the model's ability to handle diverse inputs and tasks, ensuring seamless integration of multimodal information. After this two-stage training process, VIMI demonstrates multimodal understanding capabilities, producing contextually rich and personalized videos grounded in the provided inputs, as shown in Figure1. Compared to previous subject-driven video generation methods, our generator can synthesize consistent and temporally coherent videos with large motion while retaining the semantic control. Our generator also achieves state-of-the-art text-to-video generation results on UCF101 benchmark.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Distilling Knowledge from Text-to-Image Generative Models Improves Visio-Linguistic Reasoning in CLIP

Samyadeep Basu, Shell Xu Hu, Mazar Sanjabi, Daniela Măsiceti, Soheil Feizi

Image-text contrastive models like CLIP have wide applications in zero-shot classification, image-text retrieval, and transfer learning. How-

ever, they often struggle on compositional visio-linguistic tasks (e.g., attribute-binding or object-relationships) where their performance is no better than random chance. To address this, we introduce SDS-CLIP, a lightweight and sample-efficient distillation method to enhance CLIP's compositional visio-linguistic reasoning. Our approach fine-tunes CLIP using a distillation objective borrowed from large text-to-image generative models like Stable-Diffusion, which are known for their strong visio-linguistic reasoning abilities. On the challenging Winoground benchmark, SDS-CLIP improves the visio-linguistic performance of various CLIP models by up to 7%, while on the ARO dataset, it boosts performance by up to 3%. This work underscores the potential of well-designed distillation objectives from generative models to enhance contrastive image-text models with improved visio-linguistic reasoning capabilities.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Finer: Investigating and Enhancing Fine-Grained Visual Concept Recognition in Large Vision Language Models

Jeonghwan Kim, Heng Ji

Recent advances in instruction-tuned Large Vision-Language Models (LVLMs) have imbued the models with the ability to generate high-level, image-grounded explanations with ease. While such capability is largely attributed to the rich world knowledge contained within the Large Language Models (LLMs), our work reveals their shortcomings in fine-grained visual categorization (FGVC) across six different benchmark settings. Most recent state-of-the-art LVLMs such as LLaVa-1.5, InstructBLIP and GPT-4V not only severely deteriorate in terms of classification performance, e.g., average drop of 65.58 in EM for Stanford Dogs for LLaVa-1.5, but also struggle to generate descriptive visual attributes based on a concept that appears within an input image despite their prominent zero-shot image captioning ability. In-depth analyses show that instruction-tuned LVLMs suffer from modality gap, showing discrepancy when given textual and visual inputs that correspond to the same concept. In an effort to further the community's endeavor in this direction, we propose a multiple granularity attribute-centric benchmark and training mixture, Finer, which aims to establish a ground to evaluate LVLMs' fine-grained visual comprehension ability and provide significantly improved explainability.

Nov 12 (Tue) 11:00-12:30 - Jasmine

UOUO: Uncontextualized Uncommon Objects for Measuring Knowledge Horizons of Vision Language Models

Xinyu Pi, Mingyuan Wu, Jize Jiang, Haochen Zheng, Beiting Tian, ChengXiang Zhai, Klara Nahrstedt, Zhiteng Hu

Xinyu Pi, Mingyuan Wu, Jize Jiang, Haochen Zheng, Beiting Tian, ChengXiang Zhai, Klara Nahrstedt, Zhiteng Hu
Smaller-scale Vision-Language Models (VLMs) often claim to perform on par with larger models in general-domain visual grounding and question-answering benchmarks while offering advantages in computational efficiency and storage. However, their ability to handle rare objects, which fall into the long tail of data distributions, is less understood. To rigorously evaluate this aspect, we introduce the "Uncontextualized Uncommon Objects" (UOUO) benchmark. This benchmark focuses on systematically testing VLMs with both large and small parameter counts on rare and specialized objects. Our comprehensive analysis reveals that while smaller VLMs maintain competitive performance on common datasets, they significantly underperform on tasks involving uncommon objects. We also propose an advanced, scalable pipeline for data collection and cleaning, ensuring the UOUO benchmark provides high-quality, challenging instances. These findings highlight the need to consider long-tail distributions when assessing the true capabilities of VLMs. Code and project details for UOUO can be found at <https://zoezheng126.github.io/UOUO-Website/>.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Encoding and Controlling Global Semantics for Long-form Video Question Answering

Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See-Kiong Ng, Anh Tuan Luu

Seeking answers effectively for long videos is essential to build video question answering (videoQA) systems. Previous methods adaptively select frames and regions from long videos to save computations. However, this fails to reason over the whole sequence of video, leading to sub-optimal performance. To address this problem, we introduce a state space layer (SSL) into multi-modal Transformer to efficiently integrate global semantics of the video, which mitigates the video information loss caused by frame and region selection modules. Our SSL includes a gating unit to enable controllability over the flow of global semantics into visual representations. To further enhance the controllability, we introduce a cross-modal compositional congruence objective to encourage global semantics aligned with the question. To rigorously evaluate long-form videoQA capacity, we construct two new benchmarks Ego-QA and MAD-QA featuring videos of considerably long length, i.e. 17.5 minutes and 1.9 hours, respectively. Extensive experiments demonstrate the superiority of our framework on these new as well as existing datasets.

Nov 12 (Tue) 11:00-12:30 - Jasmine

MEANT: Multimodal Encoder for Antecedent Information

Benjamin Irving, Annika Marie Schoene

The stock market provides a rich well of information that can be split across modalities, making it an ideal candidate for multimodal evaluation. Multimodal data plays an increasingly important role in the development of machine learning and has shown to positively impact performance. But information can do more than exist across modes—it can exist across time. How should we attend to temporal data that consists of multiple information types? This work introduces (i) the MEANT model, a Multimodal Encoder for Antecedent information and (ii) a new dataset called TempStock, which consists of price, Tweets, and graphical data with over a million Tweets from all of the companies in the S&P 500 Index. We find that MEANT improves performance on existing baselines by over 15%, and that the textual information affects performance far more than the visual information on our time-dependent task from our ablation study. The code and dataset will be made available upon publication.

Nov 12 (Tue) 11:00-12:30 - Jasmine

The Factuality Tax of Diversity-Intervened Text-to-Image Generation: Benchmark and Fact-Augmented Intervention

Yixin Wan, Di Wu, Haoran Wang, Kai-Wei Chang

Prompt-based "diversity interventions" are commonly adopted to improve the diversity of Text-to-Image (T2I) models depicting individuals with various racial or gender traits. However, will this strategy result in nonfunctional demographic distribution, especially when generating real historical figures? In this work, we propose **Demographic FActuality Representation (DoFaiR)**, a benchmark to systematically quantify the trade-off between using diversity interventions and preserving demographic factuality in T2I models. DoFaiR consists of 756 meticulously fact-checked test instances to reveal the factuality tax of various diversity prompts through an automated evidence-supported evaluation pipeline. Experiments on DoFaiR unveil that diversity-oriented instructions increase the number of different gender and racial groups in DALL-E-3's generations at the cost of historically inaccurate demographic distributions. To resolve this issue, we propose **Fact-Augmented Intervention** (FAI), which instructs a Large Language Model (LLM) to reflect on verbalized or retrieved factual information about gender and racial compositions of generation subjects in history, and incorporate it into the generation context of T2I models. By orienting model generations using the reflected historical truths, FAI significantly improves the demographic factuality under diversity interventions while preserving diversity.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Read Anywhere Pointed: Layout-aware GUI Screen Reading with Tree-of-Lens Grounding

Yue Fan, Lei Ding, Ching-Chen Kuo, Shan Jiang, Yang Zhao, Xinze Guan, Jie Yang, Yi Zhang, Xin Eric Wang

Graphical User Interfaces (GUIs) are central to our interaction with digital devices and growing efforts have been made to build models

for various GUI understanding tasks. However, these efforts largely overlook an important GUI-referring task: screen reading based on user-indicated points, which we name the Screen Point-and-Read (ScreenPR) task. Currently, this task is predominantly handled by rigid accessible screen reading tools, in great need of new models driven by advancements in Multimodal Large Language Models (MLLMs). In this paper, we propose a Tree-of-Lens (ToL) agent, utilizing a novel ToL grounding mechanism, to address the ScreenPR task. Based on the input point coordinate and the corresponding GUI screenshot, our ToL agent constructs a Hierarchical Layout Tree. Based on the tree, our ToL agent not only comprehends the content of the indicated area but also articulates the layout and spatial relationships between elements. Such layout information is crucial for accurately interpreting information on the screen, distinguishing our ToL agent from other screen reading tools. We also thoroughly evaluate the ToL agent against other baselines on a newly proposed ScreenPR benchmark, which includes GUIs from mobile, web, and operating systems. Last but not least, we test the ToL agent on mobile GUI navigation tasks, demonstrating its utility in identifying incorrect actions along the path of agent execution trajectories. Code and data: <https://screen-point-and-read.github.io>.

Nov 12 (Tue) 11:00-12:30 - Jasmine

If CLIP Could Talk: Understanding Vision-Language Model Representations Through Their Preferred Concept Descriptions

Reza Esfandiarpoor, Cristina Menghini, Stephen Bach

Recent works often assume that Vision-Language Model (VLM) representations are based on visual attributes like shape. However, it is unclear to what extent VLMs prioritize this information to represent concepts. We propose Extract and Explore (EX2), a novel approach to characterize textual features that are important for VLMs. EX2 uses reinforcement learning to align a large language model with VLM preferences and generates descriptions that incorporate features that are important for the VLM. Then, we inspect the descriptions to identify features that contribute to VLM representations. Using EX2, we find that spurious descriptions have a major role in VLM representations despite providing no helpful information, e.g., Click to enlarge photo of CONCEPT. More importantly, among informative descriptions, VLMs rely significantly on non-visual attributes like habitat (e.g., North America) to represent visual concepts. Also, our analysis reveals that different VLMs prioritize different attributes in their representations. Overall, we show that VLMs do not simply match images to scene descriptions and that non-visual or even spurious descriptions significantly influence their representations.

Nov 12 (Tue) 11:00-12:30 - Jasmine

MMoE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts

Haojie Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, Paul Pu Liang

Advances in multimodal models have greatly improved how interactions relevant to various tasks are modeled. Today's multimodal models mainly focus on the correspondence between images and text, using this for tasks like image-text matching. However, this covers only a subset of real-world interactions. Novel interactions, such as sarcasm expressed through opposing spoken words and gestures or humor expressed through utterances and tone of voice, remain challenging. In this paper, we introduce an approach to enhance multimodal models, which we call Multimodal Mixtures of Experts (MMoE). The key idea in MMoE is to train separate expert models for each type of multimodal interaction such as redundancy present in both modalities, uniqueness in one modality, or synergy that emerges when both modalities are fused. On a sarcasm detection task (MUSTARD) and a humor detection task (URFUNNY), we obtain new state-of-the-art results. MMoE is also able to be applied to various types of models to gain improvement.

Nov 12 (Tue) 11:00-12:30 - Jasmine

OmAgent: A Multi-modal Agent Framework for Complex Video Understanding with Task Divide-and-Conquer

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, Kyusong Lee

Recent advancements in Large Language Models (LLMs) have expanded their capabilities to multimodal contexts, including comprehensive video understanding. However, processing extensive videos such as 24-hour CCTV footage or full-length films presents significant challenges due to the vast data and processing demands. Traditional methods, like extracting key frames or converting frames to text, often result in substantial information loss. To address these shortcomings, we develop OmAgent, efficiently stores and retrieves relevant video frames for specific queries, preserving the detailed content of videos. Additionally, it features an Divide-and-Conquer Loop capable of autonomous reasoning, dynamically invoking APIs and tools to enhance query processing and accuracy. This approach ensures robust video understanding, significantly reducing information loss. Experimental results affirm OmAgent's efficacy in handling various types of videos and complex tasks. Moreover, we have endowed it with greater autonomy and a robust tool-calling system, enabling it to accomplish even more intricate tasks.

Nov 12 (Tue) 11:00-12:30 - Jasmine

VLEU: a Method for Automatic Evaluation for Generalizability of Text-to-Image Models

Jingtao Cao, Zhang Zheng, Hongru WANG, Kam-Fai Wong

Progress in Text-to-Image (T2I) models has significantly advanced the generation of images from textual descriptions. Existing metrics, such as CLIP, effectively measure the semantic alignment between single prompts and their corresponding images. However, they fall short in evaluating a model's ability to generalize across a broad spectrum of textual inputs. To address this gap, we propose the VLEU (Visual Language Evaluation Understudy) metric. VLEU leverages the power of Large Language Models (LLMs) to sample from the visual text domain, encompassing the entire range of potential inputs for the T2I task, to generate a wide variety of visual text. The images generated by T2I models from these prompts are then assessed for their alignment with the input text using the CLIP model. VLEU quantitatively measures a model's generalizability by computing the Kullback-Leibler (KL) divergence between the visual text marginal distribution and the conditional distribution over the images generated by the model. This provides a comprehensive metric for comparing the overall generalizability of T2I models, beyond single-prompt evaluations, and offers valuable insights during the finetuning process. Our experimental results demonstrate VLEU's effectiveness in evaluating the generalizability of various T2I models, positioning it as an essential metric for future research and development in image synthesis from text prompts. Our code and data will be publicly available at <https://github.com/mio7690/VLEU>.

Nov 12 (Tue) 11:00-12:30 - Jasmine

TroL: Traversal of Layers for Large Language and Vision Models

Byoung-Kwan Lee, Sangyung Chung, Chae Won Kim, Beomchan Park, Yong Man Ro

Large language and vision models (LLVMs) have been driven by the generalization power of large language models (LLMs) and the advent of visual instruction tuning. Along with scaling them up directly, these models enable LLVMs to showcase powerful vision language (VL) performances by covering diverse tasks via natural language instructions. However, existing open-source LLVMs that perform comparably to closed-source LLVMs such as GPT-4V are often considered too large (e.g., 26B, 34B, and 110B parameters), having a larger number of layers. These large models demand costly, high-end resources for both training and inference. To address this issue, we present a new efficient LLVM family with 1.8B, 3.8B, and 7B LLM model sizes, Traversal of Layers (TroL), which enables the reuse of layers in a token-wise manner. This layer traversing technique simulates the effect of looking back and retracing the answering stream while increasing the number of forward propagation layers without physically adding more layers. We demonstrate that TroL employs a simple layer traversing approach yet efficiently outperforms the open-source LLVMs with larger model sizes and rivals the performances of the closed-source LLVMs with substantial sizes.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Learning from Feedback with Coupled Comprehension and Generation

Mustafa Omer Gul, Yoav Artzi

Systems with both language comprehension and generation capabilities can benefit from the tight connection between the two. This work studies coupling comprehension and generation with focus on continually learning from interaction with users. We propose techniques to tightly integrate the two capabilities for both learning and inference. We situate our studies in two-player reference games, and deploy various models for thousands of interactions with human users, while learning from interaction feedback signals. We show dramatic improvements in performance over time, with comprehension-generation coupling leading to performance improvements up to 26% in absolute terms and up to 17% higher accuracies compared to a non-coupled system. Our analysis also shows coupling has substantial qualitative impact on the system's language, making it significantly more human-like.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering

Adjali Omar, Olivier Ferret, Sahar Ghannay, Hervé Le Borgne

The Knowledge-Aware Visual Question Answering about Entity task aims to disambiguate entities using textual and visual information, as well as knowledge. It usually relies on two independent steps, information retrieval then reading comprehension, that do not benefit each other. Retrieval Augmented Generation (RAG) offers a solution by using generated answers as feedback for retrieval training. RAG usually relies solely on pseudo-relevant passages retrieved from external knowledge bases which can lead to ineffective answer generation. In this work, we propose a multi-level information RAG approach that enhances answer generation through entity retrieval and query expansion. We formulate a joint-training RAG loss such that answer generation is conditioned on both entity and passage retrievals. We show through experiments new state-of-the-art performance on the VQuAE KB-VQA benchmark and demonstrate that our approach can help retrieve more actual relevant knowledge to generate accurate answers.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Investigating the Role of Instruction Variety and Task Difficulty in Robotic Manipulation Tasks

Amit Parekh, Nikolas Vitsakis, Alessandro Soglia, Ioannis Konstas

Evaluating the generalisation capabilities of multimodal models based solely on their performance on out-of-distribution data fails to capture their true robustness. This work introduces a comprehensive evaluation framework that systematically examines the role of instructions and inputs in the generalisation abilities of such models, considering architectural design, input perturbations across language and vision modalities, and increased task complexity. The proposed framework uncovers the resilience of multimodal models to extreme instruction perturbations and their vulnerability to observational changes, raising concerns about overfitting to spurious correlations. By employing this evaluation framework on current Transformer-based multimodal models for robotic manipulation tasks, we uncover limitations and suggest future advancements should focus on architectural and training innovations that better integrate multimodal inputs, enhancing a model's generalisation prowess by prioritising sensitivity to input content over incidental correlations.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Grounding Language in Multi-Perspective Referential Communication

Zineng Tang, Lingjun Mao, Alane Suhr

We introduce a task and dataset for referring expression generation and comprehension in multi-agent embodied environments. In this task, two agents in a shared scene must take into account one another's visual perspective, which may be different from their own, to both produce and understand references to objects in a scene and the spatial relations between them. We collect a dataset of 2,970 human-written referring expressions, each paired with human comprehension judgments, and evaluate the performance of automated models as speakers and listeners paired with human partners, finding that model performance in both reference generation and comprehension lags behind that of pairs of human agents. Finally, we experiment training an open-weight speaker model with evidence of communicative success when paired with a listener, resulting in an improvement from 58.9 to 69.3% in communicative success and even outperforming the strongest proprietary model.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Whiteboard-of-Thought: Thinking Step-by-Step Across Modalities

Sachit Menon, Richard Zemel, Carl Vondrick

When presented with questions involving visual thinking, humans naturally switch reasoning modalities, often forming mental images or drawing visual aids. Large language models have shown promising results in arithmetic and symbolic reasoning by expressing intermediate reasoning in text as a chain of thought, yet struggle to extend this capability to answer text queries that are easily solved by visual reasoning, even with extensive multimodal pretraining. We introduce a simple method, *whiteboard-of-thought* prompting, to unlock the visual reasoning capabilities of multimodal large language models across modalities. Whiteboard-of-thought prompting provides multimodal large language models with a metaphorical ‘whiteboard’ to draw out reasoning steps as images, then returns these images back to the model for further processing. We find this can be accomplished with no demonstrations or specialized modules, instead leveraging models’ existing ability to write code with libraries such as Matplotlib and Turtle. This simple approach shows state-of-the-art results on four difficult natural language tasks that involve visual and spatial reasoning. We identify multiple settings where GPT-4o using chain-of-thought fails dramatically, including more than one where it achieves 0% accuracy, while whiteboard-of-thought enables up to 92% accuracy in these same settings. We present a detailed exploration of where the technique succeeds as well as its sources of error.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Visual Text Matters: Improving Text-KVQA with Visual Text Entity Knowledge-aware Large Multimodal Assistant

Abhirama Subramanyam Penamakuri, Anand Mishra

We revisit knowledge-aware text-based visual question answering, also known as Text-KVQA in the light of modern advancements in large multimodal models (LMMs), and make the following contributions: (i) We propose VisTEL – a principled approach to perform visual text entity linking. The proposed VisTEL module harnesses a state-of-the-art visual text recognition engine and the power of a large multimodal model to jointly reason using textual and visual context obtained using surrounding cues in the image to link visual text entity to the correct knowledge base entity. (ii) We present KaLMA – knowledge-aware large multimodal assistant that augments an LMM with knowledge associated with visual text entity in the image to arrive at an accurate answer. Further, we provide a comprehensive experimental analysis and comparison of our approach with traditional visual question answering, pre-large multimodal models, and large multimodal models, as well as prior top-performing approaches. Averaging over three splits of Text-KVQA, our proposed approach surpasses the previous best approach by a substantial 23.3% on an absolute scale and establishes a new state of the art. We make our implementation publicly available.

Nov 12 (Tue) 11:00-12:30 - Jasmine

IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning

Soeun Lee, Si-Woo Kim, Taewhan Kim, Dong-Jin Kim

Recent advancements in image captioning have explored text-only training methods to overcome the limitations of paired image-text data. However, existing text-only training methods often overlook the modality gap between using text data during training and employing images during inference. To address this issue, we propose a novel approach called Image-like Retrieval, which aligns text features with visually relevant

vant features to mitigate the modality gap. Our method further enhances the accuracy of generated captions by designing a fusion module that integrates retrieved captions with input features. Additionally, we introduce a Frequency-based Entity Filtering technique that significantly improves caption quality. We integrate these methods into a unified framework, which we refer to as IFCap (**I**mage-like Retrieval and **F**requency-based Entity Filtering for Zero-shot **C**aptioning). Through extensive experimentation, our straightforward yet powerful approach has demonstrated its efficacy, outperforming the state-of-the-art methods by a significant margin in both image captioning and video captioning compared to zero-shot captioning based on text-only training.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Bayesian Example Selection Improves In-Context Learning for Speech, Text, and Visual Modalities

Siyin Wang, Chao-Han Huck Yang, Ji Wu, Chao Zhang

Large language models (LLMs) can adapt to new tasks through in-context learning (ICL) based on a few examples presented in dialogue history without any model parameter update. Despite such convenience, the performance of ICL heavily depends on the quality of the in-context examples presented, which makes the in-context example selection approach a critical choice. This paper proposes a novel eBayesian in-Context example Selection method (ByCS) for ICL. Extending the inference probability conditioned on in-context examples based on Bayes' theorem, ByCS focuses on the inverse inference conditioned on test input. Following the assumption that accurate inverse inference probability (likelihood) will result in accurate inference probability (posterior), in-context examples are selected based on their inverse inference results. Diverse and extensive cross-tasking and cross-modality experiments are performed with speech, text, and image examples. Experimental results show the efficacy and robustness of our ByCS method on various models, tasks and modalities.

Nov 12 (Tue) 11:00-12:30 - Jasmine

An Empirical Analysis on Spatial Reasoning Capabilities of Large Multimodal Models

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, Yuan-Fang Li

Large Multimodal Models (LMMs) have achieved strong performance across a range of vision and language tasks. However, their spatial reasoning capabilities are under-investigated. In this paper, we construct a novel VQA dataset, Spatial-MM, to comprehensively study LMMs' spatial understanding and reasoning capabilities. Our analyses on object-relationship and multi-hop reasoning reveal several important findings. Firstly, bounding boxes and scene graphs, even synthetic ones, can significantly enhance LMMs' spatial reasoning. Secondly, LMMs struggle more with questions posed from the human perspective than the camera perspective about the image. Thirdly, chain of thought (CoT) prompting does not improve model performance on complex multi-hop questions involving spatial relations. Moreover, spatial reasoning steps are much less accurate than non-spatial ones across MLLMs. Lastly, our perturbation analysis on GQA-spatial reveals that LMMs are much stronger at basic object detection than complex spatial reasoning. We believe our new benchmark dataset and in-depth analyses can spark further research on LMMs spatial reasoning.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Nearest Neighbor Normalization Improves Multimodal Retrieval

Neil Chowdhury, Franklin Wang, Sumeeth Shenoy, Douwe Kiela, Sarah Schwettmann, Tristan Thrush

Multimodal models leverage large-scale pretraining to achieve strong but still imperfect performance on tasks such as image captioning, visual question answering, and cross-modal retrieval. In this paper, we present a simple and efficient method for correcting errors in trained contrastive image-text retrieval models with no additional training, called Nearest Neighbor Normalization (NNN). We show an improvement on retrieval metrics in both text retrieval and image retrieval for all of the contrastive models that we tested (CLIP, BLIP, ALBEF, SigLIP, BEiT) and for both of the datasets that we used (MS-COCO and Flickr30k). NNN requires a reference database, but does not require any training on this database, and can even increase the retrieval accuracy of a model after finetuning.

Nov 12 (Tue) 11:00-12:30 - Jasmine

PropTest: Automatic Property Testing for Improved Visual Programming

Jaywon Koo, Ziyuan Yang, Paola Cascante-Bonilla, Baishakhi Ray, Vicente Ordonez

Visual Programming has recently emerged as an alternative to end-to-end black-box visual reasoning models. This type of method leverages Large Language Models (LLMs) to generate the source code for an executable computer program that solves a given problem. This strategy has the advantage of offering an interpretable reasoning path and does not require finetuning a model with task-specific data. We propose PropTest, a general strategy that improves visual programming by further using an LLM to generate code that tests for visual properties in an initial round of proposed solutions. Our method generates tests for data-type consistency, output syntax, and semantic properties. PropTest achieves comparable results to state-of-the-art methods while using publicly available LLMs. This is demonstrated across different benchmarks on visual question answering and referring expression comprehension. Particularly, PropTest improves ViperGPT by obtaining 46.1% accuracy (+6.0%) on GQA using Llama3-8B and 59.5% (+8.1%) on RefCOCO+ using CodeLlama-34B.

Nov 12 (Tue) 11:00-12:30 - Jasmine

MMedAgent: Learning to Use Medical Tools with Multi-modal Agent

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, Yixin Wang

Multi-Modal Large Language Models (MLLMs), despite being successful, exhibit limited generality and often fall short when compared to specialized models. Recently, LLM-based agents have been developed to address these challenges by selecting appropriate specialized models as tools based on user inputs. However, such advancements have not been extensively explored within the medical domain. To bridge this gap, this paper introduces the first agent explicitly designed for the medical field, named Multi-modal Medical Agent (MMedAgent). We curate an instruction-tuning dataset comprising six medical tools solving seven tasks across five modalities, enabling the agent to choose the most suitable tools for a given task. Comprehensive experiments demonstrate that MMedAgent achieves superior performance across a variety of medical tasks compared to state-of-the-art open-source methods and even the closed-source model, GPT-4o. Furthermore, MMedAgent exhibits efficiency in updating and integrating new medical tools.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Can Textual Unlearning Solve Cross-Modality Safety Alignment?

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael B. Abu-Ghazaleh, M. Salman Asif, Yue Dong, Amit Roy-Chowdhury, Chengyu Song

Recent studies reveal that integrating new modalities into large language models (LLMs), such as vision-language models (VLMs), creates a new attack surface that bypasses existing safety training techniques like supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF). While further SFT and RLHF-based safety training can be conducted in multi-modal settings, collecting multi-modal training datasets poses a significant challenge. Inspired by the structural design of recent multi-modal models, where all input modalities are ultimately fused into the language space, we explore whether unlearning solely in the textual domain can be effective for cross-modality safety alignment. Our empirical evaluation across seven datasets demonstrates promising transferability — textual unlearning in VLMs significantly reduces the Attack Success Rate (ASR) to less than 8% and in some cases, even as low as nearly 2% for both text-based and vision-text-based attacks, alongside preserving the utility. Moreover, our experiments show that unlearning with a multi-modal dataset offers no potential benefits but incurs significantly increased computational demands.

Nov 12 (Tue) 11:00-12:30 - Jasmine

What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models

Junho Kim, KIM YEONJU, Yong Man Ro

This paper presents a way of enhancing the reliability of Large Multi-modal Models (LMMs) in addressing hallucination, where the models generate cross-modal inconsistent responses. Without additional training, we propose Counterfactual Inception, a novel method that implants counterfactual thinking into LMMs using self-generated counterfactual keywords. Our method is grounded in the concept of counterfactual thinking, a cognitive process where human considers alternative realities, enabling more extensive context exploration. Bridging the human cognition mechanism into LMMs, we aim for the models to engage with and generate responses that span a wider contextual scene understanding, mitigating hallucinatory outputs. We further introduce Plausibility Verification Process (PVP), a simple yet robust keyword constraint that effectively filters out sub-optimal keywords to enable the consistent triggering of counterfactual thinking in the model responses. Comprehensive analyses across various LMMs, including both open-source and proprietary models, corroborate that counterfactual thinking significantly reduces hallucination and helps to broaden contextual understanding based on true visual clues.

Nov 12 (Tue) 11:00-12:30 - Jasmine

Beyond Single-Audio: Advancing Multi-Audio Processing in Audio Large Language Models

Yiming Chen, Xianghai Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T. Tan, Haizhou Li

Various audio-LLMs (ALLMs) have been explored recently for tackling different audio tasks simultaneously using a single, unified model. While existing evaluations of ALLMs primarily focus on single-audio tasks, real-world applications often involve processing multiple audio streams simultaneously. To bridge this gap, we propose the first multi-audio evaluation (MAE) benchmark that consists of 20 datasets from 11 multi-audio tasks encompassing both speech and sound scenarios. Comprehensive experiments on MAE demonstrate that the existing ALLMs, while being powerful in comprehending primary audio elements in individual audio inputs, struggle to handle multi-audio scenarios. To this end, we propose a novel multi-audio-LLM (MALLM) to capture audio context among multiple similar audios using discriminative learning on our proposed synthetic data. The results demonstrate that the proposed MALLM outperforms all baselines and achieves high data efficiency using synthetic data without requiring human annotations. The proposed MALLM opens the door for ALLMs towards multi-audio processing era and brings us closer to replicating human auditory capabilities in machines.

Nov 12 (Tue) 11:00-12:30 - Jasmine

LLM-A*: Large Language Model Enhanced Incremental Heuristic Search on Path Planning

Silin Meng, Yiyi Wang, Cheng-Fu Yang, Nanyun Peng, Kai-Wei Chang

Path planning is a fundamental scientific problem in robotics and autonomous navigation, requiring the derivation of efficient routes from starting to destination points while avoiding obstacles. Traditional algorithms like A* and its variants are capable of ensuring path validity but suffer from significant computational and memory inefficiencies as the state space grows. Conversely, large language models (LLMs) excel in broader environmental analysis through contextual understanding, providing global insights into environments. However, they fall short in detailed spatial and temporal reasoning, often leading to invalid or inefficient routes. In this work, we propose LLM-A*, an new LLM based route planning method that synergistically combines the precise pathfinding capabilities of A* with the global reasoning capability of LLMs. This hybrid approach aims to enhance pathfinding efficiency in terms of time and space complexity while maintaining the integrity of path validity, especially in large-scale scenarios. By integrating the strengths of both methodologies, LLM-A* addresses the computational and memory limitations of conventional algorithms without compromising on the validity required for effective pathfinding.

Nov 12 (Tue) 11:00-12:30 - Jasmine

From the Least to the Most: Building a Plug-and-Play Visual Reasoner via Data Synthesis

Chuanqi Cheng, Jian Guan, Wei Wu, Rui Yan

We explore multi-step reasoning in vision-language models (VLMs). The problem is challenging, as reasoning data consisting of multiple steps of visual and language processing are barely available. To overcome the challenge, we first introduce a least-to-most visual reasoning paradigm, which interleaves steps of decomposing a question into sub-questions and invoking external tools for resolving sub-questions. Based on the paradigm, we further propose a novel data synthesis approach that can automatically create questions and multi-step reasoning paths for an image in a bottom-up manner. Our approach divides the complex synthesis task into a few simple sub-tasks, and (almost entirely) relies on open-sourced models to accomplish the sub-tasks. Therefore, the entire synthesis process is reproducible and cost-efficient, and the synthesized data is quality guaranteed. With the approach, we construct 50k visual reasoning examples. Then, we develop a visual reasoner through supervised fine-tuning, which is capable of generally enhancing the reasoning abilities of a wide range of existing VLMs in a plug-and-play fashion. Extensive experiments indicate that the visual reasoner can consistently and significantly improve four VLMs on four VQA benchmarks.

NLP Applications 1

Nov 12 (Tue) 11:00-12:30 - Room: Riverfront Hall

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

When LLMs Meets Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, Julien Epps

Depression is a critical concern in global mental health, prompting extensive research into AI-based detection methods. Among various AI technologies, Large Language Models (LLMs) stand out for their versatility in healthcare applications. However, the application of LLMs in the identification and analysis of depressive states remains relatively unexplored, presenting an intriguing avenue for future research. In this paper, we present an innovative approach to employ an LLM in the realm of depression detection, integrating acoustic speech information into the LLM framework for this specific application. We investigate an efficient method for automatic depression detection by integrating speech signals into LLMs utilizing Acoustic Landmarks. This approach is not only valuable for the detection of depression but also represents a new perspective in enhancing the ability of LLMs to comprehend and process speech signals. By incorporating acoustic landmarks, which are specific to the pronunciation of spoken words, our method adds critical dimensions to text transcripts. This integration also provides insights into the unique speech patterns of individuals, revealing the potential mental states of individuals. By encoding acoustic landmarks information into LLMs, evaluations of the proposed approach on the DAIC-WOZ dataset reveal state-of-the-art results when compared with existing Audio-Text baselines.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

A New Pipeline for Knowledge Graph Reasoning Enhanced by Large Language Models Without Fine-Tuning

Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, Yong Dou

Conventional Knowledge Graph Reasoning (KGR) models learn the embeddings of KG components over the structure of KGs, but their performances are limited when the KGs are severely incomplete. Recent LLM-enhanced KGR models input KG structural information into LLMs. However, they require fine-tuning on open-source LLMs and are not applicable to closed-source LLMs. Therefore, in this paper, to leverage the knowledge in LLMs without fine-tuning to assist and enhance conventional KGR models, we propose a new three-stage pipeline, including knowledge alignment, KG reasoning and entity reranking. Specifically, in the alignment stage, we propose three strategies to align the knowledge in LLMs to the KG schema by explicitly associating unconnected nodes with semantic relations. Based on the enriched KGs, we train structure-aware KGR models to integrate aligned knowledge to original knowledge existing in KGs. In the reranking stage, after obtaining the results of KGR models, we rerank the top-scored entities with LLMs to recall correct answers further. Experiments show our pipeline can enhance the KGR performance in both incomplete and general situations. Code and datasets are available.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Watch Every Step! LLM Agent Learning via Iterative Step-level Process Refinement

Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, Sujian Li

Large language model agents have exhibited exceptional performance across a range of complex interactive tasks. Recent approaches have utilized tuning with expert trajectories to enhance agent performance, yet they primarily concentrate on outcome rewards, which may lead to errors or suboptimal actions due to the absence of process supervision signals. In this paper, we introduce the ***iterative step-level ***process ***R***inement ***IPR*** framework, which provides detailed step-by-step guidance to enhance agent training. Specifically, we adopt the Monte Carlo method to estimate step-level rewards. During each iteration, the agent explores along the expert trajectory and generates new actions. These actions are then evaluated against the corresponding step of expert trajectory using step-level rewards. Such comparison helps identify discrepancies, yielding contrastive action pairs that serve as training data for the agent. Our experiments on three complex agent tasks demonstrate that our framework outperforms a variety of strong baselines. Moreover, our analytical finds highlight the effectiveness of IPR in augmenting action efficiency and its applicability to diverse models.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Retrieved Sequence Augmentation for Protein Representation Learning

Chang Ma, Haiteng Zhao, Lin Zheng, Jiayi Xin, Qintong Li, Lijian Wu, Zhihong Deng, Yang Young Lu, Qi Liu, Sheng Wang, Lingpeng Kong
 Protein Language Models traditionally depend on Multiple Sequence Alignments (MSA) to incorporate evolutionary knowledge. However, MSA-based approaches suffer from substantial computational overhead and generally underperform in generalizing to de novo proteins. This study reevaluates the role of MSA, proposing it as a retrieval augmentation method and questioning the necessity of sequence alignment. We show that a simple alternative, Retrieved Sequence Augmentation (RSA), can enhance protein representation learning without the need for alignment and cumbersome preprocessing. RSA surpasses MSA Transformer by an average of 5% in both structural and property prediction tasks while being 373 times faster. Additionally, RSA demonstrates enhanced transferability for predicting de novo proteins. This methodology addresses a critical need for efficiency in protein prediction and can be rapidly employed to identify homologous sequences, improve representation learning, and enhance the capacity of Large Language Models to interpret protein structures.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Do We Need Language-Specific Fact-Checking Models? The Case of Chinese

Caigi Zhang, Zhijiang Guo, Andreas Vlachos

This paper investigates the potential benefits of language-specific fact-checking models, focusing on the case of Chinese using CHEF dataset. To better reflect real-world fact-checking, we first develop a novel Chinese document-level evidence retriever, achieving state-of-the-art performance. We then demonstrate the limitations of translation-based methods and multilingual language models, highlighting the need for language-specific systems. To better analyze token-level biases in different systems, we construct an adversarial dataset based on the CHEF dataset, where each instance has a large word overlap with the original one but holds the opposite veracity label. Experimental results on the CHEF dataset and our adversarial dataset show that our proposed method outperforms translation-based methods and multilingual language models and is more robust toward biases, emphasizing the importance of language-specific fact-checking systems.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Aligning Language Models to Explicitly Handle Ambiguity

Hyueung Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, Taeuk Kim

In interactions between users and language model agents, user utterances frequently exhibit ellipsis (omission of words or phrases) or imprecision (lack of exactness) to prioritize efficiency. This can lead to varying interpretations of the same input based on different assumptions or background knowledge. It is thus crucial for agents to adeptly handle the inherent ambiguity in queries to ensure reliability. However, even state-of-the-art large language models (LLMs) still face challenges in such scenarios, primarily due to the following hurdles: (1) LLMs are not explicitly trained to deal with ambiguous utterances; (2) the degree of ambiguity perceived by the LLMs may vary depending on the possessed knowledge. To address these issues, we propose Alignment with Perceived Ambiguity (APA), a novel pipeline that aligns LLMs to manage ambiguous queries by leveraging their own assessment of ambiguity (i.e., perceived ambiguity). Experimental results on question-answering datasets demonstrate that APA empowers LLMs to explicitly detect and manage ambiguous queries while retaining the ability to answer clear questions. Furthermore, our finding proves that APA excels beyond training with gold-standard labels, especially in out-of-distribution scenarios. The data and code are available at <https://github.com/heyjoonkim/APA>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

I Need Help! Evaluating LLMs Ability to Ask for Users Support: A Case Study on Text-to-SQL Generation

Cheng-Kuang Wu, Zhi Rui Tam, Chao-Chung Wu, Chieh-Yen Lin, Hung-yi Lee, Yun-Nung Chen

This study explores the proactive ability of LLMs to seek user support. We propose metrics to evaluate the trade-off between performance improvements and user burden, and investigate whether LLMs can determine when to request help under varying information availability. Our experiments show that without external feedback, many LLMs struggle to recognize their need for user support. The findings highlight the importance of external signals and provide insights for future research on improving support-seeking strategies. Source code: <https://github.com/appier-research/i-need-help>

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, Hung-yi Lee

Using large language models (LLMs) for automatic evaluation has become an important evaluation method in NLP research. However, it is unclear whether these LLM-based evaluators can be effectively applied in real-world classrooms to assess student assignments. This empirical report shares how we use GPT-4 as an automatic assignment evaluator in a university course with over 1000 students. Based on student responses, we found that LLM-based assignment evaluators are generally acceptable to students when they have free access to these tools. However, students also noted that the LLM sometimes fails to adhere to the evaluation instructions, resulting in unreasonable assessments. Additionally, we observed that students can easily manipulate the LLM to output specific strings, allowing them to achieve high scores without meeting the assignment rubric. Based on student feedback and our experience, we offer several recommendations for effectively integrating

LLMs into future classroom evaluations. Our observation also highlights potential directions for improving LLM-based evaluators, including their instruction-following ability and vulnerability to prompt hacking.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making

Xiutian Zhao, Ke Wang, Wei Peng

Modern large language models (LLMs) have exhibited cooperative synergy on complex task-solving, and collective decision-making (CDM) is a pivotal component in LLM-based multi-agent collaboration frameworks. Our survey on 52 recent such systems uncovers a severe lack of diversity, with a heavy reliance on dictatorial and plurality voting for CDM. Through the lens of social choice theory, we scrutinize widely-adopted CDM methods and identify their limitations. To enrich current landscape of LLM-based CDM, we present GEDI, an electoral CDM module that incorporates various ordinal preferential voting mechanisms. Our empirical case study across three benchmarks shows that the integration of certain CDM methods can markedly improve the reasoning capabilities and robustness of some leading LLMs, all without requiring intricate system designs. Additionally, we find that some CDM mechanisms generate positive synergies even with as few as three agents. The voting-based methods also demonstrate robustness against single points of failure, as well as diversity in terms of hit-rate@k and subject-wise impacts.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

LLM4Decompile: Decompiling Binary Code with Large Language Models

Hanzhuo Tan, Qi Luo, Jing Li, Yuqun Zhang

Decompilation aims to convert binary code to high-level source code, but traditional tools like Ghidra often produce results that are difficult to read and execute. Motivated by the advancements in Large Language Models (LLMs), we propose LLM4Decompile, the first and largest open-source LLM series (1.3B to 33B) trained to decompile binary code. We optimize the LLM training process and introduce the LLM4Decompile-End models to decompile binary directly. The resulting models significantly outperform GPT-4o and Ghidra on the HumanEval and ExeBenchmark benchmarks by over 100% in terms of re-executability rate. Additionally, we improve the standard refinement approach to fine-tune the LLM4Decompile-Ref models, enabling them to effectively refine the decompiled code from Ghidra and achieve a further 16.2% improvement over the LLM4Decompile-End. LLM4Decompile demonstrates the potential of LLMs to revolutionize binary code decompilation, delivering remarkable improvements in readability and executability while complementing conventional tools for optimal results.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CoEvoL: Constructing Better Responses for Instruction Finetuning through Multi-Agent Cooperation

Renhao Li, Minghuan Tan, Derek F. Wong, Min Yang

In recent years, instruction fine-tuning (IFT) on large language models (LLMs) has garnered considerable attention to enhance model performance on unseen tasks. Attempts have been made on automatic construction and effective selection for IFT data. However, we posit that previous methods have not fully harnessed the potential of LLMs for enhancing data quality. The responses within IFT data could be further enhanced by leveraging the capabilities of LLMs themselves. In this paper, we propose CoEvoL, an LLM-based multi-agent cooperation framework for the improvement of responses for instructions. To effectively refine the responses, we develop an iterative framework following a `_debate-advise-edit-judge_` paradigm. A two-stage multi-agent debate strategy is further devised to ensure the diversity and reliability of editing suggestions within the framework. Empirically, models equipped with CoEvoL outperform competitive baselines evaluated by MT-Bench and AlpacaEval, demonstrating its effectiveness in enhancing instruction-following capabilities for LLMs.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Focused Large Language Models are Stable Many-Shot Learners

Peiwei Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Heda Wang, Yao Hu, Kan Li

In-Context Learning (ICL) enables large language models (LLMs) to achieve rapid task adaptation by learning from demonstrations. With the increase in available context length of LLMs, recent experiments have shown that the performance of ICL does not necessarily scale well in many-shot (demonstration) settings. We hypothesize that the reason lies in more demonstrations dispersing the model attention from the query, hindering its understanding of key content, which we validate both theoretically and experimentally. Inspired by how humans learn from examples, we propose a training-free method FocusICL, which conducts triviality filtering to avoid attention being diverted by unimportant contents at token-level and operates hierarchical attention to further ensure sufficient attention towards current query at demonstration-level. We also design an efficient hyperparameter searching strategy for FocusICL based on model perplexity of demonstrations. Comprehensive experiments validate that FocusICL achieves an average performance improvement of 5.2% over vanilla ICL and scales well with many-shot demonstrations.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts

Zhaoxuan Tan, Zheyuan Liu, Meng Jiang

Personalized large language models (LLMs) aim to tailor interactions, content, and recommendations to individual user preferences. While parameter-efficient fine-tuning (PEFT) methods excel in performance and generalization, they are costly and limit communal benefits when used individually. To this end, we introduce Personalized Pieces (Per-Pcs), a framework that allows users to safely share and assemble personalized PEFT efficiently with collaborative efforts. Per-Pcs involves selecting sharers, breaking their PEFT into pieces, and training gates for each piece. These pieces are added to a pool, from which target users can select and assemble personalized PEFT using their history data. This approach preserves privacy and enables fine-grained user modeling without excessive storage and computation demands. Experimental results show Per-Pcs outperforms non-personalized and PEFT retrieval baselines, offering performance comparable to OPPU with significantly lower resource use across six tasks. Further analysis highlights Per-Pcs' robustness concerning sharer count and selection strategy, pieces sharing ratio, and scalability in computation time and storage space. Per-Pcs' modularity promotes safe sharing, making LLM personalization more efficient, effective, and widely accessible through collaborative efforts.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Efficient LLM Comparative Assessment: A Product of Experts Framework for Pairwise Comparisons

Adrian Liusie, Vatsal Raina, Yassir Fathullah, Mark Gales

LLM-as-a-judge approaches are a practical and effective way of assessing a range of text tasks. However, when using pairwise comparisons to rank a set of candidates, the computational cost scales quadratically with the number of candidates, which has practical limitations. This paper introduces a Product of Expert (PoE) framework for efficient LLM Comparative Assessment. Here individual comparisons are considered experts that provide information on a pair's score difference. The PoE framework combines the information from these experts to yield an expression that can be maximized with respect to the underlying set of candidates, and is highly flexible where any form of expert can be assumed. When Gaussian experts are used one can derive simple closed-form solutions for the optimal candidate ranking, as well as expressions for selecting which comparisons should be made to maximize the probability of this ranking. Our approach enables efficient comparative assessment, where by using only a small subset of the possible comparisons, one can generate score predictions that correlate

well with human judgements. We evaluate the approach on multiple NLG tasks and demonstrate that our framework can yield considerable computational savings when performing pairwise comparative assessment. With many candidate texts, using as few as 2% of comparisons the PoE solution can achieve similar performance to when all comparisons are used.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale

Junying Chen, Chi Gui, Ouyang Ruiy, Anmingze Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, Benyou Wang

The rapid development of multimodal large language models (MLLMs), such as GPT-4V, has led to significant advancements. However, these models still face challenges in medical multimodal capabilities due to limitations in the quantity and quality of medical vision-text data, stemming from data privacy concerns and high annotation costs. While pioneering approaches utilize PubMed's large-scale, de-identified medical image-text pairs to address these limitations, they often fall short due to inherent data noise. To tackle this, we refined medical image-text pairs from PubMed and employed MLLMs (GPT-4V) in an 'unblinded' capacity to denoise and reformat the data, resulting in the creation of the **PubMedVision** dataset with 1.3 million medical VQA samples. Our validation demonstrates that: (1) PubMedVision can significantly enhance the medical multimodal capabilities of MLLMs, showing significant improvement in benchmarks including the MIMU Health & Medicine track; (2) manual checks by medical experts and empirical results validate the superior data quality of our dataset compared to other data construction methods. Using PubMedVision, we train a 34B medical MLLM **HuatuuoGPT-Vision***, which shows superior performance in medical multimodal scenarios among open-source MLLMs. Our code and data are available at <https://github.com/FreedomIntelligence/HuatuuoGPT-Vision>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models

Xiaochen Wang, Jiagi Wang, Houping Xiao, Jinghui Chen, Fenglong Ma

Foundation models have demonstrated remarkable capabilities in handling diverse modalities and tasks, outperforming conventional artificial intelligence (AI) approaches that are highly task-specific and modality-reliant. In the medical domain, however, the development of comprehensive foundation models is constrained by limited access to diverse modalities and stringent privacy regulations. To address these constraints, this study introduces a novel knowledge injection approach, FedKIM, designed to scale the medical foundation model within a federated learning framework. FedKIM leverages lightweight local models to extract healthcare knowledge from private data and integrates this knowledge into a centralized foundation model using a designed adaptive Multitask Multimodal Mixture Of Experts (M^3OE) module. This method not only preserves privacy but also enhances the model's ability to handle complex medical tasks involving multiple modalities. Our extensive experiments across twelve tasks in seven modalities demonstrate the effectiveness of FedKIM in various settings, highlighting its potential to scale medical foundation models without direct access to sensitive data. Source codes are available at <https://github.com/XiaochenWang-PSU/FedKIM>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

DiVERT: Distractor Generation with Variational Errors Represented as Text for Math Multiple-choice Questions

Nigel Fernandez, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Otero Ornelas, Andrew Lan

High-quality distractors are crucial to both the assessment and pedagogical value of multiple-choice questions (MCQs), where manually crafting ones that anticipate knowledge deficiencies or misconceptions among real students is difficult. Meanwhile, automated distractor generation, even with the help of large language models (LLMs), remains challenging for subjects like math. It is crucial to not only identify plausible distractors but also understand the error behind them. In this paper, we introduce DiVERT (Distractor Generation with Variational Errors Represented as Text), a novel variational approach that learns an interpretable representation of errors behind distractors in math MCQs. Through experiments on a real-world math MCQ dataset with 1,434 questions used by hundreds of thousands of students, we show that DiVERT, despite using a base open-source LLM with 7B parameters, outperforms state-of-the-art approaches using GPT-4o on downstream distractor generation. We also conduct a human evaluation with math educators and find that DiVERT leads to error labels that are of comparable quality to human-authored ones.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

MQuinE: a Cure for Z-paradox in Knowledge Graph Embedding

Yang Liu, Huang Fang, Yunfeng Cai, Mingming Sun

Knowledge graph embedding (KGE) models achieved state-of-the-art results on many knowledge graph tasks including link prediction and information retrieval. Despite the superior performance of KGE models in practice, we discover a deficiency in the expressiveness of some popular existing KGE models called Z-paradox. Motivated by the existence of Z-paradox, we propose a new KGE model called MQuinE that does not suffer from Z-paradox while preserves strong expressiveness to model various relation patterns including symmetric/asymmetric, inverse, 1-N/1-N/1-N, and composition relations with theoretical justification. Experiments on real-world knowledge bases indicate that Z-paradox indeed degrades the performance of existing KGE models, and can cause more than 20% accuracy drop on some challenging test samples. Our experiments further demonstrate that MQuinE can mitigate the negative impact of Z-paradox and outperform existing KGE models by a visible margin on link prediction tasks.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

ARM: An Alignment-and-Replacement Module for Chinese Spelling Check Based on LLMs

Changchun Liu, Kai Zhang, Junzhe Jiang, Zirui Liu, Hanqing Tao, Min Gao, Enhong Chen

Chinese Spelling Check (CSC) aims to identify and correct spelling errors in Chinese texts, where enhanced semantic understanding of a sentence can significantly improve correction accuracy. Recently, Large Language Models (LLMs) have demonstrated exceptional mastery of world knowledge and semantic understanding, rendering them more robust against spelling errors. However, the application of LLMs in CSC is a double-edged sword, as they tend to unnecessarily alter sentence length and modify rare but correctly used phrases. In this paper, by leveraging the capabilities of LLMs while mitigating their limitations, we propose a novel plug-and-play Alignment-and-Replacement Module ARM that enhances the performance of existing CSC models and without the need for retraining or fine-tuning. Experiment results and analysis on three benchmark datasets demonstrate the effectiveness and competitiveness of the proposed module.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CARER - ClinicAI Reasoning-Enhanced Representation for Temporal Health Risk Prediction

Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, Phi Le Nguyen

The increasing availability of multimodal data from electronic health records (EHR) has paved the way for deep learning methods to improve diagnosis accuracy. However, deep learning models are data-driven, requiring large-scale datasets to achieve high generalizability. Inspired by how human experts leverage reasoning for medical diagnosis, we propose CARER, a novel health risk prediction framework, that enhances deep learning models with clinical rationales derived from medically proficient Large Language Models (LLMs). In addition, we provide a cross-view alignment loss which aligns the "local" view from the patient's health status with the "global" view from the external LLM's clinical reasoning to boost the mutual feature learning. Through extensive experiments on two predictive tasks using two popular EHR datasets,

our CARER's significantly exceeds the performance of state-of-the-art models by up to 11.2%, especially in improving data efficiency and generalizability. Our code is available at <https://github.com/tuand2812/CARER-EMNLP-2024>

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Bio-RFX: Refining Biomedical Extraction via Advanced Relation Classification and Structural Constraints

Minjia Wang, Fangzhou Liu, Xiuxing Li, Bowen Dong, Zhenyu Li, Tengyu Pan, Jianyong Wang

The ever-growing biomedical publications magnify the challenge of extracting structured data from unstructured texts. This task involves two components: biomedical entity identification (Named Entity Recognition, NER) and their interrelation determination (Relation Extraction, RE). However, existing methods often neglect unique features of the biomedical literature, such as ambiguous entities, nested proper nouns, and overlapping relation triplets, and underutilize prior knowledge, leading to an intolerable performance decline in the biomedical domain, especially with limited annotated training data. In this paper, we propose the Biomedical Relation-First eXtraction (Bio-RFX) model by leveraging sentence-level relation classification before entity extraction to tackle entity ambiguity. Moreover, we exploit structural constraints between entities and relations to guide the model's hypothesis space, enhancing extraction performance across different training scenarios. Comprehensive experimental results on biomedical datasets show that Bio-RFX achieves significant improvements on both NER and RE tasks. Even under the low-resource training scenarios, it outperforms all baselines in NER and has highly competitive performance compared to the state-of-the-art fine-tuned baselines in RE.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Enhancing High-order Interaction Awareness in LLM-based Recommender Model

Xinfeng Wang, Jin Cui, Fumiyo Fukumoto, Yoshimi Suzuki

Large language models (LLMs) have demonstrated prominent reasoning capabilities in recommendation tasks by transforming them into text-generation tasks. However, existing approaches either disregard or ineffectively model the user-item high-order interactions. To this end, this paper presents an enhanced LLM-based recommender (ELMRec). We enhance whole-word embeddings to substantially enhance LLMs' interpretation of graph-constructed interactions for recommendations, without requiring graph pre-training. This finding may inspire endeavors to incorporate rich knowledge graphs into LLM-based recommenders via whole-word embedding. We also found that LLMs often recommend items based on users' earlier interactions rather than recent ones, and present a reranking solution. Our ELMRec outperforms state-of-the-art (SOTA) methods, especially achieving a 124.3% to 293.7% improvement over SOTA LLM-based methods in direct recommendations. Our code is available online.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Unknown Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, Paolo Papotti

Computational fact-checking (FC) relies on supervised models to verify claims based on given evidence, requiring a resource-intensive process to annotate large volumes of training data. We introduce Unknown, a novel framework that generates training instances for FC systems automatically using both textual and tabular content. Unknown selects relevant evidence and generates supporting and refuting claims with advanced negation artifacts. Designed to be flexible, Unknown accommodates various strategies for evidence selection and claim generation, offering unparalleled adaptability. We comprehensively evaluate Unknown on both text-only and table+text benchmarks, including Feverous, SciFact, and MMFC, a new multi-modal FC dataset. Our results prove that Unknown examples are of comparable quality to expert-labeled data, even enabling models to achieve up to 5% higher accuracy. The code, data, and models are available at <https://github.com/disi-unibolnlp/unknown>

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

PATIENT-: Using Large Language Models to Simulate Patients for Training Mental Health Professionals

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, Zhiyu Chen

Mental illness remains one of the most critical public health issues. Despite its importance, many mental health professionals highlight a disconnect between their training and actual real-world patient practice. To help bridge this gap, we propose PATIENT-, a novel patient simulation framework for cognitive behavior therapy (CBT) training. To build PATIENT-, we construct diverse patient cognitive models based on CBT principles and use large language models (LLMs) programmed with these cognitive models to act as a simulated therapy patient. We propose an interactive training scheme, PATIENT--TRAINER, for mental health trainees to practice a key skill in CBT – formulating the cognitive model of the patient – through role-playing a therapy session with PATIENT-. To evaluate PATIENT-, we conducted a comprehensive user study of 13 mental health trainees and 20 experts. The results demonstrate that practice using PATIENT--TRAINER enhances the perceived skill acquisition and confidence of the trainees beyond existing forms of training such as textbooks, videos, and role-play with non-patients. Based on the experts' perceptions, PATIENT- is perceived to be closer to real patient interactions than GPT-4, and PATIENT--TRAINER holds strong promise to improve trainee competencies. Our code and data are released at <https://github.com/ruiyiw/patient-psi>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction

Xuren Ge, Abhishek Satpathy, Ronald Dean Williams, John Stankovic, Homa Alemzadeh

Multi-label text classification (MLTC) tasks in the medical domain often face the long-tail label distribution problem. Prior works have explored hierarchical label structures to find relevant information for few-shot classes, but mostly neglected to incorporate external knowledge from medical guidelines. This paper presents DKEC, Domain Knowledge Enhanced Classification for diagnosis prediction with two innovations: (1) automated construction of heterogeneous knowledge graphs from external sources to capture semantic relations among diverse medical entities, (2) incorporating the heterogeneous knowledge graphs in few-shot classification using a label-wise attention mechanism. We construct DKEC using three online medical knowledge sources and evaluate it on a real-world Emergency Medical Services (EMS) dataset and a public electronic health record (EHR) dataset. Results show that DKEC outperforms the state-of-the-art label-wise attention networks and transformer models of different sizes, particularly for the few-shot classes. More importantly, it helps the smaller language models achieve comparable performance to large language models.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Defending Against Social Engineering Attacks in the Age of LLMs

Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, huan liu, Julia Hirschberg

The proliferation of Large Language Models (LLMs) poses challenges in detecting and mitigating digital deception, as these models can emulate human conversational patterns and facilitate chat-based social engineering (CSE) attacks. This study investigates the dual capabilities of LLMs as both facilitators and defenders against CSE threats. We develop a novel dataset, **SEConvO**, simulating CSE scenarios in academic and recruitment contexts, and designed to examine how LLMs can be exploited in these situations. Our findings reveal that, while off-the-shelf LLMs generate high-quality CSE content, their detection capabilities are suboptimal, leading to increased operational costs for defense. In response, we propose **ConvoSentinel**, a modular defense pipeline that improves detection at both the message and the conversation levels,

offering enhanced adaptability and cost-effectiveness. The retrieval-augmented module in **ConvoSentinel** identifies malicious intent by comparing messages to a database of similar conversations, enhancing CSE detection at all stages. Our study highlights the need for advanced strategies to leverage LLMs in cybersecurity. Our code and data are available at this GitHub repository: <https://github.com/lynneai/ConvoSentinel.git>.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Target-Aware Language Modeling via Granular Data Sampling

Ernie Chang, Pin-Jie Lin, Yang Li, Changsheng Zhao, Daeil Kim, Rastislav Rabatin, Zechun Liu, Yangyang Shi, Vikas Chandra

Language model pretraining generally targets a broad range of use cases and incorporates data from diverse sources. However, there are instances where we desire a model that excels in specific areas without markedly compromising performance in other areas. A cost-effective and straightforward approach is sampling with low-dimensional data features, which allows selecting large-scale pretraining data for domain-specific use cases. In this work, we revisit importance sampling with n-gram features consisting of multi-granular tokens, which strikes a good balance between sentence compression and representation capabilities. We observed the sampled data to have a high correlation with the target downstream task performance *while preserving its effectiveness on other tasks*. This leads to the proposed data sampling paradigm where language models can be pretrained more efficiently on selected documents. On eight benchmarks we demonstrate with ~1% of the data, pretrained models perform on par with the full RefinedWeb data and outperform randomly selected samples for model sizes ranging from 125M to 1.5B.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment

Zhipeng Chen, Kun Zhou, Xin Zhao, Jingyuan Wang, Ji-Rong Wen

Large language models (LLMs) are still struggling in aligning with human preference in complex tasks and scenarios. They are prone to overfit into the unexpected patterns or superficial styles in the training data. We conduct an empirical study that only selects the top-10% most updated parameters in LLMs for alignment training, and see improvements in the convergence process and final performance. It indicates the existence of redundant neurons in LLMs for alignment training. To reduce its influence, we propose a low-redundant alignment method named **ALLO***, focusing on optimizing the most related neurons with the most useful supervised signals. Concretely, we first identify the neurons that are related to the human preference data by a gradient-based strategy, then identify the alignment-related key tokens by reward models for computing loss. Besides, we also decompose the alignment process into the forgetting and learning stages, where we first forget the tokens with unaligned knowledge and then learn aligned knowledge, by updating different ratios of neurons, respectively. Experimental results on 10 datasets have shown the effectiveness of ALLO. Our code and data will be publicly released.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

ECCO: Can We Improve Model-Generated Code Efficiency Without Sacrificing Functional Correctness?

Siddhant Waghjale, Vishruth Veerendranath, Zhiruo Wang, Daniel Fried

Although large language models (LLMs) have been largely successful in generating functionally correct programs, conditioning models to produce efficient solutions while ensuring correctness remains a challenge. Further, unreliability in benchmarking code efficiency is a hurdle across varying hardware specifications for popular interpreted languages such as Python. In this paper, we present ECCO, a reproducible benchmark for evaluating program efficiency via two paradigms: natural language (NL) based code generation and history-based code editing. On ECCO, we adapt and thoroughly investigate the three most promising existing LLM-based approaches: in-context learning, iterative refinement with execution or NL feedback, and fine-tuning conditioned on execution and editing history. While most methods degrade functional correctness and moderately increase program efficiency, we find that adding execution information often helps maintain functional correctness, and NL feedback enhances more on efficiency. We release our benchmark to support future work on LLM-based generation of efficient code.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Effective Synthetic Data and Test-Time Adaptation for OCR Correction

Shuhao Guan, Cheng Xu, Moule Lin, Derek Greene

Post-OCR technology is used to correct errors in the text produced by OCR systems. This study introduces a method for constructing post-OCR synthetic data with different noise levels using weak supervision. We define Character Error Rate (CER) thresholds for "effective" and "ineffective" synthetic data, allowing us to create more useful multi-noise level synthetic datasets. Furthermore, we propose Self-Correct-Noise Test-Time Adaptation (SCN-TTA), which combines self-correction and noise generation mechanisms. SCN-TTA allows a model to dynamically adjust to test data without relying on labels, effectively handling proper nouns in long texts and further reducing CER. In our experiments we evaluate a range of models, including multiple PLMs and LLMs. Results indicate that our method yields models that are effective across diverse text types. Notably, the ByT5 model achieves a CER reduction of 68.67% without relying on manually annotated data

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Autoregressive Multi-trait Essay Scoring via Reinforcement Learning with Scoring-aware Multiple Rewards

Heejin Do, Sangwon Ryu, Gary Lee

Recent advances in automated essay scoring (AES) have shifted towards evaluating multiple traits to provide enriched feedback. Like typical AES systems, multi-trait AES employs the quadratic weighted kappa (QWK) to measure agreement with human raters, aligning closely with the rating schema; however, its non-differentiable nature prevents its direct use in neural network training. In this paper, we propose Scoring-aware Multi-reward Reinforcement Learning (SaMRL), which integrates actual evaluation schemes into the training process by designing QWK-based rewards with a mean-squared error penalty for multi-trait AES. Existing reinforcement learning (RL) applications in AES are limited to classification models despite associated performance degradation, as RL requires probability distributions; instead, we adopt an autoregressive score generation framework to leverage token generation probabilities for robust multi-trait score predictions. Empirical analyses demonstrate that SaMRL facilitates model training, notably enhancing scoring of previously inferior prompts.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

SEGMENT+: Long Text Processing with Short-Context Language Models

Wei Shi, Shuang Li, Kerun Yu, Jinglei Chen, Zujie Liang, Xinhui Wu, Yuxi Qian, Feng Wei, Bo Zheng, Jiaqing Liang, Jiangjie Chen, Yanghua Xiao

There is a growing interest in expanding the input capacity of language models (LMs) across various domains. However, simply increasing the context window does not guarantee robust performance across diverse long-input processing tasks, such as understanding extensive documents and extracting detailed information from lengthy and noisy data. In response, we introduce Segment+, a general framework that enables LMs to handle extended inputs within limited context windows efficiently. Segment+ utilizes structured notes and a filtering module to manage information flow, resulting in a system that is both controllable and interpretable. Our extensive experiments across various model sizes, focusing on long-document question-answering and Needle-in-a-Haystack tasks, demonstrate the effectiveness of Segment+ in improving performance.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CoSafe: Evaluating Large Language Model Safety in Multi-Turn Dialogue Coreference

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, GAO Zuchen, Fei Mi, Lanqing HONG

As large language models (LLMs) constantly evolve, ensuring their safety remains a critical research issue. Previous red teaming approaches for LLM safety have primarily focused on single prompt attacks or goal hijacking. To the best of our knowledge, we are the first to study LLM safety in multi-turn dialogue coreference. We created a dataset of 1,400 questions across 14 categories, each featuring multi-turn coreference safety attacks. We then conducted detailed evaluations on five widely used open-source LLMs. The results indicated that under multi-turn coreference safety attacks, the highest attack success rate was 56% with the LLaMA2-Chat-7b model, while the lowest was 13.9% with the Mistral-7B-Instruct model. These findings highlight the safety vulnerabilities in LLMs during dialogue coreference interactions.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Automated Essay Scoring: A Reflection on the State of the Art

Shengjie Li, Vincent Ng

While steady progress has been made on the task of automated essay scoring (AES) in the past decade, much of the recent work in this area has focused on developing models that beat existing models on a standard evaluation dataset. While improving performance numbers remains an important goal in the short term, such a focus is not necessarily beneficial for the long-term development of the field. We reflect on the state of the art in AES research, discussing issues that we believe can encourage researchers to think bigger than improving performance numbers with the ultimate goal of triggering discussion among AES researchers on how we should move forward.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

Tian Liang, Zhiwei He, Wenxiang Xiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, Zhaopeng Tu

Modern large language models (LLMs) like ChatGPT have shown remarkable performance on general language tasks but still struggle on complex reasoning tasks, which drives the research on cognitive behaviors of LLMs to explore human-like problem-solving strategies. Along this direction, one representative strategy is self-reflection, which asks an LLM to refine the solution with the feedback generated by itself iteratively. However, our study shows that such reflection-style methods suffer from the Degeneration-of-Thought (DoT) problem: once the LLM has established confidence in its solutions, it is unable to generate novel thoughts later through reflection even if its initial stance is incorrect. To address the DoT problem, we propose a Multi-Agent Debate (MAD) framework, in which multiple agents express their arguments in the state of "tit for tat" and a judge manages the debate process to obtain a final solution. Clearly, our MAD framework encourages divergent thinking in LLMs which would be helpful for tasks that require deep levels of contemplation. Experiment results on two challenging datasets, commonsense machine translation and counter-intuitive arithmetic reasoning, demonstrate the effectiveness of our MAD framework. Extensive analyses suggest that the adaptive break of debate and the modest level of "tit for tat" state are required for MAD to obtain good performance. Moreover, we find that LLMs might not be a fair judge if different LLMs are used for agents.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CURE: Context- and Uncertainty-Aware Mental Disorder Detection

Migyeong Kang, goun choi, Hyolim Jeon, Ji hyun An, Daejin Choi, Jinyoung Han

As the explainability of mental disorder detection models has become important, symptom-based methods that predict disorders from identified symptoms have been widely utilized. However, since these approaches focused on the presence of symptoms, the context of symptoms can be often ignored, leading to missing important contextual information related to detecting mental disorders. Furthermore, the result of disorder detection can be vulnerable to errors that may occur in identifying symptoms. To address these issues, we propose a novel framework that detects mental disorders by leveraging symptoms and their context while mitigating potential errors in symptom identification. In this way, we propose to use large language models to effectively extract contextual information and introduce an uncertainty-aware decision fusion network that combines predictions of multiple models based on quantified uncertainty values. To evaluate the proposed method, we constructed a new Korean mental health dataset annotated by experts, named KoMOS. Experimental results demonstrate that the proposed model accurately detects mental disorders even in situations where symptom information is incomplete.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

ToolPlanner: A Tool Augmented LLM for Multi Granularity Instructions with Path Planning and Feedback

Qinzhuo Wu, Wei Liu, Jian Luan, Bin Wang

Recently, tool-augmented LLMs have gained increasing attention. Given an instruction, tool-augmented LLMs can interact with various external tools in multiple rounds and provide a final answer. However, previous LLMs were trained on overly detailed instructions, which included API names or parameters, while real users would not explicitly mention these API details. This leads to a gap between trained LLMs and real-world scenarios. In addition, most works ignore whether the interaction process follows the instruction. To address these issues, we constructed a training dataset called MGToolBench, which contains statement and category-level instructions to better reflect real-world scenarios. In addition, we propose ToolPlanner, a two-stage reinforcement learning framework that utilizes path planning and two feedback mechanisms to enhance the LLM's task completion and instruction-following capabilities. Experimental results show that ToolPlanner significantly improves the Match Rate, Pass Rate, and Win Rate by 26.8%, 20.2%, and 5.6% compared to the SOTA model. Human evaluation verifies that the multi-granularity instructions can better align with users' usage habits. Our data and code will be released upon acceptance.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

CoCoST: Automatic Complex Code Generation with Online Searching and Correctness Testing

Xinyi He, Jiaru Zou, Yun Lin, Mengyu Zhou, Shi Han, Zejian Yuan, Dongmei Zhang

Large Language Models have revolutionized code generation ability by converting natural language descriptions into executable code. However, generating complex code within real-world scenarios remains challenging due to intricate structures, subtle bugs, understanding of advanced data types, and lack of supplementary contents. To address these challenges, we introduce the CoCoST framework, which enhances complex code generation by online searching for more information with planned queries and correctness testing for code refinement. Moreover, CoCoST serializes the complex inputs and outputs to improve comprehension and generates test cases to ensure the adaptability for real-world applications. CoCoST is validated through rigorous experiments on the DS-1000 and ClassEval datasets. Experimental results show that CoCoST substantially improves the quality of complex code generation, highlighting its potential to enhance the practicality of LLMs in generating complex code.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Framework for Robust and Scalable Text Watermarking

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, Bryan Kian Hsiang Low

Protecting intellectual property (IP) of text such as articles and code is increasingly important, especially as sophisticated attacks become possible, such as paraphrasing by large language models (LLMs) or even unauthorized training of LLMs on copyrighted text to infringe such IP. However, existing text watermarking methods are not robust enough against such attacks nor scalable to millions of users for practical implementation. In this paper, we propose Waterfall, the first training-free framework for robust and scalable text watermarking applicable

across multiple text types (e.g., articles, code) and languages supportable by LLMs, for general text and LLM data provenance. Waterfall comprises several key innovations, such as being the first to use LLM as paraphrasers for watermarking along with a novel combination of techniques that are surprisingly effective in achieving robust verifiability and scalability. We empirically demonstrate that Waterfall achieves significantly better scalability, robust verifiability, and computational efficiency compared to SOTA article-text watermarking methods, and also showed how it could be directly applied to the watermarking of code.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

QuBE: Question-based Belief Enhancement for Agentic LLM

Minsoo Kim, Jongyoon Kim, Jihyuk Kim, seung-won hwang

Despite advancements in Large Language Models (LLMs), many complex tasks are not easily solved in a single inference step, requiring the use of agentic LLMs in interactive environments. However, agentic LLMs suffer from a phenomenon known as reasoning derailment, due to the indiscriminate incorporation of observations from partially observable environments. We introduce QuBE, a method that enhances agents' focus on task-relevant contexts, by constructing a belief state via question answering. We validate QuBE through experiments in two agentic LLM scenarios with partial observability: 1) a canonical interactive decision-making scenario using text-based game engines, and 2) an interactive retrieval-augmented generation (RAG) scenario using search engines. In the AlfWorld text-based game, QuBE outperforms established baselines by substantial margins, and in the search engine scenario, it achieves marked improvements on the BeIR zero-shot retrieval benchmark. The results demonstrate that QuBE significantly mitigates reasoning derailment, refining the decision-making process of LLM agents in partially observed environments.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

RAR: Retrieval Augmented Retrieval for Code Generation in Low Resource Languages

Avik Dutt, Mukul Singh, Gust Verbruggen, Sumit Gulwani, Vu Le

Language models struggle in generating code for low-resource programming languages, since these are underrepresented in training data. Either examples or documentation are commonly used for improved code generation. We propose to use both types of information together and present retrieval augmented retrieval (RAR) as a two-step method for selecting relevant examples and documentation. Experiments on three low-resource languages (Power Query M, OfficeScript and Excel formulas) show that RAR outperforms independently example and grammar retrieval (+2.81–26.14%). Interestingly, we show that two-step retrieval selects better examples and documentation when used independently as well.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Calibrating LLMs with Preference Optimization on Thought Trees for Generating Rationale in Science Question Scoring

Jiazheng Li, Hainiu Xu, ZHAOYUE SUN, Yuxiang Zhou, David West, Cesare Alasio, Yulan He

Generating rationales that justify scoring decisions has been a promising way to facilitate explainability in automated scoring systems. However, existing methods do not match the accuracy of classifier-based methods. Plus, the generated rationales often contain hallucinated information. To address these issues, we propose a novel framework capable of generating more faithful rationales and, more importantly, matching performance with classifier-based black-box scoring systems. We first mimic the human assessment process by querying Large Language Models (LLMs) to generate a thought tree. We then summarise intermediate assessment decisions from each thought tree path for creating synthetic rationale data and rationale preference data. Finally, we utilise the generated synthetic data to calibrate LLMs through a two-step training process: supervised fine-tuning and preference optimization. Extensive experimental results demonstrate that our framework achieves a 38% assessment performance improvement in the QWK score compared to prior work while producing higher-quality rationales, as recognised by human evaluators and LLMs. Our work sheds light on the effectiveness of performing preference optimization using synthetic preference data obtained from thought tree paths. Data and code are available at: https://github.com/jiazheng99/thought_tree_assessment.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Crisis counselor language and perceived genuine concern in crisis conversations

Greg Buda, Ignacio J. Tripodi, Margaret Meagher, Elizabeth A. Olson

Although clients perceptions of therapist empathy are known to correlate with therapy effectiveness, the specific ways that the therapists language use contributes to perceived empathy remain less understood. Natural Language Processing techniques, such as transformer models, permit the quantitative, automated, and scalable analysis of therapists verbal behaviors. Here, we present a novel approach to extract linguistic features from text-based crisis intervention transcripts to analyze associations between specific crisis counselor verbal behaviors and perceived genuine concern. Linguistic features associated with higher perceived genuine concern included positive emotional language and affirmations; features associated with lower perceived genuine concern included self-oriented talk and overuse of templates. These findings provide preliminary evidence toward pathways for automating real-time feedback to crisis counselors about clients' perception of the therapeutic relationship.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

MobileVLM: A Vision-Language Model for Better Intra- and Inter-UI Understanding

Qinzhu Wu, Weikai Xu, Wei Liu, Tao Tan, Lujianfeng, Ang Li, Jian Luan, Bin Wang, Shuo Shang

Recently, mobile AI agents based on VLMs have been gaining increasing attention. These works typically utilize VLM as a foundation, fine-tuning it with instruction-based mobile datasets. However, these VLMs are typically pre-trained on general-domain data, which often results in a lack of fundamental capabilities specific to the mobile domain. Therefore, they may struggle to recognize specific UI elements and understand intra-UI fine-grained information. In addition, the current fine-tuning task focuses on interacting with the most relevant element for the given instruction. These fine-tuned VLMs may still ignore the relationships between UI pages, neglect the roles of elements in page transitions and lack inter-UI understanding. To address issues, we propose a VLM called MobileVLM, which includes two additional pre-training stages to enhance both intra- and inter-UI understanding. We defined four UI-based pre-training tasks, enabling the model to better perceive fine-grained elements and capture page transition actions. To address the lack of mobile pre-training data, we built a large Chinese mobile dataset Mobile3M from scratch, which contains 3 million UI pages, and real-world transition actions, forming a directed graph structure. Experimental results show MobileVLM excels on both our test set and public mobile benchmarks, outperforming existing VLMs.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Efficient and Interpretable Grammatical Error Correction with Mixture of Experts

Muhammad Reza Qorib, Alham Fikri Ajii, Hwee Tou Ng

Error type information has been widely used to improve the performance of grammatical error correction (GEC) models, whether for generating corrections, re-ranking them, or combining GEC models. Combining GEC models that have complementary strengths in correcting different error types is very effective in producing better corrections. However, system combination incurs a high computational cost due to the need to run inference on the base systems before running the combination method itself. Therefore, it would be more efficient to have a single model with multiple sub-networks that specialize in correcting different error types. In this paper, we propose a mixture-of-experts model, MoECE, for grammatical error correction. Our model successfully achieves the performance of T5-XL with three times fewer effective parameters. Additionally, our model produces interpretable corrections by also identifying the error type during inference.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Can LLM Graph Reasoning Generalize beyond Pattern Memorization?

Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, Yulia Tsvetkov

Large language models (LLMs) demonstrate great potential for problems with implicit graphical structures, while recent works seek to enhance the graph reasoning capabilities of LLMs through specialized instruction tuning. The resulting “graph LLMs” are evaluated with in-distribution settings only, thus it remains underexplored whether LLMs are learning generalizable graph reasoning skills or merely memorizing patterns in the synthetic training data. To this end, we propose the NLGIFT benchmark, an evaluation suite of LLM graph reasoning generalization: whether LLMs could go beyond semantic, numeric, structural, reasoning patterns in the synthetic training data and improve utility on real-world graph-based tasks. Extensive experiments with two LLMs across four graph reasoning tasks demonstrate that while generalization on simple patterns (semantic, numeric) is somewhat satisfactory, LLMs struggle to generalize across reasoning and real-world patterns, casting doubt on the benefit of synthetic graph tuning for real-world tasks with underlying network structures. We explore three strategies to improve LLM graph reasoning generalization, and we find that while post-training alignment is most promising for real-world tasks, empowering LLM graph reasoning to go beyond pattern memorization remains an open research question.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Modeling News Interactions and Influence for Financial Market Prediction

Mengyu Wang, Shay B Cohen, Tiejun Ma

The diffusion of financial news into market prices is a complex process, making it challenging to evaluate the connections between news events and market movements. This paper introduces FININ (Financial Interconnected News Influence Network), a novel market prediction model that captures not only the links between news and prices but also the interactions among news items themselves. FININ effectively integrates multi-modal information from both market data and news articles. We conduct extensive experiments on two datasets, encompassing the S&P 500 and NASDAQ 100 indices over a 15-year period and over 2.7 million news articles. The results demonstrate FININ’s effectiveness, outperforming advanced market prediction models with an improvement of 0.429 and 0.341 in the daily Sharpe ratio for the two markets respectively. Moreover, our results reveal insights into the financial news, including the delayed market pricing of news, the long memory effect of news, and the limitations of financial sentiment analysis in fully extracting predictive power from news data.

Nov 12 (Tue) 11:00-12:30 - Riverfront Hall

Reference-free Hallucination Detection for Large Vision-Language Models

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, Fakhri Karray

Large vision-language models (LVLMs) have made significant progress in recent years. While LVLMs exhibit excellent ability in language understanding, question answering, and conversations of visual inputs, they are prone to producing hallucinations. While several methods are proposed to evaluate the hallucinations in LVLMs, most are reference-based and depend on external tools, which complicates their practical application. To assess the viability of alternative methods, it is critical to understand whether the reference-free approaches, which do not rely on any external tools, can efficiently detect hallucinations. Therefore, we initiate an exploratory study to demonstrate the effectiveness of different reference-free solutions in detecting hallucinations in LVLMs. In particular, we conduct an extensive study on three kinds of techniques: uncertainty-based, consistency-based, and supervised uncertainty quantification methods on four representative LVLMs across two different tasks. The empirical results show that the reference-free approaches are capable of effectively detecting non-factual responses in LVLMs, with the supervised uncertainty quantification method outperforming the others, achieving the best performance across different settings.

Session 03 - Nov 12 (Tue) 14:00-15:30

Demo

Nov 12 (Tue) 14:00-15:30 - Room: Riverfront Hall

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

ClaimLens: Automated, Explainable Fact-Checking on Voting Claims Using Frame-Semantics

Chengkai Li, Jacob Devasier, Phuong Anh Le, Rishabh Mediratta

We present ClaimLens, an automated fact-checking system focused on voting-related factual claims. Existing fact-checking solutions often lack transparency, making it difficult for users to trust and understand the reasoning behind the outcomes. In this work, we address the critical need for transparent and explainable automated fact-checking solutions. We propose a novel approach that leverages frame-semantic parsing to provide structured and interpretable fact verification. By focusing on voting-related claims, we can utilize publicly available voting records from official United States congressional sources and the established Vote semantic frame to extract relevant information from claims. Furthermore, we propose novel data augmentation techniques for frame-semantic parsing, a task known to lack robust annotated data, which leads to a +9.5% macro F1 score on frame element identification over our baseline.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

mbrs: A Library for Minimum Bayes Risk Decoding

Hidetaka Kamigaito, Hiroyuki Deguchi, Taro Watanabe, Yusuke Sakai

Minimum Bayes risk (MBR) decoding is a decision rule of text generation tasks that outperforms conventional maximum a posterior (MAP) decoding using beam search by selecting high-quality outputs based on a utility function rather than those with high-probability. Typically, it finds the most suitable hypothesis from the set of hypotheses under the sampled pseudo-references. mbrs is a library of MBR decoding, which can flexibly combine various metrics, alternative expectation estimations, and algorithmic variants. It is designed with a focus on speed measurement and calling count of code blocks, transparency, reproducibility, and extensibility, which are essential for researchers and developers. We published our mbrs as an MIT-licensed open-source project, and the code is available on GitHub^{1,2,3}. - ¹ GitHub: <https://github.com/naist-nlp/mbrs> - ² YouTube: <https://youtu.be/4qeHpg4PTn0> - ³ Demo: <https://mbrs-demo.streamlit.app>

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild

Claire Cardie, Jack Hessel, Wenjing Zhao, Xiang Ren, Yejin Choi, Yunlian Deng

The increasing availability of real-world conversation data offers exciting opportunities for researchers to study user-chatbot interactions.

However, the sheer volume of this data makes manually examining individual conversations impractical. To overcome this challenge, we introduce WildVis, an interactive tool that enables fast, versatile, and large-scale conversation analysis. WildVis provides search and visualization capabilities in the text and embedding spaces based on a list of criteria. To manage million-scale datasets, we implemented optimizations including search index construction, embedding precomputation and compression, and caching to ensure responsive user interactions within seconds. We demonstrate WildVis' utility through three case studies: facilitating chatbot misuse research, visualizing and comparing topic distributions across datasets, and characterizing user-specific conversation patterns. WildVis is open-source and designed to be extendable, supporting additional datasets and customized search and visualization functionalities.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

MarkLLM: An Open-Source Toolkit for LLM Watermarking

Aiwei Liu, Binglin Zhou, Irwin King, Leyi Pan, Lijie Wen, Philip S. Yu, Shuliang Liu, Xuandong Zhao, Xuming Hu, Yijian LU, Zhiwei He, Ziwei Gao

Watermarking for Large Language Models (LLMs), which embeds imperceptible yet algorithmically detectable signals in model outputs to identify LLM-generated text, has become crucial in mitigating the potential misuse of LLMs. However, the abundance of LLM watermarking algorithms, their intricate mechanisms, and the complex evaluation procedures and perspectives pose challenges for researchers and the community to easily understand, implement and evaluate the latest advancements. To address these issues, we introduce MarkLLM, an open-source toolkit for LLM watermarking. MarkLLM offers a unified and extensible framework for implementing LLM watermarking algorithms, while providing user-friendly interfaces to ensure ease of access. Furthermore, it enhances understanding by supporting automatic visualization of the underlying mechanisms of these algorithms. For evaluation, MarkLLM offers a comprehensive suite of 12 tools spanning three perspectives, along with two types of automated evaluation pipelines. Through MarkLLM, we aim to support researchers while improving the comprehension and involvement of the general public in LLM watermarking technology, fostering consensus and driving further advancements in research and application. Our code is available at <https://github.com/THU-BPM/MarkLLM>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Arxiv Copilot: A Self-Evolving and Efficient LLM System for Personalized Academic Assistance

Ge Liu, Guanyu Lin, Pengru Han, Tao Feng, Jiaxuan You

As scientific research proliferates, researchers face the daunting task of navigating and reading vast amounts of literature. Existing solutions, such as document QA, fail to provide personalized and up-to-date information efficiently. We present Arxiv Copilot, a self-evolving, efficient LLM system designed to assist researchers, based on thought-retrieval, user profile and high performance optimization. Specifically, Arxiv Copilot can offer personalized research services, maintaining a real-time updated database. Quantitative evaluation demonstrates that Arxiv Copilot saves 69.92% of time after efficient deployment. This paper details the design and implementation of Arxiv Copilot, highlighting its contributions to personalized academic support and its potential to streamline the research process. We have deployed Arxiv Copilot at: <https://huggingface.co/spaces/ulab-ai/ArxivCopilot>.

Discourse + Phonology + Syntax 1

Nov 12 (Tue) 14:00-15:30 - Room: Riverfront Hall

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

QUDSELECT: Selective Decoding for Questions Under Discussion Parsing

Ashima Suvarna, Xiao Liu, Tamay Parekh, Kai-Wei Chang, Nanyun Peng

Question Under Discussion (QUD) is a discourse framework that uses implicit questions to reveal discourse relationships between sentences. In QUD parsing, each sentence is viewed as an answer to a question triggered by an anchor sentence in prior context. The resulting QUD structure is required to conform to several theoretical criteria like answer compatibility (how well the question is answered), making QUD parsing a challenging task. Previous works construct QUD parsers in a pipelined manner (i.e., detect the trigger sentence in context and then generate the question). However, these parsers lack a holistic view of the task and can hardly satisfy all the criteria. In this work, we introduce QUDSELECT, a joint-training framework that selectively decodes the QUD dependency structures considering the QUD criteria criteria. Using instruction-tuning, we train models to simultaneously predict the anchor sentence and generate the associated question. To explicitly incorporate the criteria, we adopt a selective decoding strategy of sampling multiple QUD candidates during inference, followed by selecting the best one with criteria scorers. Our method outperforms the state-of-the-art baseline models by 9% in human evaluation and 4% in automatic evaluation, demonstrating the effectiveness of our framework. Code and data are in <https://github.com/asuvarna31/qudselect>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

An LLM Feature-based Framework for Dialogue Constructiveness Assessment

Lexin Zhou, Youmna Farag, Andreas Vlachos

Research on dialogue constructiveness assessment focuses on (i) analysing conversational factors that influence individuals to take specific actions, win debates, change their perspectives or broaden their open-mindedness and (ii) predicting constructiveness outcomes following dialogues for such use cases. These objectives can be achieved by training either interpretable feature-based models (which often involve costly human annotations) or neural models such as pre-trained language models (which have empirically shown higher task accuracy but lack interpretability). In this paper we propose an LLM feature-based framework for dialogue constructiveness assessment that combines the strengths of feature-based and neural approaches, while mitigating their downsides. The framework first defines a set of dataset-independent and interpretable linguistic features, which can be extracted by both prompting an LLM and simple heuristics. Such features are then used to train LLM feature-based models. We apply this framework to three datasets of dialogue constructiveness and find that our LLM feature-based models outperform or performs at least as well as standard feature-based models and neural models. We also find that the LLM feature-based model learns more robust prediction rules instead of relying on superficial shortcuts, which often trouble neural models.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Red Teaming Language Models for Processing Contradictory Dialogues

Xiaofei Wen, Bangzheng Li, Tenghao Huang, Muhaoo Chen

Most language models currently available are prone to self-contradiction during dialogues. To mitigate this issue, this study explores a novel contradictory dialogue processing task that aims to detect and modify contradictory statements in a conversation. This task is inspired by research on context faithfulness and dialogue comprehension, which have demonstrated that the detection and understanding of contradictions often necessitate detailed explanations. We develop a dataset comprising contradictory dialogues, in which one side of the conversation contradicts itself. Each dialogue is accompanied by an explanatory label that highlights the location and details of the contradiction. With this dataset, we present a Red Teaming framework for contradictory dialogue processing. The framework detects and attempts to explain the dialogue, then modifies the existing contradictory content using the explanation. Our experiments demonstrate that the framework improves

the ability to detect contradictory dialogues and provides valid explanations. Additionally, it showcases distinct capabilities for modifying such dialogues. Our study highlights the importance of the logical inconsistency problem in conversational AI.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Major Entity Identification: A Generalizable Alternative to Coreference Resolution

Kavish Manikanan, Shubham Toshniwal, Makarand Tapaswi, Vineet Gandhi

The limited generalization of coreference resolution (CR) models has been a major bottleneck in the tasks broad application. Prior work has identified annotation differences, especially for mention detection, as one of the main reasons for the generalization gap and proposed using additional annotated target domain data. Rather than relying on this additional annotation, we propose an alternative referential task, Major Entity Identification (MEI), where we: (a) assume the target entities to be specified in the input, and (b) limit the task to only the frequent entities. Through extensive experiments, we demonstrate that MEI models generalize well across domains on multiple datasets with supervised models and LLM-based few-shot prompting. Additionally, MEI fits the classification framework, which enables the use of robust and intuitive classification-based metrics. Finally, MEI is also of practical use as it allows a user to search for all mentions of a particular entity or a group of entities of interest.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

GDTB: Genre Diverse Data for English Shallow Discourse Parsing across Modalities, Text Types, and Domains

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, Amir Zeldes

Work on shallow discourse parsing in English has focused on the Wall Street Journal corpus, the only large-scale dataset for the language in the PDTB framework. However, the data is not openly available, is restricted to the news domain, and is by now 35 years old. In this paper, we present and evaluate a new open-access, multi-genre benchmark for PDTB-style shallow discourse parsing, based on the existing UD English GUM corpus, for which discourse relation annotations in other frameworks already exist. In a series of experiments on cross-domain relation classification, we show that while our dataset is compatible with PDTB, substantial out-of-domain degradation is observed, which can be alleviated by joint training on both datasets.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Argument Relation Classification through Discourse Markers and Adversarial Training

Michele Luca Contalbo, Francesco Guerra, Matteo Paganelli

Argument relation classification (ARC) identifies supportive, contrasting and neutral relations between argumentative units. The current approaches rely on transformer architectures which have proven to be more effective than traditional methods based on hand-crafted linguistic features. In this paper, we introduce DISARM, which advances the state of the art with a training procedure combining multi-task and adversarial learning strategies. By jointly solving the ARC and discourse marker detection tasks and aligning their embedding spaces into a unified latent space, DISARM outperforms the accuracy of existing approaches.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Will LLMs Replace the Encoder-Only Models in Temporal Relation Classification?

Gabriel Roccabruna, Massimo Rizzoli, Giuseppe Riccardi

The automatic detection of temporal relations among events has been mainly investigated with encoder-only models such as RoBERTa. Large Language Models (LLM) have recently shown promising performance in temporal reasoning tasks such as temporal question answering. Nevertheless, recent studies have tested the LLMs' performance in detecting temporal relations of closed-source models only, limiting the interpretability of those results. In this work, we investigate LLMs' performance and decision process in the Temporal Relation Classification task. First, we assess the performance of seven open and closed-sourced LLMs experimenting with in-context learning and lightweight fine-tuning approaches. Results show that LLMs with in-context learning significantly underperform smaller encoder-only models based on RoBERTa. Then, we delve into the possible reasons for this gap by applying explainable methods. The outcome suggests a limitation of LLMs in this task due to their autoregressive nature, which causes them to focus only on the last part of the sequence. Additionally, we evaluate the word embeddings of these two models to better understand their pre-training differences. The code and the fine-tuned models can be found respectively on GitHub.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Communicating with Speakers and Listeners of Different Pragmatic Levels

Kata Naszadi, Frans A Oliehoek, Christof Monz

This paper explores the impact of variable pragmatic competence on communicative success through simulating language learning and conversing between speakers and listeners with different levels of reasoning abilities. Through studying this interaction, we hypothesize that matching levels of reasoning between communication partners would create a more beneficial environment for communicative success and language learning. Our research findings indicate that learning from more explicit, literal language is advantageous, irrespective of the learner's level of pragmatic competence. Furthermore, we find that integrating pragmatic reasoning during language learning, not just during evaluation, significantly enhances overall communication performance. This paper provides key insights into the importance of aligning reasoning levels and incorporating pragmatic reasoning in optimizing communicative interactions.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning

Shengguang Wu, Shusheng Yang, Zhengjun Chen, Qi Su

This study addresses the challenges of assessing and enhancing social-pragmatic inference in large language models (LLMs). We first highlight the inadequacy of current accuracy-based multiple choice question answering (MCQA) formats in assessing social-pragmatic reasoning, and propose the direct evaluation of models' free-form responses as measure, which correlates better with human judgment. Furthermore, we explore methods to improve pragmatic abilities in LLMs, advocating for preference optimization (PO) over supervised finetuning (SFT), given the absence of a definitive "gold" answer in social contexts. Our results show that preferential tuning consistently outperforms SFT across pragmatic phenomena and offers a near-free launch in pragmatic abilities without compromising general capabilities. Lastly, we examine the internal structure of LLMs, revealing that the significant boost in pragmatic reasoning is tied to deeper layer representations, analogous to human high-level thinking. Our experiments span a variety of pragmatic and social reasoning datasets, as well as an image referential game requiring a multimodal theory of mind (ToM). With our refined paradigms for evaluating and enhancing pragmatic inference, this paper offers key insights into building more socially aware language models.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Any Other Thoughts, Hedgehog? Linking Deliberation Chains in Collaborative Dialogues

Abhijan Nath, Videep Venkatesha, Mariah Bradford, Aryaka Chelle, Austin Collin Youngren, Carlos Mabrey, Nathaniel Blanchard, Nikhil Krishnaswamy

Question-asking in collaborative dialogue has long been established as key to knowledge construction, both in internal and collaborative

problem solving. In this work, we examine probing questions in collaborative dialogues: questions that explicitly elicit responses from the speaker's interlocutors. Specifically, we focus on modeling the causal relations that lead directly from utterances earlier in the dialogue to the emergence of the probing question. We model these relations using a novel graph-based framework of "deliberation chains*", and realize the problem of constructing such chains as a coreference-style clustering problem. Our framework jointly models probing and causal utterances and the links between them, and we evaluate on two challenging collaborative task datasets: the Weights Task and DeliData. Our results demonstrate the effectiveness of our theoretically-grounded approach compared to both baselines and stronger coreference approaches, and establish a standard of performance in this novel task.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LLaMIPa: An Incremental Discourse Parser

Kate Thompson, Akshay Chaturvedi, Julie Hunter, Nicholas Asher

This paper provides the first discourse parsing experiments with a large language model (LLM) finetuned on corpora annotated in the style of SDRT (Segmented Discourse Representation Theory, Asher (1993), Asher and Lascaris (2003)). The result is a discourse parser, Llamipa (Llama Incremental Parser), that leverages discourse context, leading to substantial performance gains over approaches that use encoder-only models to provide local context-sensitive representations of discourse units. Furthermore, it is able to process discourse data incrementally, which is essential for the eventual use of discourse information in downstream tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Recent Trends in Linear Text Segmentation: A Survey

Iacopo Ghinassi, Lin Wang, Chris Newell, Matthew Purver

Linear Text Segmentation is the task of automatically tagging text documents with topic shifts, i.e. the places in the text where the topics change. A well-established area of research in Natural Language Processing, drawing from well-understood concepts in linguistic and computational linguistics research, the field has recently seen a lot of interest as a result of the surge of text, video, and audio available on the web, which in turn require ways of summarising and categorizing the mole of content for which linear text segmentation is a fundamental step. In this survey, we provide an extensive overview of current advances in linear text segmentation, describing the state of the art in terms of resources and approaches for the task. Finally, we highlight the limitations of available resources and of the task itself, while indicating ways forward based on the most recent literature and under-explored research directions.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Multiple Sources are Better Than One: Incorporating External Knowledge in Low-Resource Glossing

Changbing Yang, Garrett Nicolai, Miikka Silfverberg

In this paper, we address the data scarcity problem in automatic data-driven glossing for low-resource languages by coordinating multiple sources of linguistic expertise. We enhance models by incorporating both token-level and sentence-level translations, utilizing the extensive linguistic capabilities of modern LLMs, and incorporating available dictionary resources. Our enhancements lead to an average absolute improvement of 5%-points in word-level accuracy over the previous state of the art on a typologically diverse dataset spanning six low-resource languages. The improvements are particularly noticeable for the lowest-resourced language Gitksan, where we achieve a 10%-point improvement. Furthermore, in a simulated ultra-low resource setting for the same six languages, training on fewer than 100 glossed sentences, we establish an average 10%-point improvement in word-level accuracy over the previous state-of-the-art system.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Lexically Grounded Subword Segmentation

Jindrich Libovicky, Jindrich Helcl

We present three innovations in tokenization and subword segmentation. First, we propose to use unsupervised morphological analysis with Morfessor as pre-tokenization. Second, we present an algebraic method for obtaining subword embeddings grounded in a word embedding space. Based on that, we design a novel subword segmentation algorithm that uses the embeddings, ensuring that the procedure considers lexical meaning. Third, we introduce an efficient segmentation algorithm based on a subword bigram model that can be initialized with the lexically aware segmentation method to avoid using Morfessor and large embedding tables at inference time. We evaluate the proposed approaches using two intrinsic metrics and measure their performance on two downstream tasks: part-of-speech tagging and machine translation. Our experiments show significant improvements in the morphological plausibility of the segmentation when evaluated using segmentation precision on morpheme boundaries and improved Rényi efficiency in 8 languages. Although the proposed tokenization methods do not have a large impact on automatic translation quality, we observe consistent performance gains in the arguably more morphological task of part-of-speech tagging.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Distributional Properties of Subword Regularization

Marco Cognetti, Vilem Zouhar, Naoki Okazaki

Subword regularization, used widely in NLP, improves model performance by reducing the dependency on exact tokenizations, augmenting the training corpus, and exposing the model to more unique contexts during training. BPE and MaxMatch, two popular subword tokenization schemes, have stochastic dropout regularization variants. However, there has not been an analysis of the distributions formed by them. We show that these stochastic variants are heavily biased towards a small set of tokenizations per word. If the benefits of subword regularization are as mentioned, we hypothesize that biasedness artificially limits the effectiveness of these schemes. Thus, we propose an algorithm to uniformly sample tokenizations that we use as a drop-in replacement for the stochastic aspects of existing tokenizers, and find that it improves machine translation quality.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Subword Segmentation in LLMs: Looking at Inflection and Consistency

Marion Di Marco, Alexander Fraser

The role of subword segmentation in relation to capturing morphological patterns in LLMs is currently not well explored. Ideally, one would train models like GPT using various segmentations and evaluate how well word meanings are captured. Since this is not computationally feasible, we group words according to their segmentation properties and compare how well a model can solve a linguistic task for these groups. We study two criteria: (i) adherence to morpheme boundaries and (ii) the segmentation consistency of the different inflected forms of a lemma. We select word forms with high and low values for these criteria and carry out experiments on GPT-4os ability to capture verbal inflection for 10 languages. Our results indicate that in particular the criterion of segmentation consistency can help to predict the models ability to recognize and generate the lemma from an inflected form, providing evidence that subword segmentation is relevant.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Zero-Shot Cross-Lingual NER Using Phonemic Representations for Low-Resource Languages

Jimin Sohn, Haeji Jung, Alex Cheng, Jooeon Kang, Yilin Du, David R Mortensen

Existing zero-shot cross-lingual NER approaches require substantial prior knowledge of the target language, which is impractical for low-

resource languages. In this paper, we propose a novel approach to NER using phonemic representation based on the International Phonetic Alphabet (IPA) to bridge the gap between representations of different languages. Our experiments show that our method significantly outperforms baseline models in extremely low-resource languages, with the highest average F1 score (46.38%) and lowest standard deviation (12.67), particularly demonstrating its robustness with non-Latin scripts. Our codes are available at https://github.com/Gabriel819/zeroshot_ner.git

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Getting The Most Out of Your Training Data: Exploring Unsupervised Tasks for Morphological Inflection

Abhishek Purushothama, Adam Wiemerslage, Katharina von der Wense

Pre-trained transformers such as BERT have been shown to be effective in many natural language tasks. However, they are under-explored for character-level sequence to sequence tasks. In this work, we investigate pre-training transformers for the character-level task of morphological inflection in several languages. We compare various training setups and secondary tasks where unsupervised data taken directly from the target task is used. We show that training on secondary unsupervised tasks increases inflection performance even without any external data, suggesting that models learn from additional unsupervised tasks themselves—not just from additional data. We also find that this does not hold true for specific combinations of secondary task and training setup, which has interesting implications for denoising objectives in character-level tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

GE2PE: Persian End-to-End Grapheme-to-Phoneme Conversion

Elnaz Rahmati, Hossein Sameti

Text-to-Speech (TTS) systems have made significant strides, enabling the generation of speech from grapheme sequences. However, for low-resource languages, these models still struggle to produce natural and intelligible speech. Grapheme-to-Phoneme conversion (G2P) addresses this challenge by enhancing the input sequence with phonetic information. Despite these advancements, existing G2P systems face limitations when dealing with Persian texts due to the complexity of Persian transcription. In this study, we focus on enriching resources for the Persian language. To achieve this, we introduce two novel G2P training datasets: one manually labeled and the other machine-generated. These datasets comprise over five million sentences alongside their corresponding phoneme sequences. Additionally, we propose two evaluation datasets tailored for Persian sub-tasks, including Kasre-Ezafe detection, homograph disambiguation, and handling out-of-vocabulary (OOV) words. To tackle the unique challenges of the Persian language, we develop a new sentence-level End-to-End (E2E) model leveraging a two-step training approach, as outlined in our paper, to maximize the impact of manually labeled data. The results show that our model surpasses the state-of-the-art performance by 1.86% in word error rate, 4.03% in Kasre-Ezafe detection recall, and 3.42% in homograph disambiguation accuracy.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Deterministic Weighted L* Algorithm

Clemente Pasti, Talu Karagöz, Franz Nowak, Anej Svetec, Ryan Cotterell

Extracting finite state automata (FSAs) from black-box models offers a powerful approach to gaining interpretable insights into complex model behaviors. To support this pursuit, we present a weighted variant of Angluin's (1987) L^* algorithm for learning FSAs. We stay faithful to the original formulation, devising a way to exactly learn deterministic weighted FSAs whose weights support division. Furthermore, we formulate the learning process in a manner that highlights the connection with FSA minimization, showing how L^* directly learns a minimal automaton for the target language.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

On Eliciting Syntax from Language Models via Hashing

Yiran Wang, Masao Utiyama

Unsupervised parsing, also known as grammar induction, aims to infer syntactic structure from raw text. Recently, binary representation has exhibited remarkable information-preserving capabilities at both lexicon and syntax levels. In this paper, we explore the possibility of leveraging this capability to deduce parsing trees from raw text, relying solely on the implicitly induced grammars within models. To achieve this, we upgrade the bit-level CKY from zero-order to first-order to encode the lexicon and syntax in a unified binary representation space, switch training from supervised to unsupervised under the contrastive hashing framework, and introduce a novel loss function to impose stronger yet balanced alignment signals. Our model shows competitive performance on various datasets; therefore, we claim that our method is effective and efficient enough to acquire high-quality parsing trees from pre-trained language models at a low cost.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Strengthening Structural Inductive Biases by Pre-training to Perform Syntactic Transformations

Matthias Lindemann, Alexander Koller, Ivan Titov

Models need appropriate inductive biases to effectively learn from small amounts of data and generalize systematically outside of the training distribution. While Transformers are highly versatile and powerful, they can still benefit from enhanced structural inductive biases for seq2seq tasks, especially those involving syntactic transformations, such as converting active to passive voice or semantic parsing. In this paper, we propose to strengthen the structural inductive bias of a Transformer by intermediate pre-training to perform synthetically generated syntactic transformations of dependency trees given a description of the transformation. Our experiments confirm that this helps with few-shot learning of syntactic tasks such as chunking, and also improves structural generalization for semantic parsing. Our analysis shows that the intermediate pre-training leads to attention heads that keep track of which syntactic transformation needs to be applied to which token, and that the model can leverage these attention heads on downstream tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Dependency Graph Parsing as Sequence Labeling

Ana Ezquierro, David Vilares, Carlos Gómez-Rodríguez

Various linearizations have been proposed to cast syntactic dependency parsing as sequence labeling. However, these approaches do not support more complex graph-based representations, such as semantic dependencies or enhanced universal dependencies, as they cannot handle reentrancy or cycles. By extending them, we define a range of unbounded and bounded linearizations that can be used to cast graph parsing as a tagging task, enlarging the toolbox of problems that can be solved under this paradigm. Experimental results on semantic dependency and enhanced UD parsing show that with a good choice of encoding, sequence-labeling semantic dependency parsers combine high efficiency with accuracies close to the state of the art, in spite of their simplicity.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Automatic sentence segmentation of clinical record narratives in real-world data

Dongfang Xu, Davy Weissbacher, Karen O'Connor, Siddharth Rawal, Graciela Gonzalez Hernandez

Sentence segmentation is a linguistic task and is widely used as a pre-processing step in many NLP applications. The need for sentence segmentation is particularly pronounced in clinical notes, where ungrammatical and fragmented texts are common. We propose a straightforward and effective sequence labeling classifier to predict sentence spans using a dynamic sliding window based on the prediction of each input

sequence. This sliding window algorithm allows our approach to segment long text sequences on the fly. To evaluate our approach, we annotated 90 clinical notes from the MIMIC-III dataset. Additionally, we tested our approach on five other datasets to assess its generalizability and compared its performance against state-of-the-art systems on these datasets. Our approach outperformed all the systems, achieving an F1 score that is 15% higher than the next best-performing system on the clinical dataset.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LLM-supertagger: Categorial Grammar Supertagging via Large Language Models

Jinman Zhao, Gerald Penn

Supertagging is an essential task in Categorical grammar parsing and is crucial for dissecting sentence structures. Our research explores the capacity of Large Language Models (LLMs) in supertagging for both Combinatory Categorical Grammar (CCG) and Lambek Categorical Grammar (LCG). We also present a simple method that significantly boosts LLMs, enabling them to outperform LSTM and encoder-based models and achieve state-of-the-art performance. This advancement highlights LLMs' potential in classification tasks, showcasing their adaptability beyond generative capabilities. Our findings demonstrate the evolving utility of LLMs in natural language processing, particularly in complex tasks like supertagging.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Predicting generalization performance with correctness discriminators

Yuekun Yao, Alexander Koller

The ability to predict an NLP model's accuracy on unseen, potentially out-of-distribution data is a prerequisite for trustworthiness. We present a novel model that establishes upper and lower bounds on the accuracy, without requiring gold labels for the unseen data. We achieve this by training a "discriminator" which predicts whether the output of a given sequence-to-sequence model is correct or not. We show across a variety of tagging, parsing, and semantic parsing tasks that the gold accuracy is reliably between the predicted upper and lower bounds, and that these bounds are remarkably close together.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

When the Misidentified Adverbial Phrase Functions as a Complement

Yige Chen, Kyuwon Kim, KyungTae Lim, Jungyeul Park, Chulwoo Park

This study investigates the predicate-argument structure in Korean language processing. Despite the importance of distinguishing mandatory arguments and optional modifiers in sentences, research in this area has been limited. We introduce a dataset with token-level annotations which labels mandatory and optional elements as complements and adjuncts, respectively. Particularly, we reclassify certain Korean phrases, previously misidentified as adverbial phrases, as complements, addressing misuses of the term adjunct in existing Korean treebanks. Utilizing a Korean dependency treebank, we develop an automatic labeling technique for complements and adjuncts. Experiments using the proposed dataset yield satisfying results, demonstrating that the dataset is trainable and reliable.

Ethics, Bias, and Fairness 1

Nov 12 (Tue) 14:00-15:30 - Room: Riverfront Hall

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

On the Influence of Gender and Race in Romantic Relationship Prediction from Large Language Models

Abhilasha Sancheti, Haozhe An, Rachel Rudinger

We study the presence of heteronormative biases and prejudice against interracial romantic relationships in large language models by performing controlled name-replacement experiments for the task of relationship prediction. We show that models are less likely to predict romantic relationships for (a) same-gender character pairs than different-gender pairs; and (b) intra/inter-racial character pairs involving Asian names as compared to Black, Hispanic, or White names. We examine the contextualized embeddings of first names and find that gender for Asian names is less discernible than non-Asian names. We discuss the social implications of our findings, underlining the need to prioritize the development of inclusive and equitable technology.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

SHIELD: Evaluation and Defense Strategies for Copyright Compliance in LLM Text Generation

Xiaozhe Liu, Ting Sun, Tianyang Xu, Feifei Wu, Cunxiang Wang, Xiaolian Wang, Jing Gao

Large Language Models (LLMs) have transformed machine learning but raised significant legal concerns due to their potential to produce text that infringes on copyrights, resulting in several high-profile lawsuits. The legal landscape is struggling to keep pace with these rapid advancements, with ongoing debates about whether generated text might plagiarize copyrighted materials. Current LLMs may infringe on copyrights or overly restrict non-copyrighted texts, leading to these challenges: (i) the need for a comprehensive evaluation benchmark to assess copyright compliance from multiple aspects; (ii) evaluating robustness against safeguard bypassing attacks; and (iii) developing effective defenses targeted against the generation of copyrighted text. To tackle these challenges, we introduce a curated dataset to evaluate methods, test attack strategies, and propose a lightweight, real-time defense mechanism to prevent the generation of copyrighted text, ensuring the safe and lawful use of LLMs. Our experiments demonstrate that current LLMs frequently output copyrighted text, and that jailbreak attacks can significantly increase the volume of copyrighted output. Our proposed defense mechanism substantially reduces the volume of copyrighted text generated by LLMs by effectively refusing malicious requests.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Alignment-Enhanced Decoding: Defending via Token-Level Adaptive Refining of Probability Distributions

Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, Sen Su

Large language models are susceptible to jailbreak attacks, which can result in the generation of harmful content. While prior defenses mitigate these risks by perturbing or inspecting inputs, they ignore competing objectives, the underlying cause of alignment failures. In this paper, we propose Alignment-Enhanced Decoding (AED), a novel defense that employs adaptive decoding to address the root causes of jailbreak issues. We first define the Competitive Index to quantify alignment failures and utilize feedback from self-evaluation to compute post-alignment logits. Then, AED adaptively combines Competitive Index and post-alignment logits with the original logits to obtain harmless and helpful distributions. Consequently, our method enhances safety alignment while maintaining helpfulness. We conduct experiments across five models and four common jailbreaks, with the results validating the effectiveness of our approach.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Understanding "Democratization" in NLP and ML Research

Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, Zeerak Talat

Recent improvements in natural language processing (NLP) and machine learning (ML) and increased mainstream adoption have led to researchers frequently discussing the "democratization" of artificial intelligence. In this paper, we seek to clarify how democratization is understood in NLP and ML publications, through large-scale mixed-methods analyses of papers using the keyword "democra*" published in NLP and adjacent venues. We find that democratization is most frequently used to convey (ease of) access to or use of technologies, without meaningfully engaging with theories of democratization, while research using other invocations of "democra*" tends to be grounded in theories of deliberation and debate. Based on our findings, we call for researchers to enrich their use of the term democratization with appropriate theory, towards democratic technologies beyond superficial access.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Hidden Persuaders: How LLM Political Bias Could Sway Our Elections

Jujin Potter, Shiyang Lai, Junsol Kim, James Evans, Dawn Song

Do LLMs have political leanings and are LLMs able to shift our political views? This paper explores these questions in the context of the 2024 U.S. presidential election. Through a voting simulation, we demonstrate 18 open-weight and closed-source LLMs' political preference for Biden over Trump. We show how Biden-leaning becomes more pronounced in instruction-tuned and reinforced models compared to their base versions by analyzing their responses to political questions related to the two nominees. We further explore the potential impact of LLMs on voter choice by recruiting 935 U.S. registered voters. Participants interacted with LLMs (Claude-3, Llama-3, and GPT-4) over five exchanges. Intriguingly, although LLMs were not asked to persuade users to support Biden, about 20% of Trump supporters reduced their support for Trump after LLM interaction. This result is noteworthy given that many studies on the persuasiveness of political campaigns have shown minimal effects in presidential elections. Many users also expressed a desire for further interaction with LLMs on political subjects. Further research on how LLMs affect users' political views is required, as their use becomes more widespread.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

How Far Can We Extract Diverse Perspectives from Large Language Models?

Shirley Anugrah Hayati, Minhwaa Lee, Dheeraj Rajagopal, Dongyeop Kang

Collecting diverse human opinions is costly and challenging. This leads to a recent trend in exploiting large language models (LLMs) for generating diverse data for potential scalable and efficient solutions. However, the extent to which LLMs can generate diverse perspectives on subjective topics is still unclear. In this study, we explore LLMs' capacity of generating diverse perspectives and rationales on subjective topics such as social norms and argumentative texts. We introduce the problem of extracting maximum diversity from LLMs. Motivated by how humans form opinions based on values, we propose a criteria-based prompting technique to ground diverse opinions. To see how far we can extract diverse perspectives from LLMs, or called diversity coverage, we employ a step-by-step recall prompting to generate more outputs from the model iteratively. Our methods, applied to various tasks, show that LLMs can indeed produce diverse opinions according to the degree of task subjectivity. We also find that LLMs performance of extracting maximum diversity is on par with human.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Thinking Outside of the Differential Privacy Box: A Case Study in Text Privatization with Language Model Prompting

Stephan Meisenbacher, Florian Matthes

The field of privacy-preserving Natural Language Processing has risen in popularity, particularly at a time when concerns about privacy grow with the proliferation of large language models. One solution consistently appearing in recent literature has been the integration of Differential Privacy (DP) into NLP techniques. In this paper, we take these approaches into critical view, discussing the restrictions that DP integration imposes, as well as bring to light the challenges that such restrictions entail. To accomplish this, we focus on **DP-Prompt**, a recent method for text privatization leveraging language models to rewrite texts. In particular, we explore this rewriting task in multiple scenarios, both with DP and without DP. To drive the discussion on the merits of DP in NLP, we conduct empirical identity and privacy experiments. Our results demonstrate the need for more discussion on the usability of DP in NLP and its benefits over non-DP approaches.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

ChatGPT Doesn't Trust LA Chargers Fans: Guardrail Sensitivity in Context

Victoria R Li, Yida Chen, Naomi Saphra

While the biases of language models in production are extensively documented, the biases of their guardrails have been neglected. This paper studies how contextual information about the user influences the likelihood of an LLM to refuse to execute a request. By generating user biographies that offer ideological and demographic information, we find a number of biases in guardrail sensitivity on GPT-3.5. Younger, female, and Asian-American personas are more likely to trigger a refusal guardrail when requesting censored or illegal information. Guardrails are also sycophantic, refusing to comply with requests for a political position the user is likely to disagree with. We find that certain identity groups and seemingly innocuous information, e.g., sports fandom, can elicit changes in guardrail sensitivity similar to direct statements of political ideology. For each demographic category and even for American football team fandom, we find that ChatGPT appears to infer a likely political ideology and modify guardrail behavior accordingly.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

An Audit on the Perspectives and Challenges of Hallucinations in NLP

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srivastava, Koustava Goswami, Sarah Rajtmajer, Shomir Wilson

We audit how hallucination in large language models (LLMs) is characterized in peer-reviewed literature, using a critical examination of 103 publications across NLP research. Through the examination of the literature, we identify a lack of agreement with the term 'hallucination' in the field of NLP. Additionally, to compliment our audit, we conduct a survey with 171 practitioners from the field of NLP and AI to capture varying perspectives on hallucination. Our analysis calls for the necessity of explicit definitions and frameworks outlining hallucination within NLP, highlighting potential challenges, and our survey inputs provide a thematic understanding of the influence and ramifications of hallucination in society.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning

Jingyu Hu, Weiru Liu, Mengnan Du

Recent studies highlight the effectiveness of using in-context learning (ICL) to steer large language models (LLMs) in processing tabular data, a challenging task given the structured nature of such data. Despite advancements in performance, the fairness implications of these methods are less understood. This study investigates how varying demonstrations within ICL prompts influence the fairness outcomes of LLMs. Our findings reveal that deliberately including minority group samples in prompts significantly boosts fairness without sacrificing predictive accuracy. Further experiments demonstrate that the proportion of minority to majority samples in demonstrations affects the trade-off between fairness and prediction accuracy. Based on these insights, we introduce a mitigation technique that employs clustering and evolutionary strategies to curate a diverse and representative sample set from the training data. This approach aims to enhance both predictive performance and fairness in ICL applications. Experimental results validate that our proposed method dramatically improves fairness across various metrics, showing its efficacy in real-world scenarios.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Humans or LLMs as the Judge? A Study on Judgement Bias

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, Benyou Wang

Adopting human and large language models (LLMs) as judges (*a.k.a* human- and LLM-as-a-judge) for evaluating the performance of LLMs has recently gained attention. Nonetheless, this approach concurrently introduces potential biases from human and LLMs, questioning the reliability of the evaluation results. In this paper, we propose a novel framework that is free from referencing groundtruth annotations for investigating ^{**}Misinformation Oversight Bias^{**}, ^{**}Gender Bias^{**}, ^{**}Authority Bias^{**} and ^{**}Beauty Bias^{**} on LLM and human judges. We curate a dataset referring to the revised Bloom's Taxonomy and conduct thousands of evaluations. Results show that human and LLM judges are vulnerable to perturbations to various degrees, and that even the cutting-edge judges possess considerable biases. We further exploit these biases to conduct attacks on LLM judges. We hope that our work can notify the community of the bias and vulnerability of human- and LLM-as-a-judge, as well as the urgency of developing robust evaluation systems.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LoRA-Guard: Parameter-Efficient Guardrail Adaptation for Content Moderation of Large Language Models

Hayder Elesedy, Pedro M Esperanza, Silvia Vlad Oprea, Mete Ozay

Guardrails have emerged as an alternative to safety alignment for content moderation of large language models (LLMs). Existing model-based guardrails have not been designed for resource-constrained computational portable devices, such as mobile phones, more and more of which are running LLM-based applications locally. We introduce LoRA-Guard, a parameter-efficient guardrail adaptation method that relies on knowledge sharing between LLMs and guardrail models. LoRA-Guard extracts language features from the LLMs and adapts them for the content moderation task using low-rank adapters, while a dual-path design prevents any performance degradation on the generative task. We show that LoRA-Guard outperforms existing approaches with 100-1000x lower parameter overhead while maintaining accuracy, enabling on-device content moderation.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Global is Good, Local is Bad??: Understanding Brand Bias in LLMs

Mahammed Kamruzzaman, Hieu Minh Nguyen, Gene Louis Kim

Many recent studies have investigated social biases in LLMs but brand bias has received little attention. This research examines the biases exhibited by LLMs towards different brands, a significant concern given the widespread use of LLMs in affected use cases such as product recommendation and market analysis. Biased models may perpetuate societal inequalities, unfairly favoring established global brands while marginalizing local ones. Using a curated dataset across four brand categories, we probe the behavior of LLMs in this space. We find a consistent pattern of bias in this space—both in terms of disproportionately associating global brands with positive attributes and disproportionately recommending luxury gifts for individuals in high-income countries. We also find LLMs are subject to country-of-origin effects which may boost local brand preference in LLM outputs in specific contexts.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, Yang Zhang

Large vision-language models (LVLMs) have been rapidly developed and widely used in various fields, but the (potential) stereotypical bias in the model is largely unexplored. In this study, we present a pioneering measurement framework, ModSCAN, to SCAN the stereotypical bias within LVLMs from both vision and language modalities. ModSCAN examines stereotypical biases with respect to two typical stereotypical attributes (gender and race) across three kinds of scenarios: occupations, descriptors, and persona traits. Our findings suggest that 1) the currently popular LVLMs show significant stereotype biases, with CogVLM emerging as the most biased model; 2) these stereotypical biases may stem from the inherent biases in the training dataset and pre-trained models; 3) the utilization of specific prompt prefixes (from both vision and language modalities) performs well in reducing stereotypical biases. We believe our work can serve as the foundation for understanding and addressing stereotypical bias in LVLMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models

Yi Zheng, Weiyu Sun, Tran Ngoc Huynh, Dawn Song, Bo Li, Ruoxi Jia

Safety backdoor attacks in large language models (LLMs) enable harmful behaviors to be stealthily triggered while evading detection during normal interactions. The high dimensionality of the trigger search space and the diverse range of potential malicious behaviors in LLMs make this a critical open problem. This paper presents BEEAR, a novel mitigation method based on a key insight: backdoor triggers induce a uniform drift in the model's embedding space, irrespective of the trigger's form or targeted behavior. Leveraging this observation, we introduce a bi-level optimization approach. The inner level identifies universal perturbations to the decoder's embeddings that steer the model towards defender-defined unwanted behaviors; the outer level fine-tunes the model to reinforce safe behaviors against these perturbations. Our experiments demonstrate the effectiveness of this approach, reducing the success rate of safety backdoor attacks from over 95% to <1% for general harmful behaviors and from 47% to 0% for Sleeper Agents, without compromising the model's helpfulness. Notably, our method relies only on defender-defined sets of safe and unwanted behaviors without any assumptions about the trigger location or attack mechanism. This work represents the first practical framework to counter safety backdoors in LLMs and provides a foundation for future advancements in AI safety and security.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, Dan Klein

We present a large-scale study of linguistic bias exhibited by ChatGPT covering ten dialects of English (Standard American English, Standard British English, and eight widely spoken non-“standard” varieties from around the world). We prompted GPT-3.5 Turbo and GPT-4 with text by native speakers of each variety and analyzed the responses via detailed linguistic feature annotation and native speaker evaluation. We find that the models default to “standard” varieties of English; based on evaluation by native speakers, we also find that model responses to non-“standard” varieties consistently exhibit a range of issues: stereotyping (19% worse than for “standard” varieties), demeaning content (25% worse), lack of comprehension (9% worse), and condescending responses (15% worse). Moreover, if these models are asked to imitate the writing style of prompts in non-“standard” varieties, they produce text that exhibits lower comprehension of the input and is especially prone to stereotyping. GPT-4 improves on GPT-3.5 in terms of comprehension, warmth, and friendliness, but also exhibits a marked increase in stereotyping (+18%). The results indicate that GPT-3.5 Turbo and GPT-4 can perpetuate linguistic discrimination toward speakers of non-“standard” varieties.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

BiasAlert: A Plug-and-play Tool for Social Bias Detection in LLMs

Zhitong Fan, Ruijie Chen, Ruiling Xu, Zuozhu Liu

Evaluating the bias of LLMs becomes more crucial with their rapid development. However, existing evaluation approaches rely on fixed-form outputs and cannot adapt to the flexible open-text generation scenarios of LLMs (e.g., sentence completion and question answering). To address this, we introduce BiasAlert, a plug-and-play tool designed to detect social bias in open-text generations of LLMs. BiasAlert integrates external human knowledge with its inherent reasoning capabilities to detect bias reliably. Extensive experiments demonstrate that BiasAlert significantly outperforms existing state-of-the-art methods like GPT-4-as-Judge in detecting bias. Furthermore, through application studies, we showcase the utility of BiasAlert in reliable LLM fairness evaluation and bias mitigation across various scenarios. Model and code will be publicly released.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Shortcuts Arising from Contrast: Towards Effective and Lightweight Clean-Label Attacks in Prompt-Based Learning

Xiaopeng Xie, Ming YAN, Xiwen Zhou, Chenlong Zhao, Suli Wang, Yong Zhang, Joey Tianyi Zhou

Prompt-based learning paradigm has been shown to be vulnerable to backdoor attacks. Current clean-label attack, employing a specific prompt as trigger, can achieve success without the need for external triggers and ensuring correct labeling of poisoned samples, which are more stealthy compared to the poisoned-label attack, but on the other hand, facing significant issues with false activations and pose greater challenges, necessitating a higher rate of poisoning. Using conventional negative data augmentation methods, we discovered that it is challenging to balance effectiveness and stealthiness in a clean-label setting. In addressing this issue, we are inspired by the notion that a backdoor acts as a shortcut, and posit that this shortcut stems from the contrast between the trigger and the data utilized for poisoning. In this study, we propose a method named Contrastive Shortcut Injection (CSI), by leveraging activation values, integrates trigger design and data selection strategies to craft stronger shortcut features. With extensive experiments on full-shot and few-shot text classification tasks, we empirically validate CSIs high effectiveness and high stealthiness at low poisoning rates.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Images Speak Louder than Words: Understanding and Mitigating Bias in Vision-Language Model from a Causal Mediation Perspective

Zhaotian Weng, Zijun Gao, Jerome Andrews, Jieyu Zhao

Vision-language models (VLMs) pre-trained on extensive datasets can inadvertently learn biases by correlating gender information with specific objects or scenarios. Current methods, which focus on modifying inputs and monitoring changes in the models output probability scores, often struggle to comprehensively understand bias from the perspective of model components. We propose a framework that incorporates causal mediation analysis to measure and map the pathways of bias generation and propagation within VLMs. Our framework is applicable to a wide range of vision-language and multimodal tasks. In this work, we apply it to the object detection task and implement it on the GLIP model. This approach allows us to identify the direct effects of interventions on model bias and the indirect effects of interventions on bias mediated through different model components. Our results show that image features are the primary contributors to bias, with significantly higher impacts than text features, specifically accounting for 32.57% and 12.63% of the bias in the MSCOCO and PASCAL-SENTENCE datasets, respectively. Notably, the image encoder's contribution surpasses that of the text encoder and the deep fusion encoder. Further experimentation confirms that contributions from both language and vision modalities are aligned and non-conflicting. Specifically, focusing on blurring gender representations within the image encoder which contributes most to the model bias, reduces bias efficiently by 22.03% and 9.04% in the MSCOCO and PASCAL-SENTENCE datasets, respectively, with minimal performance loss or increased computational demands.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi XIE, Rui Pan, Qing LIAN, Hanze Dong, Jipeng Zhang, Tong Zhang

The deployment of multimodal large language models (MLLMs) has brought forth a unique vulnerability: susceptibility to malicious attacks through visual inputs. This paper investigates the novel challenge of defending MLLMs against such attacks. Compared to large language models (LLMs), MLLMs include an additional image modality. We discover that images act as a "foreign language" that is not considered during safety alignment, making MLLMs more prone to producing harmful responses. Unfortunately, unlike the discrete tokens considered in text-based LLMs, the continuous nature of image signals presents significant alignment challenges, which poses difficulty to thoroughly cover all possible scenarios. This vulnerability is exacerbated by the fact that most state-of-the-art MLLMs are fine-tuned on limited image-text pairs that are much fewer than the extensive text-based pretraining corpus, which makes the MLLMs more prone to catastrophic forgetting of their original abilities during safety fine-tuning. To tackle these challenges, we introduce MLLM-Protector, a plug-and-play strategy that solves two subtasks: 1) identifying harmful responses via a lightweight harm detector, and 2) transforming harmful responses into harmless ones via a detoxifier. This approach effectively mitigates the risks posed by malicious visual inputs without compromising the original performance of MLLMs. Our results demonstrate that MLLM-Protector offers a robust solution to a previously unaddressed aspect of MLLM security.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

RAFT: Realistic Attacks to Fool Text Detectors

James Liyuan Wang, Ran Li, Junfeng Yang, Chengzhi Mao

Large language models (LLMs) have exhibited remarkable fluency across various tasks. However, their unethical applications, such as disseminating disinformation, have become a growing concern. Although recent works have proposed a number of LLM detection methods, their robustness and reliability remain unclear. In this paper, we present RAFT: a grammar error-free black-box attack against existing LLM detectors. In contrast to previous attacks for language models, our method exploits the transferability of LLM embeddings at the word-level while preserving the original text quality. We leverage an auxiliary embedding to greedily select candidate words to perturb against the target detector. Experiments reveal that our attack effectively compromises all detectors in the study across various domains by up to 99%, and are transferable across source models. Manual human evaluation studies show our attacks are realistic and indistinguishable from original human-written text. We also show that examples generated by RAFT can be used to train adversarially robust detectors. Our work shows that current LLM detectors are not adversarially robust, underscoring the urgent need for more resilient detection mechanisms.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

From Descriptive Richness to Bias: Unveiling the Dark Side of Generative Image Caption Enrichment

Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, Yuta Nakashima

Large language models (LLMs) have enhanced the capacity of vision-language models to caption visual text. This generative approach to image caption enrichment further makes textual captions more descriptive, improving alignment with the visual context. However, while many studies focus on the benefits of generative caption enrichment (GCE), are there any negative side effects? We compare standard-format captions and recent GCE processes from the perspectives of gender bias and hallucination, showing that enriched captions suffer from increased gender bias and hallucination. Furthermore, models trained on these enriched captions amplify gender bias by an average of 30.9% and increase hallucination by 59.5%. This study serves as a caution against the trend of making captions more descriptive.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerero-Arenas, Luisa Bentivogli

Gender bias in machine translation (MT) is recognized as an issue that can harm people and society. And yet, advancements in the field rarely involve people, the final MT users, or inform how they might be impacted by biased technologies. Current evaluations are often restricted to automatic methods, which offer an opaque estimate of what the downstream impact of gender disparities might be. We conduct an extensive human-centered study to examine if and to what extent bias in MT brings harms with tangible costs, such as quality of service gaps across women and men. To this aim, we collect behavioral data from 90 participants, who post-edited MT outputs to ensure correct gender translation. Across multiple datasets, languages, and types of users, our study shows that feminine post-editing demands significantly more technical and temporal effort, also corresponding to higher financial costs. Existing bias measurements, however, fail to reflect the found disparities. Our findings advocate for human-centered approaches that can inform the societal impact of bias.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Jailbreaking LLMs with Arabic Transliteration and Arabizi

Mansour Al Ghannim, saleh almohaimeed, Mengxin Zheng, Yan Solihin, Qian Lou

This study identifies the potential vulnerabilities of Large Language Models (LLMs) to 'jailbreak' attacks, specifically focusing on the Arabic language and its various forms. While most research has concentrated on English-based prompt manipulation, our investigation broadens the scope to investigate the Arabic language. We initially tested the AdvBench benchmark in Standardized Arabic, finding that even with prompt manipulation techniques like prefix injection, it was insufficient to provoke LLMs into generating unsafe content. However, when using Arabic transliteration and chatspeak (or arabizi), we found that unsafe content could be produced on platforms like OpenAI GPT-4 and Anthropic Claude 3 Sonnet. Our findings suggest that using Arabic and its various forms could expose information that might remain hidden, potentially increasing the risk of jailbreak attacks. We hypothesize that this exposure could be due to the model's learned connection to specific words, highlighting the need for more comprehensive safety training across all language forms.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Who is better at math, Jenny or Jingzhen? Uncovering Stereotypes in Large Language Models

Zara Siddique, Liam Turner, Luis Espinosa-Anke

Large language models (LLMs) have been shown to propagate and amplify harmful stereotypes, particularly those that disproportionately affect marginalized communities. To understand the effect of these stereotypes more comprehensively, we introduce GlobalBias, a dataset of 876k sentences incorporating 40 distinct gender-by-ethnicity groups alongside descriptors typically used in bias literature, which enables us to study a broad set of stereotypes from around the world. We use GlobalBias to directly probe a suite of LMs via perplexity, which we use as a proxy to determine how certain stereotypes are represented in the model's internal representations. Following this, we generate character profiles based on given names and evaluate the prevalence of stereotypes in model outputs. We find that the demographic groups associated with various stereotypes remain consistent across model likelihoods and model outputs. Furthermore, larger models consistently display higher levels of stereotypical outputs, even when explicitly instructed not to.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

The Generation Gap: Exploring Age Bias Underlying in the Value Systems of Large Language Models

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, Rada Mihalcea

We explore the alignment of values in Large Language Models (LLMs) with specific age groups, leveraging data from the World Value Survey across thirteen categories. Through a diverse set of prompts tailored to ensure response robustness, we find a general inclination of LLM values towards younger demographics, especially when compared to the US population. Although a general inclination can be observed, we also found that this inclination toward younger groups can be different across different value categories. Additionally, we explore the impact of incorporating age identity information in prompts and observe challenges in mitigating value discrepancies with different age cohorts. Our findings highlight the age bias in LLMs and provide insights for future work. Materials for our analysis will be available via <https://github.com/anonymous>

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Evaluating Short-Term Temporal Fluctuations of Social Biases in Social Media Data and Masked Language Models

Yi Zhou, Danushka Bollegala, Jose Camacho-Collados

Social biases such as gender or racial biases have been reported in language models (LMs), including Masked Language Models (MLMs). Given that MLMs are continuously trained with increasing amounts of additional data collected over time, an important yet unanswered question is how the social biases encoded with MLMs vary over time. In particular, the number of social media users continues to grow at an exponential rate, and it is a valid concern for the MLMs trained specifically on social media data whether their social biases (if any) would also amplify over time. To empirically analyse this problem, we use a series of MLMs pretrained on chronologically ordered temporal snapshots of corpora. Our analysis reveals that, although social biases are present in all MLMs, most types of social bias remain relatively stable over time (with a few exceptions). To further understand the mechanisms that influence social biases in MLMs, we analyse the temporal corpora used to train the MLMs. Our findings show that some demographic groups, such as male, obtain higher preference over the other, such as female on the training corpora constantly.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Gender Bias in Decision-Making with Large Language Models

Sharon Levy, William Adler, Tahlin Sanchez Karver, Mark Dredze, Michelle R Kaufman

Large language models (LLMs) acquire beliefs about gender from training data and can therefore generate text with stereotypical gender attitudes. Prior studies have demonstrated model generations favor one gender or exhibit stereotypes about gender, but have not investigated the complex dynamics that can influence model reasoning and decision-making involving gender. We study gender equity within LLMs through a decision-making lens with a new dataset, DeMET Prompts, containing scenarios related to intimate, romantic relationships. We explore nine relationship configurations through name pairs across three name lists (men, women, neutral). We investigate equity in the context of gender roles through numerous lenses: typical and gender-neutral names, with and without model safety enhancements, same and mixed-gender relationships, and egalitarian versus traditional scenarios across various topics. While all models exhibit the same biases (women favored, then those with gender-neutral names, and lastly men), safety guardrails reduce bias. In addition, models tend to circumvent traditional male dominance stereotypes and side with "traditionally female" individuals more often, suggesting relationships are viewed as a female domain by the models.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions

Angana Borah, Rada Mihalcea

As Large Language Models (LLMs) continue to evolve, they are increasingly being employed in numerous studies to simulate societies and execute diverse social tasks. However, LLMs are susceptible to societal biases due to their exposure to human-generated data. Given that LLMs are being used to gain insights into various societal aspects, it is essential to mitigate these biases. To that end, our study investigates

the presence of implicit gender biases in multi-agent LLM interactions and proposes two strategies to mitigate these biases. We begin by creating a dataset of scenarios where implicit gender biases might arise, and subsequently develop a metric to assess the presence of biases. Our empirical analysis reveals that LLMs generate outputs characterized by strong implicit bias associations ($\geq \approx 50\%$ of the time). Furthermore, these biases tend to escalate following multi-agent interactions. To mitigate them, we propose two strategies: self-reflection with in-context examples (ICE); and supervised fine-tuning. Our research demonstrates that both methods effectively mitigate implicit biases, with the ensemble of fine-tuning and self-reflection proving to be the most successful.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Beyond Perplexity: Multi-dimensional Safety Evaluation of LLM Compression

Zhiyao Xu, Ashim Gupta, Tao Li, Oliver Bentham, Vivek Srikanth

Increasingly, model compression techniques enable large language models (LLMs) to be deployed in real-world applications. As a result of this momentum towards local deployment, compressed LLMs will interact with a large population. Prior work on compression typically prioritizes preserving perplexity, which is directly analogous to training loss. The impact of compression method on other critical aspects of model behavior—particularly safety—requires systematic assessment. To this end, we investigate the impact of model compression along four dimensions: (1) degeneration harm, i.e., bias and toxicity in generation; (2) representational harm, i.e., biases in discriminative tasks; (3) dialect bias; and (4) language modeling and downstream task performance. We examine a wide spectrum of LLM compression techniques, including unstructured pruning, semi-structured pruning, and quantization. Our analysis reveals that compression can lead to unexpected consequences. Although compression may unintentionally alleviate LLMs' degeneration harm, it can still exacerbate representational harm. Furthermore, increasing compression produces a divergent impact on different protected groups. Finally, different compression methods have drastically different safety impacts: for example, quantization mostly preserves bias while pruning degrades quickly. Our findings underscore the importance of integrating safety assessments into the development of compressed LLMs to ensure their reliability across real-world applications.

Interpretability and Analysis of Models for NLP 2

Nov 12 (Tue) 14:00-15:30 - Room: Jasmine

Nov 12 (Tue) 14:00-15:30 - Jasmine

Insights into LLM Long-Context Failures: When Transformers Know but Don't Tell

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, Daniel Khashabi

Large Language Models (LLMs) exhibit positional bias, struggling to utilize information from the middle or end of long contexts. Our study explores LLMs' long-context reasoning by probing their hidden representations. We find that while LLMs encode the position of target information, they often fail to leverage this in generating accurate responses. This reveals a disconnect between information retrieval and utilization, a "know but don't tell" phenomenon. We further analyze the relationship between extraction time and final accuracy, offering insights into the underlying mechanics of transformer models.

Nov 12 (Tue) 14:00-15:30 - Jasmine

On Diversified Preferences of Large Language Model Alignment

Dun Zeng, Yong Dai, Pengyu Cheng, Longyue Wang, Tianhuo Hu, Wanshun CHEN, nan du, Zenglin Xu

Aligning large language models (LLMs) with human preferences has been recognized as the key to improving LLMs' interaction quality. However, in this pluralistic world, human preferences can be diversified due to annotators' different tastes, which hinders the effectiveness of LLM alignment methods. This paper presents the first quantitative analysis of the experimental scaling law for reward models with varying sizes, from 1.3 billion to 7 billion parameters, trained with human feedback exhibiting diverse preferences. Our analysis reveals that the impact of diversified human preferences depends on both model size and data size. Larger models with sufficient capacity mitigate the negative effects of diverse preferences, while smaller models struggle to accommodate them. To mitigate the impact of diverse preferences, we introduce a new metric, Expected Calibration Error (ECE), to evaluate RMs and show their obvious positive correlation with the alignment performance of LLMs. Furthermore, we propose a Multi-Objective Reward learning method (MORE) to enhance the calibration performance of RMs on shared preferences. Through experiments on four models and five human preference datasets, we find the calibration error can be adopted as a key metric for evaluating RMs and MORE can obtain superior alignment performance.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Evaluating Readability and Faithfulness of Concept-based Explanations

Meng Li, Haoran Jin, Ruixuan HUANG, Zhihao Xu, Defu Liang, Zijia Lin, Di ZHANG, Xiting Wang

With the growing popularity of general-purpose Large Language Models (LLMs), comes a need for more global explanations of model behaviors. Concept-based explanations arise as a promising avenue for explaining high-level patterns learned by LLMs. Yet their evaluation poses unique challenges, especially due to their non-local nature and high dimensional representation in a model's hidden space. Current methods approach concepts from different perspectives, lacking a unified formalization. This makes evaluating the core measures of concepts, namely faithfulness or readability, challenging. To bridge this gap, we introduce a formal definition of concepts generalizing to diverse concept-based explanations' settings. Based on this, we quantify the faithfulness of a concept explanation via perturbation. We ensure adequate perturbation in the high-dimensional space for different concepts via an optimization problem. Readability is approximated via an automatic and deterministic measure, quantifying the coherence of patterns that maximally activate a concept while aligning with human understanding. Finally, based on measurement theory, we apply a meta-evaluation method for evaluating these measures, generalizable to other types of explanations or tasks as well. Extensive experimental analysis has been conducted to inform the selection of explanation evaluation measures.

Nov 12 (Tue) 14:00-15:30 - Jasmine

From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP

Marius Moshach, Vagrant Gautam, Tomáš Vergara Browne, Dietrich Klakow, Mor Geva

Interpretability and analysis (IA) research is a growing subfield within NLP with the goal of developing a deeper understanding of the behavior or inner workings of NLP systems and methods. Despite growing interest in the subfield, a criticism of this work is that it lacks actionable insights and therefore has little impact on NLP. In this paper, we seek to quantify the impact of IA research on the broader field of NLP. We approach this with a mixed-methods analysis of: (1) a citation graph of 185K+ papers built from all papers published at ACL and EMNLP conferences from 2018 to 2023, and their references and citations, and (2) a survey of 138 members of the NLP community. Our quantitative results show that IA work is well-cited outside of IA, and central in the NLP citation graph. Through qualitative analysis of survey responses and manual annotation of 556 papers, we find that NLP researchers build on findings from IA work and perceive it as important for progress in NLP, multiple subfields, and rely on its findings and terminology for their own work. Many novel methods are proposed based on IA findings and highly influenced by them, but highly influential non-IA work cites IA findings without being driven by them. We end by summarizing

what is missing in IA work today and provide a call to action, to pave the way for a more impactful future of IA research.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Estimating Knowledge in Large Language Models Without Generating a Single Token

Daniela Göttszman, Mor Geva

To evaluate knowledge in large language models (LLMs), current methods query the model and then evaluate its generated responses. In this work, we ask whether evaluation can be done *before* the model has generated any text. Concretely, is it possible to estimate how knowledgeable a model is about a certain entity, only from its internal computation? We study this question with two tasks: given a subject entity, the goal is to predict (a) the ability of the model to answer common questions about the entity, and (b) the factuality of open-ended responses generated by the model about the entity. Experiments with a variety of LLMs show that KEEN, a simple probe trained over internal subject representations, succeeds at both tasks — correlating with both the QA accuracy of the model per-subject and FActScore, a recent factuality metric in open-ended generation. Moreover, KEEN naturally aligns with the model's hedging behavior and faithfully reflects changes in the model's knowledge after fine-tuning. Lastly, we show a more interpretable yet equally performant variant of KEEN, which highlights a small set of tokens indicative of clusters and gaps in the model's knowledge. Being simple and lightweight, KEEN can be leveraged to guide decisions such as when it is appropriate to apply further training or augment queries with retrieval.

Nov 12 (Tue) 14:00-15:30 - Jasmine

EVEDIT: Event-based Knowledge Editing for Deterministic Knowledge Propagation

Jiateng Liu, Pengfei Yu, Yufi Zhang, Sha Li, Zixuan Zhang, Ruhui Sarikaya, Kevin Small, Heng Ji

The dynamic nature of real-world information necessitates knowledge editing (KE) in large language models (LLMs). The edited knowledge should propagate and facilitate the deduction of new information based on existing model knowledge. We term the existing related knowledge in LLM serving as the origination of knowledge propagation as "deduction anchors". However, current KE approaches, which only operate on (subject, relation, object) triple. We both theoretically and empirically observe that this simplified setting often leads to uncertainty when determining the deduction anchors, causing low confidence in their answers. To mitigate this issue, we propose a novel task of event-based knowledge editing that pairs facts with event descriptions. This task manifests not only a closer simulation of real-world editing scenarios but also a more logically sound setting, implicitly defining the deduction anchor and enabling LLMs to propagate knowledge confidently. We curate a new benchmark dataset Evedit derived from the CounterFact dataset and validate its superiority in improving model confidence. Moreover, while we observe that the event-based setting is significantly challenging for existing approaches, we propose a novel approach Self>Edit that showcases stronger performance, achieving 55.6% consistency improvement while maintaining the naturalness of generation.

Nov 12 (Tue) 14:00-15:30 - Jasmine

SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaozhe Liu, Xingyao Wang, Yangyi Chen, Jing Gao

Large language models (LLMs) often generate inaccurate or fabricated information and generally fail to indicate their confidence, which limits their broader applications. Previous work has elicited confidence from LLMs by direct or self-consistency prompting, or constructing specific datasets for supervised finetuning. The prompting-based approaches have inferior performance, and the training-based approaches are limited to binary or inaccurate group-level confidence estimates. In this work, we present SaySelf, a novel training framework that teaches LLMs to express more fine-grained confidence estimates. In addition, beyond the confidence scores, SaySelf initiates the process of directing LLMs to produce self-reflective rationales that clearly identify gaps in their parametric knowledge and explain their uncertainty. This is achieved by using an LLM to automatically summarize the uncertainties in specific knowledge via natural language. The summarization is based on the analysis of the inconsistency in multiple sampled reasoning chains, and the resulting data is utilized for supervised fine-tuning. Moreover, we utilize reinforcement learning with a meticulously crafted reward function to calibrate the confidence estimates, motivating LLMs to deliver accurate, high-confidence predictions and to penalize overconfidence in erroneous outputs. Experimental results demonstrate the effectiveness of SaySelf in reducing the confidence calibration error and maintaining the task performance. The generated self-reflective rationales are also reasonable and can further contribute to the calibration. The code is made public at <https://github.com/xu1868/SaySelf>.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Investigating Mysteries of CoT-Augmented Distillation

Somita Wadhwa, Silvio Amir, Byron C Wallace

Eliciting chain of thought (CoT) rationales - sequences of token that convey a "reasoning" process has been shown to consistently improve LLM performance on tasks like question answering. More recent efforts have shown that such rationales can also be used for model distillation. Including CoT sequences (elicited from a large "teacher" model) in addition to target labels when fine-tuning a small student model yields (often substantial) improvements. In this work we ask: Why and how does this additional training signal help in model distillation? We perform ablations to interrogate this, and report some potentially surprising results. Specifically: (1) Placing CoT sequences after labels (rather than before) realizes consistently better downstream performance – this means that no student "reasoning" is necessary at test time to realize gains. (2) When rationales are appended in this way, they need not be coherent reasoning sequences to yield improvements; performance increases are robust to permutations of CoT tokens, for example. In fact, (3) a small number of key tokens are sufficient to achieve improvements equivalent to those observed when full rationales are used in model distillation.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Personas as a Way to Model Truthfulness in Language Models

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, He He

Large language models (LLMs) are trained on vast amounts of text from the internet, which contains both factual and misleading information about the world. While unintuitive from a classic view of LMs, recent work has shown that the truth value of a statement can be elicited from the models representations. This paper presents an explanation for why LMs appear to know the truth despite not being trained with truth labels. We hypothesize that the pretraining data is generated by groups of (un)truthful agents whose outputs share common features, and they form a (un)truthful persona. By training on this data, LMs can infer and represent the persona in its activation space. This allows the model to separate truth from falsehoods and controls the truthfulness of its generation. We show evidence for the persona hypothesis via two observations: (1) we can probe whether a models answer will be truthful before it is generated; (2) finetuning a model on a set of facts improves its truthfulness on unseen topics. Next, using arithmetics as a synthetic environment, we show that structures of the pretraining data are crucial for the model to infer the truthful persona. Overall, our findings suggest that models can exploit hierarchical structures in the data to learn abstract concepts like truthfulness.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Discovering Knowledge-Critical Subnetworks in Pretrained Language Models

Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, Antoine Bosselut

Pretrained language models (LMs) encode implicit representations of knowledge in their parameters. However, localizing these representations and disentangling them from each other remains an open problem. In this work, we investigate whether pretrained language models contain various *knowledge-critical* subnetworks: particular sparse computational subgraphs that can, if removed, precisely suppress specific

knowledge the model has memorized. We propose a multi-objective differentiable masking scheme that can be applied to both weights and neurons to discover such subnetworks and show that we can use them to precisely remove specific knowledge from models while minimizing adverse effects on the behavior of the original model. We demonstrate our method on multiple GPT2 variants, uncovering highly sparse subnetworks (98%+ sparsity) that are critical for expressing specific collections of relational knowledge. When these subnetworks are removed, the remaining network maintains most of its initial abilities but struggles to represent the suppressed knowledge.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, Jiliang Tang

Large language models (LLMs) are susceptible to a type of attack known as jailbreaking, which misleads LLMs to output harmful contents. Although there are diverse jailbreak attack strategies, there is no unified understanding on why some methods succeed and others fail. This paper explores the behavior of harmful and harmless prompts in the LLM's representation space to investigate the intrinsic properties of successful jailbreak attacks. We hypothesize that successful attacks share some similar properties: They are effective in moving the representation of the harmful prompt towards the direction to the harmless prompts. We leverage hidden representations into the objective of existing jailbreak attacks to move the attacks along the acceptance direction, and conduct experiments to validate the above hypothesis using the proposed objective. We hope this study provides new insights into understanding how LLMs understand harmfulness information.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Does Large Language Model Contain Task-Specific Neurons?

Ran Song, Shizhu He, Shuting Jiang, Yanjuan Xian, Shengxiang Gao, Kang Liu, Zhengqiao Yu

Large language models (LLMs) have demonstrated remarkable capabilities in comprehensively handling various types of natural language processing (NLP) tasks. However, there are significant differences in the knowledge and abilities required for different tasks. Therefore, it is important to understand whether the same LLM processes different tasks in the same way. Are there specific neurons in a LLM for different tasks? Inspired by neuroscience, this paper pioneers the exploration of whether distinct neurons are activated when a LLM handles different tasks. Compared with current research exploring the neurons of language and knowledge, task-specific neurons present a greater challenge due to their abstractness, diversity, and complexity. To address these challenges, this paper proposes a method for task-specific neuron localization based on Causal Gradient Variation with Special Tokens (CGVST). CGVST identifies task-specific neurons by concentrating on the most significant tokens during task processing, thereby eliminating redundant tokens and minimizing interference from non-specific neurons. Compared to traditional neuron localization methods, our approach can more effectively identify task-specific neurons. We conduct experiments across eight different public tasks. Experiments involving the inhibition and amplification of identified neurons demonstrate that our method can accurately locate task-specific neurons.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Unveiling Factual Recall Behaviors of Large Language Models through Knowledge Neurons

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Lingjing Li, Daniel Dajun Zeng

In this paper, we investigate whether Large Language Models (LLMs) actively recall or retrieve their internal repositories of factual knowledge when faced with reasoning tasks. Through an analysis of LLMs' internal factual recall at each reasoning step via Knowledge Neurons, we reveal that LLMs fail to harness the critical factual associations under certain circumstances. Instead, they tend to opt for alternative, shortcut-like pathways to answer reasoning questions. By manually manipulating the recall process of parametric knowledge in LLMs, we demonstrate that enhancing this recall process directly improves reasoning performance whereas suppressing it leads to notable degradation. Furthermore, we assess the effect of Chain-of-Thought (CoT) prompting, a powerful technique for addressing complex reasoning tasks. Our findings indicate that CoT can intensify the recall of factual knowledge by encouraging LLMs to engage in orderly and reliable reasoning. Furthermore, we explored how contextual conflicts affect the retrieval of facts during the reasoning process to gain a comprehensive understanding of the factual recall behaviors of LLMs. Code and data will be available soon.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment

Vyas Raina, Aditan Liu, Mark Gales

Large Language Models (LLMs) are powerful zero-shot assessors used in real-world situations such as assessing written exams and benchmarking systems. Despite these critical applications, no existing work has analyzed the vulnerability of judge-LLMs to adversarial manipulation. This work presents the first study on the adversarial robustness of assessment LLMs, where we demonstrate that short universal adversarial phrases can be concatenated to deceive judge-LLMs to predict inflated scores. Since adversaries may not know or have access to the judge-LLMs, we propose a simple surrogate attack where a surrogate model is first attacked, and the learned attack phrase then transferred to unknown judge-LLMs. We propose a practical algorithm to determine the short universal attack phrases and demonstrate that when transferred to unseen models, scores can be drastically inflated such that irrespective of the assessed text, maximum scores are predicted. It is found that judge-LLMs are significantly more susceptible to these adversarial attacks when used for absolute scoring, as opposed to comparative assessment. Our findings raise concerns on the reliability of LLM-as-a-judge methods, and emphasize the importance of addressing vulnerabilities in LLM assessment methods before deployment in high-stakes real-world scenarios.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?

Gal Youn, Roee Aharoni, Mor Geva

We posit that large language models (LLMs) should be capable of expressing their intrinsic uncertainty in natural language. For example, if the LLM is equally likely to output two contradicting answers to the same question, then its generated response should reflect this uncertainty by hedging its answer (e.g., "I'm not sure, but I think..."). We formalize faithful response uncertainty based on the gap between the model's intrinsic confidence in the assertions it makes and the decisiveness by which they are conveyed. This example-level metric reliably indicates whether the model reflects its uncertainty, as it penalizes both excessive and insufficient hedging. We evaluate a variety of aligned LLMs at faithfully conveying uncertainty on several knowledge-intensive question answering tasks. Our results provide strong evidence that modern LLMs are poor at faithfully conveying their uncertainty, and that better alignment is necessary to improve their trustworthiness.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Attribute or Abstain: Large Language Models as Long Document Assistants

Jan Buchmann, Xiao Liu, Iryna Gurevych

LLMs can help humans working with long documents, but are known to hallucinate. *Attribution* can increase trust in LLM responses: The LLM provides evidence that supports its response, which enhances verifiability. Existing approaches to attribution have only been evaluated in RAG settings, where the initial retrieval confounds LLM performance. This is crucially different from the long document setting, where retrieval is not needed, but could help. Thus, a long document specific evaluation of attribution is missing. To fill this gap, we present LAB, a benchmark of 6 diverse long document tasks with attribution, and experiments with different approaches to attribution on 5 LLMs of different sizes. We find that *citation*, i.e. response generation and evidence extraction in one step, performs best for large and fine-tuned models,

while additional retrieval can help for small, prompted models. We investigate whether the "Lost in the Middle" phenomenon exists for attribution, but do not find this. We also find that evidence quality can predict response quality on datasets with simple responses, but not so for complex responses, as models struggle with providing evidence for complex claims. We release code and data for further investigation. [Link](<https://github.com/UKPLab/arxiv2024-attribute-or-abstain>)

Nov 12 (Tue) 14:00-15:30 - Jasmine

Beyond Label Attention: Transparency in Language Models for Automated Medical Coding via Dictionary Learning

John Wu, David Wu, Jimeng Sun

Medical coding, the translation of unstructured clinical text into standardized medical codes, is a crucial but time-consuming healthcare practice. Though large language models (LLM) could automate the coding process and improve the efficiency of such tasks, interpretability remains paramount for maintaining patient trust. Current efforts in interpretability of medical coding applications rely heavily on label attention mechanisms, which often leads to the highlighting of extraneous tokens irrelevant to the ICD code. To facilitate accurate interpretability in medical language models, this paper leverages dictionary learning that can efficiently extract sparsely activated representations from dense language model embeddings in superposition. Compared with common label attention mechanisms, our model goes beyond token-level representations by building an interpretable dictionary which enhances the mechanistic-based explanations for each ICD code prediction, even when the highlighted tokens are medically irrelevant. We show that dictionary features are human interpretable, can elucidate the hidden meanings of upwards of 90% of medically irrelevant tokens, and steer model behavior.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs

Sheridan Feucht, David Atkinson, Byron C Wallace, David Bau

LLMs process text as sequences of tokens that roughly correspond to words, where less common words are represented by multiple tokens. However, individual tokens are often semantically unrelated to the meanings of the words/concepts they comprise. For example, Llama-2-7b's tokenizer splits the word "patrolling" into two tokens, "pat" and "rolling", neither of which correspond to semantically meaningful units like "patrol" or "-ing." Similarly, the overall meanings of named entities like "Neil Young" and multi-word expressions like "break a leg" cannot be directly inferred from their constituent tokens. Mechanistically, how do LLMs convert such arbitrary groups of tokens into useful higher-level representations? In this work, we find that last token representations of named entities and multi-token words exhibit a pronounced "erasure" effect, where information about previous and current tokens is rapidly forgotten in early layers. Using this observation, we propose a method to "read out" the implicit vocabulary of an autoregressive LLM by examining differences in token representations across layers, and present results of this method for Llama-2-7b and Llama-3-8B. To our knowledge, this is the first attempt to probe the implicit vocabulary of an LLM.

Nov 12 (Tue) 14:00-15:30 - Jasmine

When Parts are Greater Than Sums: Individual LLM Components Can Outperform Full Models

Ting-Yun Chang, Jesse Thomason, Robin Jia

This paper studies in-context learning by decomposing the output of large language models into the individual contributions of attention heads and MLPs (components). We observe curious components: good-performing ones that individually do well on a classification task, even when the model performs poorly; bad-performing ones that do much worse than chance; and label-biased components that always predict the same label. We find that component accuracies are well-correlated across different demonstration sets and perturbations of prompt templates. Based on our findings, we propose component reweighting, which learns to linearly re-scale the component activations from a few labeled examples. Given 24 labeled examples, our method improves by an average of 6.0% accuracy points over 24-shot ICL across 8 tasks on Llama-2-7B. Overall, this paper both enriches our understanding of ICL and provides a practical method for improvement by examining model internals.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Encourage or Inhibit Monosemanticity? Revisit Monosemanticity from a Feature Decorrelation Perspective

Hanqi Yang, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, Yulan He

To better interpret the intrinsic mechanism of large language models (LLMs), recent studies focus on monosemanticity on its basic units. A monosemantic neuron is dedicated to a single and specific concept, which forms a one-to-one correlation between neurons and concepts. Despite extensive research in monosemanticity probing, it remains unclear whether monosemanticity is beneficial or harmful to model capacity. To explore this question, we revisit monosemanticity from the feature decorrelation perspective and advocate for its encouragement. We experimentally observe that the current conclusion by wang2024learning, which suggests that decreasing monosemanticity enhances model performance, does not hold when the model changes. Instead, we demonstrate that monosemanticity consistently exhibits a positive correlation with model capacity, in the preference alignment process. Consequently, we apply feature correlation as a proxy for monosemanticity and incorporate a feature decorrelation regularizer into the dynamic preference optimization process. The experiments show that our method not only enhances representation diversity and activation sparsity but also improves preference alignment performance.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Demystifying Verbatim Memorization in Large Language Models

Jing Huang, Diyi Yang, Christopher Potts

Large Language Models (LLMs) frequently memorize long sequences verbatim, often with serious legal and privacy implications. Much prior work has studied such verbatim memorization using observational data. To complement such work, we develop a framework to study verbatim memorization in a controlled setting by continuing pre-training from Pythia checkpoints with injected sequences. We find that (1) non-trivial amounts of repetition are necessary for verbatim memorization to happen; (2) later (and presumably better) checkpoints are more likely to verbatim memorize sequences, even for out-of-distribution sequences; (3) the generation of memorized sequences is triggered by distributed model states that encode high-level features and makes important use of general language modeling capabilities. Guided by these insights, we develop stress tests to evaluate unlearning methods and find they often fail to remove the verbatim memorized information, while also degrading the LM. Overall, these findings challenge the hypothesis that verbatim memorization stems from specific model weights or mechanisms. Rather, verbatim memorization is intertwined with the LM's general capabilities and thus will be very difficult to isolate and suppress without degrading model quality.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Stable Language Model Pre-training by Reducing Embedding Variability

Woojin Chung, Jiwoo Hong, Na Min An, James Thorne, Se-Young Yun

Stable pre-training is essential for achieving better-performing language models. However, tracking pre-training stability is impractical due to high computational costs. We study Token Embedding Variability as a simple proxy to estimate pre-training stability. We theoretically and empirically demonstrate that Multi-head Low-Rank Attention acts as a fundamental approach to reducing instability. This is supported by empirical findings on variants on GPT-2, demonstrating improved stability and lower perplexities, even at deeper layer counts.

Nov 12 (Tue) 14:00-15:30 - Jasmine

A Multi-Perspective Analysis of Memorization in Large Language Models

Bowen Chen, Namgi Han, Yusuke Miyao

Large Language Models (LLMs) can generate the same sequences contained in the pre-train corpora, known as memorization. Previous research studied it at a macro level, leaving micro yet important questions under-explored, e.g., what makes sentences memorized, the dynamics when generating memorized sequence, its connection to unmemorized sequence, and its predictability. We answer the above questions by analyzing the relationship of memorization with outputs from LLM, namely, embeddings, probability distributions, and generated tokens. A memorization score is calculated as the overlap between generated tokens and actual continuations when the LLM is prompted with a context sequence from the pre-train corpora. Our findings reveal: (1) The inter-correlation between memorized/unmemorized sentences, model size, continuation size, and context size, as well as the transition dynamics between sentences of different memorization scores, (2) A sudden drop and increase in the frequency of input tokens when generating memorized/unmemorized sequences (boundary effect), (3) Cluster of sentences with different memorization scores in the embedding space, (4) An inverse boundary effect in the entropy of probability distributions for generated memorized/unmemorized sequences, (5) The predictability of memorization is related to model size and continuation length. In addition, we show a Transformer model trained by the hidden states of LLM can predict unmemorized tokens.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Experimental Contexts Can Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently

Kaniska Misra, Allyson Ettinger, Kyle Mahowald

Recent zero-shot evaluations have highlighted important limitations in the abilities of language models (LMs) to perform meaning extraction. However, it is now well known that LMs can demonstrate radical improvements in the presence of experimental contexts such as in-context examples and instructions. How well does this translate to previously studied meaning-sensitive tasks? We present a case-study on the extent to which experimental contexts can improve LMs' robustness in performing property inheritance—predicting semantic properties of novel concepts, a task that they have been previously shown to fail on. Upon carefully controlling the nature of the in-context examples and the instructions, our work reveals that they can indeed lead to non-trivial property inheritance behavior in LMs. However, this ability is inconsistent: with a minimal reformulation of the task, some LMs were found to pick up on shallow, non-semantic heuristics from their inputs, suggesting that the computational principles of semantic property inference are yet to be mastered by LMs.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Latent Concept-based Explanation of NLP Models

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, Hassan Sajjad

Interpreting and understanding the predictions made by deep learning models poses a formidable challenge due to their inherently opaque nature. Many previous efforts aimed at explaining these predictions rely on input features, specifically, the words within NLP models. However, such explanations are often less informative due to the discrete nature of these words and their lack of contextual verbosity. To address this limitation, we introduce the Latent Concept Attribution method (LACOAT), which generates explanations for predictions based on latent concepts. Our foundational intuition is that a word can exhibit multiple facets, contingent upon the context in which it is used. Therefore, given a word in context, the latent space derived from our training process reflects a specific facet of that word. LACOAT functions by mapping the representations of salient input words into the training latent space, allowing it to provide latent context-based explanations of the prediction.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Cluster-Norm for Unsupervised Probing of Knowledge

Walter Laurito, Sharan Maiya, Grégoire DHIMOÏLA, Owen Ho Wan Yeung, Kaarel Hänni

The deployment of language models brings challenges in generating reliable text, especially when these models are fine-tuned with human preferences. To extract the encoded knowledge in these models without (potentially) biased human labels, unsupervised probing techniques like Contrast-Consistent Search (CCS) have been developed (Burns et al., 2022). However, salient but unrelated features in activation space can mislead these probes (Farquhar et al., 2023). Addressing this, we propose a cluster-normalization method to minimize the impact of such features by clustering and normalizing activations of contrast pairs before applying unsupervised probing techniques. While this approach does not address the issue of distinguishing between latent knowledge and that portrayed by a simulated agent, a major issue in the literature of eliciting latent knowledge (Paul Christiano and Xu, 2021) it still significantly improves the accuracy of probes in identifying the intended knowledge amidst distractions.

Nov 12 (Tue) 14:00-15:30 - Jasmine

LLM Task Interference: An Initial Study on the Impact of Task-Switch in Conversational History

Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, Mario Fritz

With the recent emergence of powerful instruction-tuned large language models (LLMs), various helpful conversational Artificial Intelligence (AI) systems have been deployed across many applications. When prompted by users, these AI systems successfully perform a wide range of tasks as part of a conversation. To provide some sort of memory and context, such approaches typically condition their output on the entire conversational history. Although this sensitivity to the conversational history can often lead to improved performance on subsequent tasks, we find that performance can in fact also be negatively impacted, if there is a _task-switch_. To the best of our knowledge, our work makes the first attempt to formalize the study of such vulnerabilities and interference of tasks in conversational LLMs caused by task-switches in the conversational history. Our experiments across 5 datasets with 15 task switches using popular LLMs reveal that many of the task-switches can lead to significant performance degradation.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Layer by Layer: Uncovering Where Multi-Task Learning Happens in Instruction-Tuned Large Language Models

Zhen Zhao, Yifah Ziser, Shay B Cohen

Fine-tuning pre-trained large language models (LLMs) on a diverse array of tasks has become a common approach for building models that can solve various natural language processing (NLP) tasks. However, where and to what extent these models retain task-specific knowledge remains largely unexplored. This study investigates the task-specific information encoded in pre-trained LLMs and the effects of instruction tuning on their representations across a diverse set of over 60 NLP tasks. We use a set of matrix analysis tools to examine the differences between the way pre-trained and instruction-tuned LLMs store task-specific information. Our findings reveal that while some tasks are already encoded within the pre-trained LLMs, others greatly benefit from instruction tuning. Additionally, we pinpointed the layers in which the model transitions from high-level general representations to more task-oriented representations. This finding extends our understanding of the governing mechanisms of LLMs and facilitates future research in the fields of parameter-efficient transfer learning and multi-task learning. Our code is available at: https://github.com/zsquaredz/layer_by_layer/

Nov 12 (Tue) 14:00-15:30 - Jasmine

CMR Scaling Law: Predicting Critical Mixture Ratios for Continual Pre-training of Language Models

Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, Fei Tan

Large Language Models (LLMs) excel in diverse tasks but often underperform in specialized fields due to limited domain-specific or proprietary corpora. Continual pre-training (CPT) enhances LLM capabilities by imbuing new domain-specific or proprietary knowledge while

replaying general corpus to prevent catastrophic forgetting. The data mixture ratio of general corpus and domain-specific corpus, however, has been chosen heuristically, leading to sub-optimal training efficiency in practice. In this context, we attempt to re-visit the scaling behavior of LLMs under the hood of CPT, and discover a power-law relationship between loss, mixture ratio, and training tokens scale. We formalize the trade-off between general and domain-specific capabilities, leading to a well-defined Critical Mixture Ratio (CMR) of general and domain data. By striking the balance, CMR maintains the model's general ability and achieves the desired domain transfer, ensuring the highest utilization of available resources. Considering the balance between efficiency and effectiveness, CMR can be regarded as the optimal mixture ratio. Through extensive experiments, we ascertain the predictability of CMR, propose CMR scaling law and have substantiated its generalization. These findings offer practical guidelines for optimizing LLM training in specialized domains, ensuring both general and domain-specific performance while efficiently managing training resources.

Nov 12 (Tue) 14:00-15:30 - Jasmine

On the Robustness of Editing Large Language Models

Xinbei Ma, Tianjie Ju, Jiayang Qiu, Zhuosheng Zhang, hai zhao, lifeng Liu, Yulong Wang

Large language models (LLMs) have played a pivotal role in building communicative AI, yet they encounter the challenge of efficient updates. Model editing enables the manipulation of specific knowledge memories and the behavior of language generation without retraining. However, the robustness of model editing remains an open question. This work seeks to understand the strengths and limitations of editing methods, facilitating practical applications of communicative AI. We focus on three key research questions. RQ1: Can edited LLMs behave consistently resembling communicative AI in realistic situations? RQ2: To what extent does the rephrasing of prompts lead LLMs to deviate from the edited knowledge memory? RQ3: Which knowledge features are correlated with the performance and robustness of editing? Our empirical studies uncover a substantial disparity between existing editing methods and the practical application of LLMs. On rephrased prompts that are flexible but common in realistic applications, the performance of editing experiences a significant decline. Further analysis shows that more popular knowledge is memorized better, easier to recall, and more challenging to edit effectively.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Information Flow Routes: Automatically Interpreting Language Models at Scale

Javier Ferrando, Elena Voita

Information flows by routes inside the network via mechanisms implemented in the model. These routes can be represented as graphs where nodes correspond to token representations and edges to computations. We automatically build these graphs in a top-down manner, for each prediction leaving only the most important nodes and edges. In contrast to the existing workflows relying on activation patching, we do this through attribution, thus allowing us to efficiently uncover existing circuits with just a single forward pass. Unlike with patching, we do not need a human to carefully design prediction templates, and we can extract information flow routes for any prediction (not just the ones among the allowed templates). As a result, we can analyze model behavior in general, for specific types of predictions, or different domains. We experiment with Llama 2 and show that some attention head roles are overall important, e.g. previous token heads and subword merging heads. Next, we find similarities in Llama 2 behavior when handling tokens of the same part of speech. Finally, we show that some model components can be specialized on domains such as coding or multilingual texts.

Nov 12 (Tue) 14:00-15:30 - Jasmine

A linguistically-motivated evaluation methodology for unraveling model's abilities in reading comprehension tasks

Elie Antoine, Frédéric Bechet, Géraldine Dammé, Philippe Langlais

We introduce an evaluation methodology for reading comprehension tasks based on the intuition that certain examples, by the virtue of their linguistic complexity, consistently yield lower scores regardless of model size or architecture. We capitalize on semantic frame annotation for characterizing this complexity, and study seven complexity factors that may account for model's difficulty. We first deploy this methodology on a carefully annotated French reading comprehension benchmark showing that two of those complexity factors are indeed good predictors of models' failure, while others are less so. We further deploy our methodology on a well studied English benchmark by using chatGPT as a proxy for semantic annotation. Our study reveals that fine-grained linguistically-motivated automatic evaluation of a reading comprehension task is not only possible, but helps understand models' abilities to handle specific linguistic characteristics of input examples. It also shows that current state-of-the-art models fail with some for those characteristics which suggests that adequately handling them requires more than merely increasing model size.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Transformers are Multi-State RNNs

Matanet Oren, Michael Hassid, Nir Yarden, Yossi Adi, Roy Schwartz

Transformers are considered conceptually different than the previous generation of state-of-the-art NLP models - recurrent neural networks (RNNs). In this work, we demonstrate that decoder-only transformers can in fact be conceptualized as *unbounded* multi-state RNNs—an RNN variant with unlimited hidden state size. We further show that transformers can be converted into *bounded* multi-state RNNs by fixing the size of their hidden state, effectively compressing their key-value cache. We introduce a novel, training-free compression policy - *Token Omission Via Attention (TOVA)*. Our experiments with four long range tasks and several LLMs show that TOVA outperforms several baseline compression policies. Particularly, our results are nearly on par with the full model, using in some cases only $\frac{1}{8}$ of the original cache size, which translates to 4.8X higher throughput. Our results shed light on the connection between transformers and RNNs, and help mitigate one of LLMs' most painful computational bottlenecks—the size of their key-value cache. We will publicly release our code upon publication.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Fuse to Forget: Bias Reduction and Selective Memorization through Model Fusion

Kerem Zaman, Leshem Choshen, Shashank Srivastava

Model fusion research aims to aggregate the knowledge of multiple individual models to enhance performance by combining their weights. In this work, we study the inverse problem: investigating whether model fusion can be used to reduce unwanted knowledge. We investigate the effects of model fusion in three scenarios: the learning of shortcuts, social biases, and memorization of training data in fine-tuned language models. Through experiments covering classification and generation tasks, our analysis highlights that shared knowledge among models is enhanced during model fusion, while unshared knowledge is usually forgotten. Based on this observation, we demonstrate the potential of model fusion as a debiasing tool and showcase its efficacy in addressing privacy concerns associated with language models.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Towards Faithful Knowledge Graph Explanation Through Deep Alignment in Commonsense Question Answering

WEIHE ZHAI, Arkaitz Zubiaaga, Bingquan Liu, Chengjie Sun, Yalong Zhao

The fusion of language models (LMs) and knowledge graphs (KGs) is widely used in commonsense question answering, but generating faithful explanations remains challenging. Current methods often overlook path decoding faithfulness, leading to divergence between graph encoder outputs and model predictions. We identify confounding effects and LM-KG misalignment as key factors causing spurious explanations. To address this, we introduce the LM-KG Fidelity metric to assess KG representation reliability and propose the LM-KG Distribution-

aware Alignment (LKDA) algorithm to improve explanation faithfulness. Without ground truth, we evaluate KG explanations using the proposed Fidelity-Sparsity Trade-off Curve. Experiments on CommonsenseQA and OpenBookQA show that LKDA significantly enhances explanation fidelity and model performance, highlighting the need to address distributional misalignment for reliable commonsense reasoning.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Factuality of Large Language Models in the Year 2024

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, Preslav Nakov
Large language models (LLMs), especially when instruction-tuned for chat, have become part of our daily lives, freeing people from the process of searching, extracting, and integrating information from multiple sources by offering a straightforward answer to a variety of questions in a single place. Unfortunately, in many cases, LLM responses are factually incorrect, which limits their applicability in real-world scenarios. As a result, research on evaluating and improving the factuality of LLMs has attracted a lot of research attention recently. In this survey, we critically analyze existing work with the aim to identify the major challenges and their associated causes, pointing out to potential solutions for improving the factuality of LLMs, and analyzing the obstacles to automated factuality evaluation for open-ended text generation. We further offer an outlook on where future research should go.

Nov 12 (Tue) 14:00-15:30 - Jasmine

DISCERN: Decoding Systematic Errors in Natural Language for Text Classifiers

Rakesh R Menon, Shashank Srivastava
Despite their high predictive accuracies, current machine learning systems often exhibit systematic biases stemming from annotation artifacts or insufficient support for certain classes in the dataset. Recent work proposes automatic methods for identifying and explaining systematic biases using keywords. We introduce DISCERN, a framework for interpreting systematic biases in text classifiers using language explanations. DISCERN iteratively generates precise natural language descriptions of systematic errors by employing an interactive loop between two large language models. Finally, we use the descriptions to improve classifiers by augmenting classifier training sets with synthetically generated instances or annotated examples via active learning. On three text-classification datasets, we demonstrate that language explanations from our framework induce consistent performance improvements that go beyond what is achievable with exemplars of systematic bias. Finally, in human evaluations, we show that users can interpret systematic biases more effectively (by over 25% relative) and efficiently when described through language explanations as opposed to cluster exemplars.

Nov 12 (Tue) 14:00-15:30 - Jasmine

MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangning Pan, William Yang Wang
Large Language Models (LLMs) have shown exceptional results on current benchmarks when working individually. The advancement in their capabilities, along with a reduction in parameter size and inference times, has facilitated the use of these models as agents, enabling interactions among multiple models to execute complex tasks. Such collaborations offer several advantages, including the use of specialized models (e.g., coding), improved confidence through multiple computations, and enhanced divergent thinking, leading to more diverse outputs. Thus, the collaborative use of language models is expected to grow significantly in the coming years. In this work, we evaluate the behavior of a network of models collaborating through debate under the influence of an adversary. We introduce pertinent metrics to assess the adversary's effectiveness, focusing on system accuracy and model agreement. Our findings highlight the importance of a models persuasive ability in influencing others. Additionally, we explore inference-time methods to generate more compelling arguments and evaluate the potential of prompt-based mitigation as a defensive strategy.

Nov 12 (Tue) 14:00-15:30 - Jasmine

InternalInspector^{1.2}: Robust Confidence Estimation in LLMs through Internal States

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, Lifu Huang
Despite their vast capabilities, Large Language Models (LLMs) often struggle with generating reliable outputs, frequently producing high-confidence inaccuracies known as hallucinations. Addressing this challenge, our research introduces InternalInspector, a novel framework designed to enhance confidence estimation in LLMs by leveraging contrastive learning on internal states including attention states, feed-forward states, and activation states of all layers. Unlike existing methods that primarily focus on the final activation state, InternalInspector conducts a comprehensive analysis across all internal states of every layer to accurately identify both correct and incorrect prediction processes. By benchmarking InternalInspector against existing confidence estimation methods across various natural language understanding and generation tasks, including factual question answering, commonsense reasoning, and reading comprehension, InternalInspector achieves significantly higher accuracy in aligning the estimated confidence scores with the correctness of the LLM's predictions and lower calibration error. Furthermore, InternalInspector excels at HaluEval, a hallucination detection benchmark, outperforming other internal-based confidence estimation methods in this task.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Semantic Token Reweighting for Interpretable and Controllable Text Embeddings in CLIP

Eunji Kim, Kyuhong Shim, Simyoung Chang, Sungroh Yoon
A text encoder within Vision-Language Models (VLMs) like CLIP plays a crucial role in translating textual input into an embedding space shared with images, thereby facilitating the interpretative analysis of vision tasks through natural language. Despite the varying significance of different textual elements within a sentence depending on the context, efforts to account for variation of importance in constructing text embeddings have been lacking. We propose a framework of Semantic Token Reweighting to build Interpretable text embeddings (SToRI), which incorporates controllability as well. SToRI refines the text encoding process in CLIP by differentially weighting semantic elements based on contextual importance, enabling finer control over emphasis responsive to data-driven insights and user preferences. The efficacy of SToRI is demonstrated through comprehensive experiments on few-shot image classification and image retrieval tailored to user preferences.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Infer-then-Verbalize: How do LMs Map true/false to cat/dog During In-Context Learning?

Junyi Tao, Xiaoyan Chen, Nelson F. Liu
Large language models (LMs) are capable of in-context learning from a few demonstrations (example-label pairs) to solve new tasks during inference. Despite the intuitive importance of high-quality demonstrations, previous work has observed that, in some settings, ICL performance is minimally affected by irrelevant labels (Min et al., 2022). We hypothesize that LMs perform ICL with irrelevant labels via two sequential processes: an inference function that solves the task, followed by a verbalization function that maps the inferred answer to the label space. Importantly, we hypothesize that the inference function is invariant to remappings of the label space (e.g., true/false to cat/dog), enabling LMs to share the same inference function across settings with different label words. We empirically validate this hypothesis with controlled layer-wise interchange intervention experiments. Our findings confirm the hypotheses on multiple datasets and tasks (natural language inference, sentiment analysis, and topic classification) and further suggest that the two functions can be localized in specific layers across various open-sourced models, including GEMMA-7B, MISTRAL-7B-V0.3, GEMMA-2-27B, and LLAMA-3.1-70B.

Language Modeling 2

Nov 12 (Tue) 14:00-15:30 - Room: Riverfront Hall

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Rethinking the Reversal Curse of LLMs: a Prescription from Human Knowledge Reversal

Zhicong Lu, Li Jin, Peiguang Li, Yu Tian, Linhao Zhang, Sirui Wang, Guangluan Xu, Changyuan Tian, Xunliang Cai

Large Language Models (LLMs) have exhibited exceptional performance across diverse domains. However, recent studies reveal that LLMs are plagued by the "reversal curse". Most existing methods rely on aggressive sample permutation and pay little attention to delving into the underlying reasons for this issue, resulting in only partial mitigation. In this paper, inspired by human knowledge reversal, we investigate and quantify the individual influence of three potential reasons on the reversal curse: 1) knowledge clarity, 2) entity correlation modeling, and 3) pairwise relationship reasoning capability. Motivated by the analysis of these reasons, we propose a novel **P**airwise entity **R**eversal-and-**R**elationship-**E**nhanced (**PORE**) data strategy, which facilitates bidirectional entity correlation modeling and pairwise relationship reasoning to overcome the reversal curse. Specifically, PORE augments the samples with entity order-reversal and semantically preserved question-answer pairs, enhancing the encoding of entity correlations in both directions. PORE also employs entity-interleaved pairwise relationship data, which elevates the model's capability for relationship reasoning. Additionally, to improve the recall of reverse relationships, we leverage knowledge clarity to construct high-clarity data for PORE. Extensive experimental results on available and two newly assembled datasets demonstrate the effectiveness and generalization of our method in both data-sufficient and -constrained situations.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

NumeroLogic: Number Encoding for Enhanced LLMs' Numerical Reasoning

Eli Schwartz, Leshem Choshen, Joseph Shitok, Sivan Dovoh, Leonid Karlinsky, Assaf Arbel

Language models struggle with handling numerical data and performing arithmetic operations. We hypothesize that this limitation can be partially attributed to non-intuitive textual numbers representation. When a digit is read or generated by a causal language model it does not know its place value (e.g. thousands vs. hundreds) until the entire number is processed. To address this issue, we propose a simple adjustment to how numbers are represented by including the count of digits before each number. For instance, instead of "42", we suggest using "2|42" as the new format. This approach, which we term NumeroLogic, offers an added advantage in number generation by serving as a Chain of Thought (CoT). By requiring the model to consider the number of digits first, it enhances the reasoning process before generating the actual number. We use arithmetic tasks to demonstrate the effectiveness of the NumeroLogic formatting. We further demonstrate NumeroLogic applicability to general natural language modeling, improving language understanding performance in the MMLU benchmark.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models

Zihan Wang, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, Yu Wu

Parameter-efficient fine-tuning (PEFT) is crucial for customizing Large Language Models (LLMs) with constrained resource. Although there have been various PEFT methods for dense-architecture LLMs, PEFT for sparse-architecture LLMs is still underexplored. In this work, we study the PEFT method for LLMs with the Mixture-of-Experts (MoE) architecture and the contents of this work are mainly threefold: (1) We investigate the dispersion degree of the activated experts in customized tasks, and found that the routing distribution for specific task tend to be highly concentrated, while the distribution of activated experts varies significantly across different tasks. (2) We propose the expert-specialized fine-tuning method, which tunes the experts most relevant to downstream tasks while freezing the other experts; experimental results demonstrate that our method not only improves the tuning efficiency, but also matches or even surpasses the performance of full-parameter fine-tuning. (3) We further analyze the impact of the MoE architecture on expert-specialized fine-tuning. We find that MoE models with finer-grained experts are more advantageous in selecting the combination of experts that are most relevant to downstream tasks, thereby enhancing the both the training efficiency and effectiveness.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

A Survey on In-context Learning

Qingxu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, Zifeng Sui

With the increasing capabilities of large language models (LLMs), in-context learning (ICL) has emerged as a new paradigm for natural language processing (NLP), where LLMs make predictions based on contexts augmented with a few examples. It has been a significant trend to explore ICL to evaluate and extrapolate the ability of LLMs. In this paper, we aim to survey and summarize the progress and challenges of ICL. We first present a formal definition of ICL and clarify its correlation to related studies. Then, we organize and discuss advanced techniques, including training strategies, prompt designing strategies, and related analysis. Additionally, we explore various ICL application scenarios, such as data engineering and knowledge updating. Finally, we address the challenges of ICL and suggest potential directions for further research. We hope that our work can encourage more research on uncovering how ICL works and improving ICL.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, Ahmad Beirami

Aligning language models (LMs) based on human-annotated preference data is a crucial step in obtaining practical and performant LM-based systems. However, multilingual human preference data are difficult to obtain at scale, making it challenging to extend this framework to diverse languages. In this work, we evaluate a simple approach for zero-shot cross-lingual alignment, where a reward model is trained on preference data in one source language and directly applied to other target languages. On summarization and open-ended dialog generation, we show that this method is consistently successful under comprehensive evaluation settings, including human evaluation: cross-lingually aligned models are preferred by humans over unaligned models on up to >70% of evaluation instances. We moreover find that a different-language reward model sometimes yields better aligned models than a same-language reward model. We also identify best practices when there is no language-specific data for even supervised finetuning, another component in alignment.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

DecorateLM: Data Engineering through Corpus Rating, Tagging, and Editing with Language Models

Ranchi Zhao, Zhen Leng Thai, Yifan Zhang, Shengding Hu, Jie Zhou, Yunqi Ba, Jie Cai, Zhiyuan Liu, Maosong Sun

The performance of Large Language Models (LLMs) is substantially influenced by the pretraining corpus, which consists of vast quantities of unsupervised data processed by the models. Despite its critical role in model performance, ensuring the quality of this data is challenging due to its sheer volume and the absence of sample-level quality annotations and enhancements. In this paper, we introduce DecorateLM, a data engineering method designed to refine the pretraining corpus through data rating, tagging and editing. Specifically, DecorateLM rates

texts against quality criteria, tags texts with hierarchical labels, and edits texts into a more formalized format. Due to the massive size of the pretraining corpus, adopting an LLM fordecorating the entire corpus is less efficient. Therefore, to balance performance with efficiency, we curate a meticulously annotated training corpus for DecorateLM using a large language model and distill data engineering expertise into a compact 1.2 billion parameter small language model (SLM). We then apply DecorateLM to enhance 100 billion tokens of the training corpus, selecting 45 million tokens that exemplify high quality and diversity for the further training of another 1.2 billion parameter LLM. Our results demonstrate that employing such high-quality data can significantly boost model performance, showcasing a powerful approach to enhance the quality of the pretraining corpus.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LEMoE: Advanced Mixture of Experts Adaptor for Lifelong Model Editing of Large Language Models

Renchi Wang, Piji Li

Large language models (LLMs) require continual knowledge updates to stay abreast of the ever-changing world facts, prompting the formulation of lifelong model editing task. While recent years have witnessed the development of various techniques for single and batch editing, these methods either fail to apply or perform sub-optimally when faced with lifelong editing. In this paper, we introduce LEMoE, an advanced Mixture of Experts (MoE) adaptor for lifelong model editing. We first analyze the factors influencing the effectiveness of conventional MoE adaptor in lifelong editing, including catastrophic forgetting, inconsistent routing and order sensitivity. Based on these insights, we propose a tailored module insertion method to achieve lifelong editing, incorporating a novel KV anchor routing to enhance routing consistency between training and inference stage, along with a concise yet effective clustering-based editing order planning. Experimental results demonstrate the effectiveness of our method in lifelong editing, surpassing previous model editing techniques while maintaining outstanding performance in batch editing task. Our code will be available.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

CUTE: Measuring LLMs' Understanding of Their Tokens

Lukas Edman, Helmut Schmid, Alexander Fraser

Large Language Models (LLMs) show remarkable performance on a wide variety of tasks. Most LLMs split text into multi-character tokens and process them as atomic units without direct access to individual characters. This raises the question: To what extent can LLMs learn orthographic information? To answer this, we propose a new benchmark, CUTE, which features a collection of tasks designed to test the orthographic knowledge of LLMs. We evaluate popular LLMs on CUTE, finding that most of them seem to know the spelling of their tokens, yet fail to use this information effectively to manipulate text, calling into question how much of this knowledge is generalizable.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Academics Can Contribute to Domain-Specialized Language Models

Mark Dredze, Genta Indra Winata, Prabhansan Kambadur, Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, David S Rosenberg, Sebastian Gehrmann

Commercially available models dominate academic leaderboards. While impressive, this has concentrated research on creating and adapting general-purpose models to improve NLP leaderboard standings for large language models. However, leaderboards collect many individual tasks and general-purpose models often underperform in specialized domains; domain-specific or adapted models yield superior results. This focus on large general-purpose models excludes many academics and draws attention away from areas where they can make important contributions. We advocate for a renewed focus on developing and evaluating domain- and task-specific models, and highlight the unique role of academics in this endeavor.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Scaling Laws Across Model Architectures: A Comparative Analysis of Dense and MoE Models in Large Language Models

Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, Jingang Wang

The scaling of large language models (LLMs) is a critical research area for the efficiency and effectiveness of model training and deployment. Our work investigates the transferability and discrepancies of scaling laws between Dense Models and Mixture of Experts (MoE) models. Through a combination of theoretical analysis and extensive experiments, including consistent loss scaling, optimal batch size/learning rate scaling, and resource allocation strategies scaling, our findings reveal that the power-law scaling framework also applies to MoE Models, indicating that the fundamental principles governing the scaling behavior of these models are preserved, even though the architecture differs. Additionally, MoE Models demonstrate superior generalization, resulting in lower testing losses with the same training compute budget compared to Dense Models. These findings indicate the scaling consistency and transfer generalization capabilities of MoE Models, providing new insights for optimizing MoE Model training and deployment strategies.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Learning to Retrieve Iteratively for In-Context Learning

Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, Benjamin Van Durme

We introduce iterative retrieval, a novel framework that empowers retrievers to make iterative decisions through policy optimization. Finding an optimal portfolio of retrieved items is a combinatorial optimization problem, generally considered NP-hard. This approach provides a learned approximation to such a solution, meeting specific task requirements under a given family of large language models (LLMs). We propose a training procedure based on reinforcement learning, incorporating feedback from LLMs. We instantiate an iterative retriever for composing in-context learning (ICL) exemplars and apply it to various semantic parsing tasks that demand synthesized programs as outputs. By adding only 4M additional parameters for state encoding, we convert an off-the-shelf dense retriever into a stateful iterative retriever, outperforming previous methods in selecting ICL exemplars on semantic parsing datasets such as CalFlow, TreeDST, and MTOP. Additionally, the trained iterative retriever generalizes across different inference LLMs beyond the one used during training.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LIONS: An Empirically Optimized Approach to Align Language Models

Xiao Yu, Qingyang Wu, Yu Li, Zhou Yu

Alignment is a crucial step to enhance the instruction-following and conversational abilities of language models. Despite many recent works proposing new algorithms, datasets, and training pipelines, there is a lack of comprehensive studies measuring the impact of various design choices throughout the whole training process. We first conduct a rigorous analysis over a three-stage training pipeline consisting of supervised fine-tuning, offline preference learning, and online preference learning. We have found that using techniques like sequence packing, loss masking in SFT, increasing the preference dataset size in DPO, and online DPO training can significantly improve the performance of language models. We then train from Gemma-2b-base and LLaMA-3-8b-base, and find that our best models exceed the performance of the official instruct models tuned with closed-source data and algorithms. Our code and models can be found at <https://github.com/Columbia-NLP-Lab/LionAlignment>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Birdie: Advancing State Space Models with a Minimalist Architecture and Novel Pre-training Objectives

Sam Blouir, Jimmy T.H. Smith, Antonios Anastasopoulos, Amarda Shehu

Efficient state space models (SSMs), including linear recurrent neural networks and linear attention variants, have emerged as potential alternative language models to Transformers. While efficient, SSMs struggle with tasks requiring in-context retrieval, such as text copying and associative recall, limiting their usefulness in practical settings. Prior work on how to meet this challenge has focused on the internal model architecture and not investigated the role of the training procedure. This paper proposes a new training procedure that improve the performance of SSMs on retrieval-intensive tasks. This novel pre-training procedure combines a bidirectional processing of the input with dynamic mixtures of pre-training objectives to improve the utilization of the SSM's fixed-size state. Our experimental evaluations show that this procedure significantly improves performance on retrieval-intensive tasks that challenge current SSMs, such as phone book lookup, long paragraph question-answering, and infilling tasks. Our findings offer insights into a new direction to advance the training of SSMs to close the performance gap with Transformers.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Can Transformer Language Models Learn n -gram Language Models?

Anej Sveté, Nadav Borenstein, Mike Zhou, Ryan Cotterell

Much theoretical work has described the ability of transformers to represent formal languages. However, linking theoretical results to empirical performance is not straightforward due to the complex interplay between the architecture, the learning algorithm, and training data. To test whether theoretical lower bounds imply *learnability* of formal languages, we turn to recent work relating transformers to n -gram language models (LMs). We study transformers' ability to learn random n -gram LMs of two kinds: ones with arbitrary next-symbol probabilities and ones where those are defined with shared parameters. We find that classic estimation techniques for n -gram LMs such as add- λ smoothing outperform transformers on the former, while transformers perform better on the latter, outperforming methods specifically designed to learn n -gram LMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

StablePrompt : Automatic Prompt Tuning using Reinforcement Learning for Large Language Model

Minchan Kwon, Gaeun Kim, Jongsuik Kim, Haeil Lee, Junmo Kim

Finding appropriate prompts for the specific task has become an important issue as the usage of Large Language Models (LLM) have expanded. However, the variety of input-output formats complicate finding the prompts. Reinforcement Learning (RL) is a promising for prompt tuning due to its ability to incrementally produce better results through interaction with the environment. But its inherent training instability and environmental dependency make it difficult to use in practice. In this paper, we propose StablePrompt, a prompt tuning method based on RL. We formulate prompt tuning as RL problem between agent and target LLM, and introduce Adaptive Proximal Policy Optimization (APPO), an modified version of PPO for prompt tuning. APPO introduces an anchor model and updates it adaptively based on the training trajectory. Using this anchor model for the KL divergence term in PPO keeps the search space flexible and ensures training stability. We evaluate StablePrompt on various tasks, including text classification, question answering, and text generation. StablePrompt achieves State-of-The-Art performance across diverse tasks. We demonstrate that StablePrompt performs well across various types and sizes of LLMs. Furthermore, we present TTE-StablePrompt, an extension for generating input-dependent prompts. It outperforms StablePrompt in tasks that are hard to solve with a single prompt. This shows that StablePrompt is an extensible and stable RL framework for LLM.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

SCOI: Syntax-augmented Coverage-based In-context Example Selection for Machine Translation

Chenhang Tang, Zhixiang Wang, Yunfang Wu

In-context learning (ICL) greatly improves the performance of large language models (LLMs) on various down-stream tasks, where the improvement highly depends on the quality of demonstrations. In this work, we introduce syntactic knowledge to select better in-context examples for machine translation (MT). We propose a new strategy, namely Syntax-augmented COverage-based In-context example selection (SCOI), leveraging the deep syntactic structure beyond conventional word matching. Specifically, we measure the set-level syntactic coverage by computing the coverage of polynomial terms with the help of a simplified tree-to-polynomial algorithm, and lexical coverage using word overlap. Furthermore, we devise an alternate selection approach to combine both coverage measures, taking advantage of syntactic and lexical information. We conduct experiments with two multi-lingual LLMs on six translation directions. Empirical results show that our proposed SCOI obtains the highest average COMET score among all learning-free methods, indicating that combining syntactic and lexical coverage successfully helps to select better in-context examples for MT. Our code is available at <https://github.com/JamyDon/SCOI>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Turn Waste into Worth: Rectifying Top- k Router of MoE

Zhiyuan Meng, Qipeng Guo, Maohe Fei, Zhangye Yin, Yunhua Zhou, Linyang Li, Tianxiang Sun, Hang Yan, Dahua Lin, Xipeng Qiu

Sparse Mixtures of Experts (MoE) models are popular for training large language models due to their computational efficiency. However, the commonly used top- k routing mechanism suffers from redundancy computation and memory costs due to the unbalanced routing. Some experts are overflow, where the exceeding tokens are dropped. While some experts are empty, which are padded with zeros, negatively impacting model performance. To address the dropped tokens and padding, we propose the Rectify-Router, comprising the Intra-GPU Rectification and the Fill-in Rectification. The Intra-GPU Rectification handles dropped tokens, efficiently routing them to experts within the GPU where they are located to avoid inter-GPU communication. The Fill-in Rectification addresses padding by replacing padding tokens with the tokens that have high routing scores. Our experimental results demonstrate that the Intra-GPU Rectification and the Fill-in Rectification effectively handle dropped tokens and padding, respectively. Furthermore, the combination of them achieves superior performance, surpassing the accuracy of the vanilla top-1 router by 4.7%.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Null-Shot Prompting: Rethinking Prompting Large Language Models With Hallucination

Pittawat Takeekitworachai, Febri Abdullah, Ruck Thawonmas

This paper presents a series of investigations into an interesting phenomenon where we observe performance increases in large language models (LLMs) when providing a prompt that causes and exploits hallucination. We propose null-shot prompting, a counter-intuitive approach where we intentionally instruct LLMs to look at and utilize information from a null section. We investigate null-shot prompting on a wide range of tasks, including arithmetic reasoning, commonsense reasoning, and reading comprehension. We observe a substantial increase in performance in arithmetic reasoning tasks for various models, with up to a 44.62% increase compared to a baseline in one model. Therefore, we investigate deeper into this task by utilizing a more challenging mathematics problem-solving benchmark. We observe that LLMs benefit from hallucination in null-shot prompting in this task and discuss the mathematical topics that benefit the most from introducing hallucination in the prompt. We continue our investigation by evaluating hallucination detection abilities of the LLMs when using null-shot prompting. We find surprising results where hallucination in prompts can improve hallucination detection abilities of many LLMs. We also examine the effects of introducing both reasoning, which is known to mitigate hallucination, and hallucination simultaneously in the prompt and observe another surprising turn for the mathematics problem-solving benchmark with many performance improvements. We hope this paper will spark more interest, investigations, and discussions on how hallucination in prompts LLMs and even bolsters them in certain cases.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

How to Leverage Demonstration Data in Alignment for Large Language Model? A Self-Imitation Learning Perspective

Teng Xiao, Mingxiao Li, Yige Yuan, Huaiseng Zhu, Chao Cui, Vasant Honavar

This paper introduces a novel generalized self-imitation learning GSIL framework, which effectively and efficiently aligns large language models with offline demonstration data. We develop GSIL by deriving a surrogate objective of imitation learning with density ratio estimates, facilitating the use of self-generated data and optimizing the imitation learning objective with simple classification losses. GSIL eliminates the need for complex adversarial training in standard imitation learning, achieving lightweight and efficient fine-tuning for large language models. In addition, GSIL encompasses a family of offline losses parameterized by a general class of convex functions for density ratio estimation and enables a unified view for alignment with demonstration data. Extensive experiments show that GSIL consistently and significantly outperforms baselines in many challenging benchmarks, such as coding (HuannEval), mathematical reasoning (GSM8K) and instruction-following benchmark (MT-Bench). Code is public available at <https://github.com/tengxiao1/GSIL>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Large Language Models in the Clinic: A Comprehensive Benchmark

Fengling Liu, Zheng Li, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Hongjian Zhou, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, Bing Yin, David A. Clifton

The adoption of large language models (LLMs) to assist clinicians has attracted remarkable attention. Existing works mainly adopt the close-ended question-answering (QA) task with answer options for evaluation. However, many clinical decisions involve answering open-ended questions without pre-set options. To better understand LLMs in the clinic, we construct a benchmark ClinicBench. We first collect eleven existing datasets covering diverse clinical language generation, understanding, and reasoning tasks. Furthermore, we construct six novel datasets and clinical tasks that are complex but common in real-world practice, e.g., open-ended decision-making, long document processing, and emerging drug analysis. We conduct an extensive evaluation of twenty-two LLMs under both zero-shot and few-shot settings. Finally, we invite medical experts to evaluate the clinical usefulness of LLMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Re-RST: Reflection-Reinforced Self-Training for Language Agents

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, Nanyun Peng

Finetuning language agents with reasoning-action trajectories is effective, but obtaining these trajectories from human annotations or stronger models is costly and sometimes impractical. In this paper, we investigate the use of self-training in language agents, which can generate supervision from the agent itself, offering a promising alternative without relying on human or stronger model demonstrations. Self-training, however, requires high-quality model-generated samples, which are hard to obtain for challenging language agent tasks. To address this, we present Reflect-Reinforced Self-Training (Re-RST), which uses a *reflector* to refine low-quality generated samples during self-training. The reflector takes the agent's output and feedback from an external environment (e.g., unit test results in code generation) to produce improved samples. This technique enhances the quality of inferior samples and efficiently enriches the self-training dataset with higher-quality samples. We conduct extensive experiments on open-source language agents across tasks, including multi-hop question answering, sequential decision-making, code generation, visual question answering, and text-to-image generation. The results demonstrate the effectiveness of self-training and Re-RST in language agent tasks, with self-training improving baselines by 7.6% on HotpotQA and 28.4% on AlWorld, and Re-RST further boosting performance by 2.0% and 14.1%, respectively. Our studies also confirm the efficiency of using a reflector to generate high-quality samples for self-training. Moreover, we demonstrate a method to employ reflection during inference without ground-truth feedback, addressing the limitation of previous reflection work.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Small LLMs Are Weak Tool Learners: A Multi-LLM Agent

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, Fei Huang

Large Language Model (LLM) agents significantly extend the capabilities of standalone LLMs, empowering them to interact with external tools (e.g., APIs, functions) and complete various tasks in a self-directed fashion. The challenge of tool use demands that LLMs not only understand user queries and generate answers accurately but also excel in task planning, tool invocation, and result summarization. While traditional works focus on training a single LLM with all these capabilities, performance limitations become apparent, particularly with smaller models. To overcome these challenges, we propose a novel approach that decomposes the aforementioned capabilities into a planner, caller, and summarizer. Each component is implemented by a single LLM that focuses on a specific capability and collaborates with others to accomplish the task. This modular framework facilitates individual updates and the potential use of smaller LLMs for building each capability. To effectively train this framework, we introduce a two-stage training paradigm. First, we fine-tune a backbone LLM on the entire dataset without discriminating sub-tasks, providing the model with a comprehensive understanding of the task. Second, the fine-tuned LLM is used to instantiate the planner, caller, and summarizer respectively, which are continually fine-tuned on respective sub-tasks. Evaluation across various tool-use benchmarks illustrates that our proposed multi-LLM framework surpasses the traditional single-LLM approach, highlighting its efficacy and advantages in tool learning.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, Nanyun Peng

Model editing is a technique that edits the large language models (LLMs) with updated knowledge to alleviate hallucinations without resource-intensive retraining. While current model editing methods can effectively modify a model's behavior within a specific area of interest, they often overlook the potential unintended side effects on the general abilities of LLMs such as reasoning, natural language inference, and question answering. In this paper, we raise concern that model editing improvements on factuality may come at the cost of a significant degradation of the models' general abilities. We systematically analyze the side effects by evaluating four popular editing methods on three LLMs across eight representative tasks. Our extensive empirical experiments show that it is challenging for current editing methods to simultaneously improve factuality of LLMs and maintain their general abilities. Our analysis reveals that the side effects are caused by model editing altering the original model weights excessively, leading to overfitting to the edited facts. To mitigate this, a method named RECT is proposed to regularize the edit update weights by imposing constraints on their complexity based on the RElative Change in weightT. Evaluation results show that RECT can significantly mitigate the side effects of editing while still maintaining over 94% editing performance.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

T-FREE: Tokenizer-Free Generative LLMs via Sparse Representations for Memory-Efficient Embeddings

Björn Deiseroth, Manuel Brack, Samuel Weinbach, Patrick Schramowski, Kristian Kersting

Tokenizers are crucial for encoding information in Large Language Models, but their development has recently stagnated, and they contain inherent weaknesses. Major limitations include computational overhead, ineffective vocabulary use, and unnecessarily large embedding and head layers. Additionally, their performance is biased towards a reference corpus, leading to reduced effectiveness for underrepresented languages. To remedy these issues, we propose T-Free, which directly embeds words through sparse activation patterns over character triplets and does not require a reference corpus. T-Free inherently exploits morphological similarities and allows for strong compression of embedding

layers. In our exhaustive experimental evaluation, we achieve competitive downstream performance with a parameter reduction of more than 85% on these layers. Further, T-Free shows significant improvements in cross-lingual transfer learning.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Dynamic Rewarding with Prompt Optimization Enables Tuning-free Self-Alignment of Language Models

Somanshu Singla, Zhen Wang, Tianyang Liu, Abdullah Ashfaq, Zhiting Hu, Eric P. Xing

Aligning Large Language Models (LLMs) traditionally relies on complex and costly training processes like supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). To address the challenge of achieving alignment without these extensive tuning costs and expensive annotations, we present a novel, tuning-free approach for self-alignment called Dynamic Rewarding with Prompt Optimization (DRPO). Our approach enables self-alignment through a search-based prompt optimization framework, allowing the model to self-improve and generate optimized prompts without additional training or human supervision. The core of DRPO leverages a dynamic rewarding mechanism to identify and rectify model-specific alignment weaknesses, enabling LLMs to adapt quickly to various alignment challenges. Empirical evaluations on eight recent LLMs, including both open- and closed-source, reveal that DRPO significantly enhances alignment performance, enabling base models to outperform their SFT/RLHF-tuned counterparts. Moreover, DRPO's automatically optimized prompts surpass those curated by human experts, demonstrating its superior alignment capabilities. Our findings envision a highly cost-effective and adaptable solution for future alignment research to be further explored.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Improving Multi-Agent Debate with Sparse Communication Topology

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, Eugene Ie

Multi-agent debate has proven effective in improving large language models quality for reasoning and factuality tasks. While various role-playing strategies in multi-agent debates have been explored, in terms of the communication among agents, existing approaches adopt a brute force algorithm – each agent can communicate with all other agents. In this paper, we systematically investigate the effect of communication connectivity in multi-agent systems. Our experiments on GPT and Mistral models reveal that multi-agent debates leveraging sparse communication topology can achieve comparable or superior performance while significantly reducing computational costs. Furthermore, we extend the multi-agent debate framework to multi-modal reasoning and alignment labeling tasks, showcasing its broad applicability and effectiveness. Our findings underscore the importance of communication connectivity on enhancing the efficiency and effectiveness of the “society of minds” approach.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Detecting Machine-Generated Long-Form Content with Latent-Space Variables

Yufei Tian, Zeyu Pan, Nanyun Peng

The increasing capability of large language models (LLMs) to generate fluent long-form texts is presenting new challenges in distinguishing these outputs from those of humans. Existing zero-shot detectors that primarily focus on token-level distributions are vulnerable to real-world domain shift including different decoding strategies, variations in prompts, and attacks. We propose a more robust method that incorporates abstract elements—such as topic or event transitions—as key deciding factors, by training a latent-space model on sequences of events or topics derived from human-written texts. On three different domains, machine generations which are originally inseparable from humans' on the token level can be better distinguished with our latent-space model, leading to a 31% improvement over strong baselines such as DetectGPT. Our analysis further reveals that unlike humans, modern LLMs such as GPT-4 selecting event triggers and transitions differently, and inherent disparity regardless of the generation configurations adopted in real-time.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues

Makesh Narasimhan Sreedhar, Traian Rebedea, Shaona Ghosh, Jiaqi Zeng, Christopher Parisien

Recent advancements in instruction-tuning datasets have predominantly focused on specific tasks like mathematical or logical reasoning. There has been a notable gap in data designed for aligning language models to maintain topic relevance in conversations - a critical aspect for deploying chatbots to production. We introduce the CanTalkAboutThis dataset to help language models remain focused on the subject at hand during task-oriented interactions. It consists of synthetic dialogues on a wide range of conversation topics from different domains. These dialogues are interspersed with distractor turns that intentionally divert the chatbot from the predefined topic. Fine-tuning language models on this dataset helps make them resilient to deviating from the assigned role and improves their ability to maintain topical coherence compared to general-purpose instruction-tuned LLMs like gpt-4-turbo and Mixtral-Instruct. Additionally, preliminary observations suggest that training models on this dataset also enhance their performance on fine-grained instruction following tasks, including safety alignment.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LongAlign: A Recipe for Long Context Alignment of Large Language Models

Yushu Bai, Xin Lv, Jiajia Zhang, Yuzhe He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, Juanzi Li

Extending large language models to effectively handle long contexts requires instruction fine-tuning on input sequences of similar length. To address this, we present LongAlign—a recipe of the instruction data, training, and evaluation for long context alignment. First, we construct a long instruction-following dataset using Self-Instruct. To ensure the data diversity, it covers a broad range of tasks from various long context sources. Second, we adopt the packing and sorted batching strategies to speed up supervised fine-tuning on data with varied length distributions. Additionally, we develop a loss weighting method to balance the contribution to the loss across different sequences during packing training. Third, we introduce the LongBench-Chat benchmark for evaluating instruction-following capabilities on queries of 10k-100k in length. Experiments show that LongAlign outperforms existing recipes for LLMs in long context tasks by up to 30%, while also maintaining their proficiency in handling short, generic tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Adaptive BPE Tokenization for Enhanced Vocabulary Adaptation in Finetuning Pretrained Language Models

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, Niloy Ganguly

In this work, we show a fundamental limitation in vocabulary adaptation approaches that use Byte-Pair Encoding (BPE) tokenization scheme for fine-tuning pretrained language models (PLMs) to expert domains. Current approaches trivially append the target domain-specific vocabulary ($V_D\bar{O}M\bar{A}IN$) at the end of the PLM vocabulary. This approach leads to a lower priority score and causes sub-optimal tokenization in BPE that iteratively uses merge rules to tokenize a given text. To mitigate this issue, we propose ADAPT-BPE where the BPE tokenization initialization phase is modified to first perform the longest string matching on the added (target) vocabulary before tokenizing at the character level. We perform an extensive evaluation of ADAPT-BPE versus the standard BPE over various classification and summarization tasks; ADAPT-BPE improves by 3.57% (in terms of accuracy) and 1.87% (in terms of Rouge-L), respectively. ADAPT-BPE for MEDVOC works particularly well when reference summaries have high OOV concentration or are longer in length. We also conduct a human evaluation, revealing that ADAPT-BPE generates more relevant and more faithful summaries as compared to MEDVOC. We make our codebase publicly available at <https://github.com/gb-kgp/adaptbpe>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Enhancing Language Model Alignment: A Confidence-Based Approach to Label Smoothing

Baile Huang, Hiteshi Sharma, Yi Mao

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains. Within the training pipeline of LLMs, the Reinforcement Learning with Human Feedback (RLHF) phase is crucial for aligning LLMs with human preferences and values. Label smoothing, a technique that replaces hard labels with soft labels, emerges as promising techniques to enhance RLHF training. Despite the benefits, the choice of label smoothing parameters often relies on heuristic approaches and lack theoretical understanding. This paper addresses the challenge of selecting the label smoothing parameter in a principled manner. We introduce Confidence Aware Label Smoothing (CALS), a method that iteratively updates the label smoothing parameter based on preference labels and model forecasts. Our theoretical analysis characterizes the optimal label smoothing parameter, demonstrates its dependence on the confidence level, and reveals its influence on training dynamics and equilibrium. Empirical evaluations on state-of-the-art alignment tasks show that CALS achieves competitive performance, highlighting its potential for improving alignment.

Machine Learning for NLP 1

Nov 12 (Tue) 14:00-15:30 - Room: Jasmine

Nov 12 (Tue) 14:00-15:30 - Jasmine

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Inskip Lee, Osbert Bastani, Edgar Dobriban

Language Models (LMs) have shown promising performance in natural language generation. However, as LMs often generate incorrect or hallucinated responses, it is crucial to correctly quantify their uncertainty in responding to given inputs. In addition to verbalized confidence elicited via prompting, many uncertainty measures (e.g., semantic entropy and affinity-graph-based measures) have been proposed. However, these measures can differ greatly, and it is unclear how to compare them, partly because they take values over different ranges (e.g., $[0, \infty)$ or $[0, 1]$). In this work, we address this issue by developing a novel and practical framework, termed "Rank-Calibration*", to assess uncertainty and confidence measures for LMs. Our key tenet is that higher uncertainty (or lower confidence) should imply lower generation quality, on average. Rank-calibration quantifies deviations from this ideal relationship in a principled manner, without requiring ad hoc binary thresholding of the correctness score (e.g., ROUGE or METEOR). The broad applicability and the granular interpretability of our methods are demonstrated empirically.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Mitigating the Alignment Tax of RLHF

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, Tong Zhang

LLMs acquire a wide range of abilities during pre-training, but aligning LLMs under Reinforcement Learning with Human Feedback (RLHF) can lead to forgetting pretrained abilities, which is also known as the alignment tax. To investigate alignment tax, we conducted experiments with existing RLHF algorithms using OpenLLaMA-3B, which revealed a pronounced alignment tax in NLP tasks. Whereas, despite various techniques to mitigate forgetting, they are often at odds with the RLHF performance, leading to a trade-off between alignment performance and forgetting mitigation, leading to an alignment-forgetting trade-off. In this paper we show that model averaging, which simply interpolates between pre and post RLHF model weights, surprisingly achieves the most strongest alignment-forgetting Pareto front among a wide range of competing methods. To understand its effectiveness, we offer theoretical insights into model averaging, revealing that it enhances performance Pareto front by increasing feature diversity on the layers where tasks share overlapped feature spaces. Empirical evidence corroborates our analysis by showing the benefits of averaging low-level transformer layers. Building on the analysis and the observation that averaging different layers of the transformer leads to significantly different alignment-forgetting trade-offs, we propose Heterogeneous Model Averaging (HMA) to Heterogeneously find various combination ratios of model layers. HMA seeks to maximize the alignment performance while incurring minimal alignment tax. Moreover, we validate HMA's performance across a range of RLHF algorithms over OpenLLaMA-3B and further extend our findings to Mistral-7B which is evaluated by open-sourced preference model and GPT4. Code available here².

Nov 12 (Tue) 14:00-15:30 - Jasmine

Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, Maosong Sun

Alignment in artificial intelligence pursues the consistency between model responses and human preferences as well as values. In practice, the multifaceted nature of human preferences inadvertently introduces what is known as the "alignment tax"—a compromise where enhancements in alignment within one objective (e.g., harmlessness) can diminish performance in others (e.g., helpfulness). However, existing alignment techniques are mostly unidirectional, leading to suboptimal trade-offs and poor flexibility over various objectives. To navigate this challenge, we argue the prominence of grounding LLMs with evident preferences. We introduce controllable preference optimization (CPO), which explicitly specifies preference scores for different objectives, thereby guiding the model to generate responses that meet the requirements. Our experimental analysis reveals that the aligned models can provide responses that match various preferences among the "3H" (helpfulness, honesty, harmlessness) desiderata. Furthermore, by introducing diverse data and alignment goals, we surpass baseline methods in aligning with single objectives, hence mitigating the impact of the alignment tax and achieving improvements in multi-objective alignment.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Direct Multi-Turn Preference Optimization for Language Agents

Wentao Shi, Mengqi Yuan, Junkang Wu, Qijian Wang, Fuli Feng

Adapting Large Language Models (LLMs) for agent tasks is critical in developing language agents. Direct Preference Optimization (DPO) is a promising technique for this adaptation with the alleviation of compounding errors, offering a means to directly optimize Reinforcement Learning (RL) objectives. However, applying DPO to multi-turn tasks presents challenges due to the inability to cancel the partition function. Overcoming this obstacle involves making the partition function independent of the current state and addressing length disparities between preferred and dis-preferred trajectories. In this light, we replace the policy constraint with the state-action occupancy measure constraint in the RL objective and add length normalization to the Bradley-Terry model, yielding a novel loss function named DMPO for multi-turn agent tasks with theoretical explanations. Extensive experiments on three multi-turn agent task datasets confirm the effectiveness and superiority of the DMPO loss.

²<https://github.com/avalonstrel/Mitigating-the-Alignment-Tax-of-RLHF.git>

Nov 12 (Tue) 14:00-15:30 - Jasmine

MoDULA: Mixture of Domain-Specific and Universal LoRA for Multi-Task Learning

Yifei Ma, Zihan Liang, Huangyu Dai, Ben Chen, Dehong Gao, Zhuoran Ran, Zihan Wang, Linbo Jin, Wen Jiang, Guannan Zhang, Xiaoyan Cai, Libin Yang

The growing demand for larger-scale models in the development of Large Language Models (LLMs) poses challenges for efficient training within limited computational resources. Traditional fine-tuning methods often exhibit instability in multi-task learning and rely heavily on extensive training resources. Here, we propose MoDULA (Mixture of Domain-Specific and Universal LoRA), a novel Parameter Efficient Fine-Tuning (PEFT) Mixture-of-Expert (MoE) paradigm for improved fine-tuning and parameter efficiency in multi-task learning. The paradigm effectively improves the multi-task capability of the model by training universal experts, domain-specific experts, and routers separately. MoDULA-Res is a new method within the MoDULA paradigm, which maintains the model's general capability by connecting universal and task-specific experts through residual connections. The experimental results demonstrate that the overall performance of the MoDULA-Plan and MoDULA-Res methods surpasses that of existing fine-tuning methods on various LLMs. Notably, MoDULA-Res achieves more significant performance improvements in multiple tasks while reducing training costs by over 80% without losing general capability. Moreover, MoDULA displays flexible pluggability, allowing for the efficient addition of new tasks without retraining existing experts from scratch. This progressive training paradigm circumvents data balancing issues, enhancing training efficiency and model stability. Overall, MoDULA provides a scalable, cost-effective solution for fine-tuning LLMs with enhanced parameter efficiency and generalization capability.

Nov 12 (Tue) 14:00-15:30 - Jasmine

SEEKR: Selective Attention-Guided Knowledge Retention for Continual Learning of Large Language Models

Jinghan He, Hayun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, Jingqiao Wang

Continual learning (CL) is crucial for language models to dynamically adapt to the evolving real-world demands. To mitigate the catastrophic forgetting problem in CL, data replay has been proven a simple and effective strategy, and the subsequent data-replay-based distillation can further enhance the performance. However, existing methods fail to fully exploit the knowledge embedded in models from previous tasks, resulting in the need for a relatively large number of replay samples to achieve good results. In this work, we first explore and emphasize the importance of attention weights in knowledge retention, and then propose a SElective attEntion-guided Knowledge Retention method (SEEKR) for data-efficient replay-based continual learning of large language models (LLMs). Specifically, SEEKR performs attention distillation on the selected attention heads for finer-grained knowledge retention, where the proposed forgettability-based and task-sensitivity-based measures are used to identify the most valuable attention heads. Experimental results on two continual learning benchmarks for LLMs demonstrate the superiority of SEEKR over the existing methods on both performance and efficiency. Explicitly, SEEKR achieves comparable or even better performance with only 1/10 of the replayed data used by other methods, and reduces the proportion of replayed data to 1%. The code is available at <https://github.com/jinghanhe/SEEKR>.

Nov 12 (Tue) 14:00-15:30 - Jasmine

SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Difenderfer, Bhavya Kailkhura, Sijia Liu

Large Language Models (LLMs) have highlighted the necessity of effective unlearning mechanisms to comply with data regulations and ethical AI practices. LLM unlearning aims at removing undesired data influences and associated model capabilities without compromising utility beyond the scope of unlearning. While interest in studying LLM unlearning is growing, the impact of the optimizer choice for LLM unlearning remains unexplored. In this work, we shed light on the significance of optimizer selection in LLM unlearning for the first time, establishing a clear connection between second-order optimization and influence unlearning (a classical approach using influence functions to update the model for data influence removal). This insight propels us to develop a second-order optimization-based LLM unlearning framework, termed Second-Order UnLearning (SOUL), which extends the static, one-shot model update using influence unlearning to a dynamic, iterative unlearning process. Our extensive experiments show that SOUL consistently outperforms conventional first-order methods across various unlearning tasks, models, and metrics, indicating that second-order optimization offers an effective and broadly applicable solution for LLM unlearning.

Nov 12 (Tue) 14:00-15:30 - Jasmine

ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback

Ju-Seung Byun, Jiyun Chun, Jihyung Kil, Andrew Perrault

Large Multimodal Models (LMMs) excel at comprehending human instructions and demonstrate remarkable results across a broad spectrum of tasks. Reinforcement Learning from Human Feedback (RLHF) and AI Feedback (RLAIF) further refine LLMs by aligning them with specific preferences. These methods primarily use ranking-based feedback for entire generations. With advanced AI models (Teacher), such as GPT-4 and Claude 3 Opus, we can request various types of detailed feedback that are expensive for humans to provide. We propose a two-stage algorithm ARES that Alternates REinforcement Learning (RL) and Supervised Fine-Tuning (SFT). First, we ask the Teacher to score how much each sentence contributes to solving the problem in a Chain-of-Thought (CoT). This sentence-level feedback allows us to consider individual valuable segments, providing more granular rewards for the RL procedure. Second, we ask the Teacher to correct wrong reasoning after the RL stage. The RL procedure requires substantial hyperparameter tuning and often generates errors such as repetitive words and incomplete sentences. With correction feedback, we stabilize the RL fine-tuned model through SFT. We conduct experiments on the multi-modal datasets ScienceQA and A-OKVQA to demonstrate the effectiveness of our proposal. The ARES rationale achieves around 70% win rate compared to baseline models judged by GPT-4o. Additionally, we observe that the improved rationale reasoning leads to a 2.5% increase in inference answer accuracy on average for the multi-modal datasets.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Concept-skill Transferability-based Data Selection for Large Vision-Language Models

Jaewoo Lee, Boyang Li, Sung Ju Hwang

Instruction tuning, or supervised finetuning on extensive task-specific data, is necessary for Large Vision-Language Models (LVLMs) to generalize well across a broad range of vision-language (VL) tasks. However, training on large VL datasets can become prohibitively expensive. In this work, we introduce COINCIDE, an effective and scalable data selection technique that uses a small model as a reference model to select visual instruction tuning data for efficient finetuning of a target LVLM, focusing on diversity and transferability. Specifically, we cluster the training data using internal activations from a small model, which identifies VL concept-skill compositions needed by a target LVLM. We then sample data from these diverse clusters by considering their density and transferability, or the ability to transfer well to other concept-skill compositions. This approach ensures the diversity of these compositions, which is vital for LVLM generalization. Extensive experiments demonstrate that COINCIDE achieves superior performance and data selection efficiency against 8 strong baselines on two distinct datasets: LLava-1.5 and Vision-Flan. Using only 20% of the LLava-1.5 dataset, COINCIDE achieves performance comparable to the LVLM finetuned on the whole dataset, with 70% reduction of the wall-clock running time. On the Vision-Flan dataset, our method achieves superior results with only 16.7% of the training data.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Context-Aware Assistant Selection for Improved Inference Acceleration with Large Language Models

Jerry Huang, Prasanna Parthasarathi, Mehdi Rezagholizadeh, Sarah Chandar

Despite their widespread adoption, large language models (LLMs) remain prohibitive to use under resource constraints, with their ever growing sizes only increasing the barrier for use. One particular issue stems from the high latency associated with auto-regressive generation in LLMs, rendering the largest LLMs difficult to use without advanced computing infrastructure. Assisted decoding, where a smaller draft model guides a larger expert model's generation, has helped alleviate this concern, but remains dependent on alignment between the two models. Thus if the draft model is insufficiently capable on some domain of interest relative to the target model, performance can degrade. Alternatively, one can leverage multiple draft models to better cover the expertise of the target, but when multiple black-box draft models are available, selecting an assistant without details about its construction can be difficult. To better understand this decision making problem, we observe it as a contextual bandit, where a policy must choose a draft model based on a context. We show that even without prior knowledge of the draft models, creating an offline dataset from only outputs of independent draft/target models and training a policy over the alignment of these outputs can accelerate performance on multiple domains as long as an individual draft model is effective. We observe these results hold on various settings with multiple assisted decoding candidates, highlighting its flexibility and the advantageous role that such decision making can play.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Do Large Language Models Know How Much They Know?

Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, Sarah Chandar

Large Language Models (LLMs) have emerged as highly capable systems and are increasingly being integrated into various uses. Nevertheless, the rapid advancement in their deployment trails a comprehensive understanding of their internal mechanisms, as well as a delineation of their capabilities and limitations. A desired characteristic of an intelligent system is its ability to recognize the scope of its own knowledge. To investigate whether LLMs embody this attribute, we develop a benchmark that challenges these models to enumerate all information they possess on specific topics. This benchmark assesses whether the models recall excessive, insufficient, or the precise amount of required information, thereby indicating their awareness of how much they know about the given topic. Our findings reveal that the emergence of this property varies across different architectures and manifests at diverse rates. However, with sufficient scaling, all tested models are ultimately capable of performing this task. The insights gained from this research advance our understanding of LLMs, shedding light on their operational capabilities and contributing to the ongoing exploration of their intricate dynamics.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Language-to-Code Translation with a Single Labeled Example

Kai Bostrom, Harsh Jhamtani, Hao Fang, Sam Thomson, Richard Shin, Patrick Xia, Benjamin Van Durme, Jason Eisner, Jacob Andreas

Tools for translating natural language into code promise natural, open-ended interaction with databases, web APIs, and other software systems. However, this promise is complicated by the diversity and continual development of these systems, each with its own interface and distinct set of features. Building a new language-to-code translator, even starting with a large language model (LM), typically requires annotating a large set of natural language commands with their associated programs. In this paper, we describe ICIP (In-Context Inverse Programming), a method for bootstrapping a language-to-code system using mostly (or entirely) unlabeled programs written using a potentially unfamiliar (but human-readable) library or API. ICIP uses a pre-trained LM to assign candidate natural language descriptions to these programs, then iteratively refines the descriptions to ensure global consistency. Across nine different application domains from the Overnight and Spider benchmarks and text-davinci-003 and CodeLlama-7b-Instruct models, ICIP outperforms a number of prompting baselines. Indeed, in a nearly unsupervised setting with only a single annotated program and 100 unlabeled examples, it achieves up to 85% of the performance of a fully supervised system.

Nov 12 (Tue) 14:00-15:30 - Jasmine

WPO: Enhancing RLHF with Weighted Preference Optimization

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqi Song, Silei Xu, Chenguang Zhu

Reinforcement learning from human feedback (RLHF) is a promising solution to align large language models (LLMs) more closely with human values. Off-policy preference optimization, where the preference data is obtained from other models, is widely adopted due to its cost efficiency and scalability. However, off-policy preference optimization often suffers from a distributional gap between the policy used for data collection and the target policy, leading to suboptimal optimization. In this paper, we propose a novel strategy to mitigate this problem by simulating on-policy learning with off-policy preference data. Our Weighted Preference Optimization (WPO) method adapts off-policy data to resemble on-policy data more closely by reweighting preference pairs according to their probability under the current policy. This method not only addresses the distributional gap problem but also enhances the optimization process without incurring additional costs. We validate our method on instruction following benchmarks including Alpaca Eval 2 and MT-bench. WPO not only outperforms Direct Preference Optimization (DPO) by up to 5.6% on Alpaca Eval 2 but also establishes a remarkable length-controlled winning rate against GPT-4-turbo of 76.7% based on Gemma-2-9b-it. We release the code and models at <https://github.com/wzhouad/WPO>.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation

Zhen Lin, Shubhendu Trivedi, Jimeng Sun

The advent of large language models (LLMs) has dramatically advanced the state-of-the-art in numerous natural language generation tasks. For LLMs to be applied reliably, it is essential to have an accurate measure of their confidence. Currently, the most commonly used confidence score function is the likelihood of the generated sequence, which, however, conflates semantic and syntactic components. For instance, in question-answering (QA) tasks, an awkward phrasing of the correct answer might result in a lower probability prediction. Additionally, different tokens should be weighted differently depending on the context. In this work, we propose enhancing the predicted sequence probability by assigning different weights to various tokens using attention values elicited from the base LLM. By employing a validation set, we can identify the relevant attention heads, thereby significantly improving the reliability of the vanilla sequence probability confidence measure. We refer to this new score as the Contextualized Sequence Likelihood (CSL). CSL is easy to implement, fast to compute, and offers considerable potential for further improvement with task-specific prompts. Across several QA datasets and a diverse array of LLMs, CSL has demonstrated significantly higher reliability than state-of-the-art baselines in predicting generation quality, as measured by the AUROC or AUARC.

Nov 12 (Tue) 14:00-15:30 - Jasmine

SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers

Viktoriya A. Chekalina, Anna Rudenko, Gleb Mezentev, Aleksandr Mikhalev, Alexander Panchenko, Ivan Oseledets

The performance of Transformer models has been enhanced by increasing the number of parameters and the length of the processed text. Consequently, fine-tuning the entire model becomes a memory-intensive process. High-performance methods for parameter-efficient fine-tuning (PEFT) typically work with Attention blocks and often overlook MLP blocks, which contain about half of the model parameters. We

propose a new selective PEFT method, namely SparseGrad, that performs well on MLP blocks. We transfer layer gradients to a space where only about 1% of the layer's elements remain significant. By converting gradients into a sparse structure, we reduce the number of updated parameters. We apply SparseGrad to fine-tune BERT and RoBERTa for the NLU task and LLaMa-2 for the Question-Answering task. In these experiments, with identical memory requirements, our method outperforms LoRA and MeProp, robust popular state-of-the-art PEFT approaches.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Explicit Memory Learning with Expectation Maximization

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang

Large Language Models (LLMs) have revolutionized the landscape of natural language processing, demonstrating remarkable abilities across various complex tasks. However, their stateless nature limits the capability to retain information across interactions, hindering performance in scenarios requiring historical context recall. To mitigate this, current approaches primarily use explicit memory to allow LLMs to store useful information, which is accessible, readable, and interpretable. Nevertheless, explicit memory lacks the reliable learning mechanisms of implicit memory, which can be optimized end-to-end. To harness the benefits of both, we introduce EM², a novel framework enhancing explicit memory updates via the Expectation-Maximization (EM) algorithm. EM² treats memory as a latent variable, ensuring continual learning and improvement during updates. Experimental results on streaming inference tasks demonstrate that EM² significantly enhances performance across various backbones and memory strategies, providing a robust solution for advancing LLM memory management and enabling explicit memory to learn and improve similarly to implicit memory.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Semformer: Transformer Language Models with Semantic Planning

Yongjing Yin, Junran Ding, Kai Song, Yue Zhang

Next-token prediction serves as the dominant component in current neural language models. During the training phase, the model employs teacher forcing, which predicts tokens based on all preceding ground truth tokens. However, this approach has been found to create shortcuts, utilizing the revealed prefix to spuriously fit future tokens, potentially compromising the accuracy of the next-token predictor. In this paper, we introduce Semformer, a novel method of training a Transformer language model that explicitly models the semantic planning of response. Specifically, we incorporate a sequence of planning tokens into the prefix, guiding the planning token representations to predict the latent semantic representations of the response, which are induced by an autoencoder. In a minimal planning task (i.e., graph path-finding), our model exhibits near-perfect performance and effectively mitigates shortcut learning, a feat that standard training methods and baseline models have been unable to accomplish. Furthermore, we pretrain Semformer from scratch with 125M parameters, demonstrating its efficacy through measures of perplexity, in-context learning, and fine-tuning on summarization tasks.

Nov 12 (Tue) 14:00-15:30 - Jasmine

VerifyMatch: A Semi-Supervised Learning Paradigm for Natural Language Inference with Confidence-Aware MixUp

Seo Yeon Park, Cornelia Caragea

While natural language inference (NLI) has emerged as a prominent task for evaluating a model's capability to perform natural language understanding, creating large benchmarks for training deep learning models imposes a significant challenge since it requires extensive human annotations. To overcome this, we propose to construct pseudo-generated samples (premise-hypothesis pairs) using class-specific fine-tuned large language models (LLMs) thereby reducing the human effort and the costs in annotating large amounts of data. However, despite the impressive performance of LLMs, it is necessary to verify that the pseudo-generated labels are actually correct. Towards this goal, in this paper, we propose VerifyMatch, a semi-supervised learning (SSL) approach in which the LLM pseudo-labels guide the training of the SSL model and, at the same time, the SSL model acts as a verifier of the LLM-generated data. In our approach, we retain all pseudo-labeled samples, but to ensure unlabeled data quality, we further propose to use MixUp whenever the verifier does not agree with the LLM-generated label or when they both agree on the label but the verifier has a low confidence—lower than an adaptive confidence threshold. We achieve competitive accuracy compared to strong baselines for NLI datasets in low-resource settings.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Defending Jailbreak Prompts via In-Context Adversarial Game

Yujun Zhou, Yafei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, Xiangliang Zhang

Large Language Models (LLMs) demonstrate remarkable capabilities across diverse applications. However, concerns regarding their security, particularly the vulnerability to jailbreak attacks, persist. Drawing inspiration from adversarial training in deep learning and LLM agent learning processes, we introduce the In-Context Adversarial Game (ICAG) for defending against jailbreaks without the need for fine-tuning. ICAG leverages agent learning to conduct an adversarial game, aiming to dynamically extend knowledge to defend against jailbreaks. Unlike traditional methods that rely on static datasets, ICAG employs an iterative process to enhance both the defense and attack agents. This continuous improvement process strengthens defenses against newly generated jailbreak prompts. Our empirical studies affirm ICAG's efficacy, where LLMs safeguarded by ICAG exhibit significantly reduced jailbreak success rates across various attack scenarios. Moreover, ICAG demonstrates remarkable transferability to other LLMs, indicating its potential as a versatile defense mechanism. The code is available at <https://github.com/YujunZhou/In-Context-Adversarial-Game>.

Nov 12 (Tue) 14:00-15:30 - Jasmine

Towards Aligning Language Models with Textual Feedback

Saïc Abadal Lloret, Shehzad Dhuliawala, Keerthiram Murugesan, Mrinmaya Sachan

We present ALT (ALignment with Textual feedback), an approach that aligns language models with user preferences expressed in text. We argue that text offers greater expressiveness, enabling users to provide richer feedback than simple comparative preferences and this richer feedback can lead to more efficient and effective alignment. ALT aligns the model by conditioning its generation on the textual feedback. Our method relies solely on language modeling techniques and requires minimal hyper-parameter tuning, though it still presents the main benefit of RL-based algorithms and can effectively learn from textual feedback. We explore the efficacy and efficiency of textual feedback across different tasks such as toxicity reduction, summarization, and dialog response. We find that ALT outperforms PPO for the task of toxicity reduction while being able to match its performance on summarization with only 20% of the samples. We also explore how ALT can be used with feedback provided by an existing LLM.

Nov 12 (Tue) 14:00-15:30 - Jasmine

GottBERT: a pure German Language Model

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patrik Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, Martin Boeker
Pre-trained language models have significantly advanced natural language processing (NLP), especially with the introduction of BERT and its optimized version, RoBERTa. While initial research focused on English, single-language models can be advantageous compared to multilingual ones in terms of pre-training effort, overall resource efficiency or downstream task performance. Despite the growing popularity of

prompt-based LLMs, more compute-efficient BERT-like models remain highly relevant. In this work, we present the first German single-language RoBERTa model, GottBERT, pre-trained exclusively on the German portion of the OSCAR dataset. Additionally, we investigated the impact of filtering the OSCAR corpus. GottBERT was pre-trained using fairseq and standard hyperparameters. We evaluated its performance on two Named Entity Recognition (NER) tasks (CoNLL 2003 and GermEval 2014) and three text classification tasks (GermEval 2018 fine and coarse, and 10kGNAD) against existing German BERT models and two multilingual models. Performance was measured using the F_1 score and accuracy. The GottBERT base and large models showed competitive performance, with GottBERT leading among the base models in 4 of 6 tasks. Contrary to our expectation, the applied filtering did not significantly affect the results. To support the German NLP research community, we are releasing the GottBERT models under the MIT license.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

ALVIN: Active Learning Via INterpolation

Michalis Korakakis, Andreas Vlachos

Active Learning aims to minimize annotation effort by selecting the most useful instances from a pool of unlabeled data. However, typical active learning methods overlook the presence of distinct example groups within a class, whose prevalence may vary, e.g., in occupation classification datasets certain demographics are disproportionately represented in specific classes. This oversight causes models to rely on shortcuts for predictions, i.e., spurious correlations between input attributes and labels occurring in well-represented groups. To address this issue, we propose Active Learning Via INterpolation (ALVIN), which conducts intra-class interpolations between examples from under-represented and well-represented groups to create anchors, i.e., artificial points situated between the example groups in the representation space. By selecting instances close to the anchors for annotation, ALVIN identifies informative examples exposing the model to regions of the representation space that counteract the influence of shortcuts. Crucially, since the model considers these examples to be of high certainty, they are likely to be ignored by typical active learning methods. Experimental results on six datasets encompassing sentiment analysis, natural language inference, and paraphrase detection demonstrate that ALVIN outperforms state-of-the-art active learning methods in both in-distribution and out-of-distribution generalization.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Filtered Direct Preference Optimization

Tetsuro Morimura, Mitsuaki Sakamoto, Yuu Jinrai, Kenshi Abe, Kaito Ariu

Reinforcement learning from human feedback (RLHF) plays a crucial role in aligning language models with human preferences. While the significance of dataset quality is generally recognized, explicit investigations into its impact within the RLHF framework, to our knowledge, have been limited. This paper addresses the issue of text quality within the preference dataset by focusing on direct preference optimization (DPO), an increasingly adopted reward-model-free RLHF method. We confirm that text quality significantly influences the performance of models optimized with DPO more than those optimized with reward-model-based RLHF. Building on this new insight, we propose an extension of DPO, termed filtered direct preference optimization (fDPO). fDPO uses a trained reward model to monitor the quality of texts within the preference dataset during DPO training. Samples of lower quality are discarded based on comparisons with texts generated by the model being optimized, resulting in a more accurate dataset. Experimental results demonstrate that fDPO enhances the final model performance. Our code is available at <https://github.com/CyberAgentAILab/fILTERED-dpo>.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

SQFT: Low-cost Model Adaptation in Low-precision Sparse Foundation Models

Juan Pablo Munoz, Jinjie Yuan, Nilesh Jain

Large pre-trained models (LPMs), such as large language models, have become ubiquitous and are employed in many applications. These models are often adapted to a desired domain or downstream task through a fine-tuning stage. This paper proposes SQFT, an end-to-end solution for low-precision sparse parameter-efficient fine-tuning of LPMs, allowing for effective model manipulation in resource-constrained environments. Additionally, an innovative strategy enables the merging of sparse weights with low-rank adapters without losing sparsity and accuracy, overcoming the limitations of previous approaches. SQFT also addresses the challenge of having quantized weights and adapters with different numerical precisions, enabling merging in the desired numerical format without sacrificing accuracy. Multiple adaptation scenarios, models, and comprehensive sparsity levels demonstrate the effectiveness of SQFT.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Regression (and Scoring) Aware Inference with LLMs

Mihai Lukasi, Harikrishna Narasimhan, Aditya Krishna Menon, Felix Yu, Sanjiv Kumar

Large language models (LLMs) have shown strong results on a range of applications, including regression and scoring tasks. Typically, one obtains outputs from an LLM via autoregressive sampling from the model's output distribution. We show that this inference strategy can be sub-optimal for common regression and scoring evaluation metrics. As a remedy, we build on prior work on Minimum Bayes Risk decoding and propose alternate inference strategies that estimate the Bayes-optimal solution for regression and scoring metrics in closed-form from sampled responses. We show that our proposal significantly improves over baselines across datasets and models.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

Reconfounding LLMs from the Grouping Loss Perspective

Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, Gael Varoquaux

Large Language Models (LLMs), such as GPT and LLaMA, are susceptible to generating hallucinated answers in a confident tone. While previous efforts to elicit and calibrate confidence scores have shown some success, they often overlook biases towards certain groups, such as specific nationalities. Existing calibration methods typically focus on average performance, failing to address this disparity. In our study, we demonstrate that the concept of grouping loss is an effective metric for understanding and correcting the heterogeneity in confidence levels. We introduce a novel evaluation dataset, derived from a knowledge base, specifically designed to assess the confidence scores of LLM responses across different groups. Our experimental results highlight significant variations in confidence, which are accurately captured by grouping loss. To tackle this issue, we propose a new method to calibrate the confidence scores of LLMs by considering different groups, a process we term *reconfounding*. Our findings indicate that this approach effectively mitigates biases against minority groups, contributing to the development of fairer LLMs.

Nov 12 (Tue) 14:00-15:30 - *Jasmine*

LaRS: Latent Reasoning Skills for Chain-of-Thought Reasoning

Zifan Xu, Haozhu Wang, Dmitry Bespalov, Xian Wu, Peter Stone, Yanjun Qi

Chain-of-thought (CoT) prompting is a popular in-context learning (ICL) approach for large language models (LLMs), especially when tackling complex reasoning tasks. Traditional ICL approaches construct prompts using examples that contain questions similar to the input question. However, CoT prompting, which includes crucial intermediate reasoning steps (rationales) within its examples, necessitates selecting examples based on these rationales rather than the questions themselves. Existing methods require human experts or pre-trained LLMs to describe the skill, a high-level abstraction of rationales, to guide the selection. These methods, however, are often costly and difficult to scale. Instead, this paper introduces a new approach named Latent Reasoning Skills (LaRS) that employs unsupervised learning to create a latent

space representation of rationales, with a latent variable called a reasoning skill. Concurrently, LaRS learns a reasoning policy to determine the required reasoning skill for a given question. Then the ICL examples are selected by aligning the reasoning skills between past examples and the question. This approach is theoretically grounded and compute-efficient, eliminating the need for auxiliary LLM inference or manual prompt design. Empirical results demonstrate that LaRS consistently outperforms SOTA skill-based selection methods, processing example banks four times faster, reducing LLM inferences during the selection stage by half, and showing greater robustness to sub-optimal example banks.

Nov 12 (Tue) 14:00-15:30 - Jasmine

CoBa: Convergence Balancer for Multitask Finetuning of Large Language Models

Zi Gong, Hang Yu, Cong Liao, Bingchang Liu, Chaoyu Chen, Jianguo Li

Multi-task learning (MTL) benefits the fine-tuning of large language models (LLMs) by providing a single model with improved performance and generalization ability across tasks, presenting a resource-efficient alternative to developing separate models for each task. Yet, existing MTL strategies for LLMs often fall short by either being computationally intensive or failing to ensure simultaneous task convergence. This paper presents CoBa, a new MTL approach designed to effectively manage task convergence balance with minimal computational overhead. Utilizing Relative Convergence Scores (RCS), Absolute Convergence Scores (ACS), and a Divergence Factor (DF), CoBa dynamically adjust task weights during the training process, ensuring that the validation loss of all tasks progress towards convergence at an even pace while mitigating the issue of individual task divergence. The results of our experiments involving three disparate datasets underscore that this approach not only fosters equilibrium in task improvement but enhances the LLMs' performance by up to 13% relative to the second-best baselines. Code is open-sourced at <https://github.com/codefuse-ai/MFTCoder>.

Multilinguality and Language Diversity 1

Nov 12 (Tue) 14:00-15:30 - Room: Riverfront Hall

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Zero-shot Cross-Lingual Transfer for Synthetic Data Generation in Grammatical Error Detection

Gaëtan Lopez Latouche, Marc-André Carboneau, Benjamin Swanson

Grammatical Error Detection (GED) methods rely heavily on human annotated error corpora. However, these annotations are unavailable in many low-resource languages. In this paper, we investigate GED in this context. Leveraging the zero-shot cross-lingual transfer capabilities of multilingual pre-trained language models, we train a model using data from a diverse set of languages to generate synthetic errors in other languages. These synthetic error corpora are then used to train a GED model. Specifically we propose a two-stage fine-tuning pipeline where the GED model is first fine-tuned on multilingual synthetic data from target languages followed by fine-tuning on human-annotated GED corpora from source languages. This approach outperforms current state-of-the-art annotation-free GED methods. We also analyse the errors produced by our method and other strong baselines, finding that our approach produces errors that are more diverse and more similar to human errors.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Cross-lingual Transfer for Automatic Question Generation by Learning Interrogative Structures in Target Languages

Seonjeong Hwang, Yunsu Kim, Gary Lee

Automatic question generation (QG) serves a wide range of purposes, such as augmenting question-answering (QA) corpora, enhancing chatbot systems, and developing educational materials. Despite its importance, most existing datasets predominantly focus on English, resulting in a considerable gap in data availability for other languages. Cross-lingual transfer for QG (XLT-QG) addresses this limitation by allowing models trained on high-resource language datasets to generate questions in low-resource languages. In this paper, we propose a simple and efficient XLT-QG method that operates without the need for monolingual, parallel, or labeled data in the target language, utilizing a small language model. Our model, trained solely on English QA datasets, learns interrogative structures from a limited set of question exemplars, which are then applied to generate questions in the target language. Experimental results show that our method outperforms several XLT-QG baselines and achieves performance comparable to GPT-3.5-turbo across different languages. Additionally, the synthetic data generated by our model proves beneficial for training multilingual QA models. With significantly fewer parameters than large language models and without requiring additional training for target languages, our approach offers an effective solution for QG and QA tasks across various languages.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, Ben Bergen

Multilingual language models are widely used to extend NLP systems to low-resource languages. However, concrete evidence for the effects of multilinguality on language modeling performance in individual languages remains scarce. Here, we pre-train over 10,000 monolingual and multilingual language models for over 250 languages, including multiple language families that are under-studied in NLP. We assess how language modeling performance in each language varies as a function of (1) monolingual dataset size, (2) added multilingual dataset size, (3) linguistic similarity of the added languages, and (4) model size (up to 45M parameters). We find that in moderation, adding multilingual data improves low-resource language modeling performance, similar to increasing low-resource dataset sizes by up to 33%. Improvements depend on the syntactic similarity of the added multilingual data, with marginal additional effects of vocabulary overlap. However, high-resource languages consistently perform worse in multilingual pre-training scenarios. As dataset sizes increase, adding multilingual data begins to hurt performance for both low-resource and high-resource languages, likely due to limited model capacity (the "curse of multilinguality"). These results suggest that massively multilingual pre-training may not be optimal for any languages involved, but that more targeted models can significantly improve performance.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

An Analysis of Multilingual FActScore

Vu Trong Kim, Michael Krumdick, Varshini Reddy, Franck Dernoncourt, Viet Dac Lai

FActScore has gained popularity as a metric to estimate the factuality of long-form texts generated by Large Language Models (LLMs) in English. However, there has not been any work in studying the behavior of FActScore in other languages. This paper studies the limitations of each component in the four-component pipeline of FActScore in the multilingual setting. We introduce a new dataset for FActScore on texts generated by strong multilingual LLMs. Our evaluation shows that LLMs exhibit distinct behaviors in both fact extraction and fact scoring tasks. No LLM produces consistent and reliable FActScore across languages of varying levels of resources. We also find that the knowledge source plays an important role in the quality of the estimated FActScore. Using Wikipedia as the knowledge source may hinder the true FActScore of long-form text due to its limited coverage in medium- and low-resource languages. We also incorporate 3 mitigations to our knowledge source that ultimately improve FActScore estimation across all languages.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Voices Unheard: NLP Resources and Models for Yorùbá Regional Dialects

Orevaoghene Aha, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, Yulia Tsvetkov
Yorùbá—an African language with roughly 47 million speakers—encompasses a continuum with several dialects. Recent efforts to develop NLP technologies for African languages have focused on their standard dialects, resulting in disparities for dialects and varieties for which there are little to no resources or tools. We take steps towards bridging this gap by introducing a new high-quality parallel text and speech corpus; YORULECT across three domains and four regional yoruba dialects. To develop this corpus, we engaged native speakers, traveling to communities where these dialects are spoken, to collect text and speech data. Using our newly created corpus, we conducted extensive experiments on (text) machine translation, automatic speech recognition, and speech-to-text translation. Our results reveal substantial performance disparities between standard yoruba and the other dialects across all tasks. However, we also show that with dialect-adaptive finetuning, we are able to narrow this gap. We believe our dataset and experimental analysis will contribute greatly to developing NLP tools for Yorùbá and its dialects, and potentially for other African languages, by improving our understanding of existing challenges and offering a high-quality dataset for further development. We will release YORULECT dataset and models publicly under an open license.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Using Language Models to Disambiguate Lexical Choices in Translation

Josh Barua, Sanjay Subramanian, Kavv Yin, Alane Suhr

In translation, a concept represented by a single word in a source language can have multiple variations in a target language. The task of lexical selection requires using context to identify which variation is most appropriate for a source text. We work with native speakers of nine languages to create DTaILS, a dataset of 1,377 sentence pairs that exhibit cross-lingual concept variation when translating from English. We evaluate recent LLMs and neural machine translation systems on DTaILS, with the best-performing model, GPT-4, achieving from 67 to 85% accuracy across languages. Finally, we use language models to generate English rules describing target-language concept variations. Providing weaker models with high-quality lexical rules improves accuracy substantially, in some cases reaching or outperforming GPT-4.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages

Holly Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santos, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onna P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Railey Montalan, Ryan Ignatius Hadiwijaya, Joainto Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Dianadar, Yuce GAO, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaize Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adhya Ryanda, Sonny Lazarudi Hernawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johannes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirkhodjai Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muenninghoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjihi, Peerat Limkanchitwatt, Alham Fikri Ajie, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, Samuel Cahyawijaya

Southeast Asia (SEA) is a region rich in linguistic diversity and cultural variety, with over 1,300 indigenous languages and a population of 671 million people. However, prevailing AI models suffer from a significant lack of representation of texts, images, and audio datasets from SEA, compromising the quality of AI models for SEA languages. Evaluating models for SEA languages is challenging due to the scarcity of high-quality datasets, compounded by the dominance of English training data, raising concerns about potential cultural misrepresentation. To address these challenges, through a collaborative movement, we introduce SEACrowd, a comprehensive resource center that fills the resource gap by providing standardized corpora in nearly 1,000 SEA languages across three modalities. Through our SEACrowd benchmarks, we assess the quality of AI models on 36 indigenous languages across 13 tasks, offering valuable insights into the current AI landscape in SEA. Furthermore, we propose strategies to facilitate greater AI advancements, maximizing potential utility and resource equity for the future of AI in Southeast Asia.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Concept Space Alignment in Multilingual LLMs

Qiwet Peng, Anders Søgaard

Multilingual large language models (LLMs) seem to generalize somewhat across languages. We hypothesize this is a result of implicit vector space alignment. Evaluating such alignment, we see that larger models exhibit very high-quality linear alignments between corresponding concepts in different languages. Our experiments show that multilingual LLMs suffer from two familiar weaknesses: generalization works best for languages with similar typology, and for abstract concepts. For some models, e.g., the Llama-2 family of models, prompt-based embeddings align better than word embeddings, but the projections are less linear – an observation that holds across almost all model families, indicating that some of the implicitly learned alignments are broken somewhat by prompt-based methods.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Methods of Automatic Matrix Language Determination for Code-Switched Speech

Olga Lakovenko, Thomas Hain

Code-switching (CS) is the process of speakers interchanging between two or more languages which in the modern world becomes increasingly common. In order to better describe CS speech the Matrix Language Frame (MLF) theory introduces the concept of a Matrix Language, which is the language that provides the grammatical structure for a CS utterance. In this work the MLF theory was used to develop systems for Matrix Language Identity (MLID) determination. The MLID of English/Mandarin and English/Spanish CS text and speech was compared to acoustic language identity (LID), which is a typical way to identify a language in monolingual utterances. MLID predictors from audio show higher correlation with the textual principles than LID in all cases while also outperforming LID in an MLID recognition task based on F1 macro (60%) and correlation score (0.38). This novel approach has identified that non-English languages (Mandarin and Spanish) are preferred over the English language as the ML contrary to the monolingual choice of LID.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Verba volant, scripta volant? Don't worry! There are computational solutions for protoword reconstruction

Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Ioan-Bogdan Jordache, Teodor-George Marchitan, Simona Georgescu, Laurentiu Zeica

We introduce a new database of cognate words and etymons for the five main Romance languages, the most comprehensive one to date. We propose a strong benchmark for the automatic reconstruction of protowords for Romance languages, by applying a set of machine learning models and features on these data. The best results reach 90% accuracy in predicting the protoword of a given cognate set, surpassing existing state-of-the-art results for this task and showing that computational methods can be very useful in assisting linguists with protoword reconstruction.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Understanding and Mitigating Language Confusion in LLMs

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, Sebastian Ruder

We investigate a surprising limitation of LLMs: their inability to consistently generate text in a user's desired language. We create the Language Confusion Benchmark (LCB) to evaluate such failures, covering 15 typologically diverse languages with existing and newly-created English and multilingual prompts. We evaluate a range of LLMs on monolingual and cross-lingual generation reflecting practical use cases, finding that Llama Instruct and Mistral models exhibit high degrees of language confusion and even the strongest models fail to consistently respond in the correct language. We observe that base and English-centric instruct models are more prone to language confusion, which is aggravated by complex prompts and high sampling temperatures. We find that language confusion can be partially mitigated via few-shot prompting, multilingual SFT and preference tuning. We release our language confusion benchmark, which serves as a first layer of efficient, scalable multilingual evaluation.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

An Empirical Study of Multilingual Reasoning Distillation for Question Answering

Patomporn Payoungkhamdee, Peerat Limkachotiwat, Jinheon Baek, Potsavee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuvanich, Sarana Nutanong

Reasoning is one crucial capability in Large Language Models (LLMs), allowing them to perform complex tasks such as solving math problems and multi-step planning. Whilst reasoning capability can emerge in larger models, smaller ones usually have to rely on distillation to transfer this capability from a larger model. However, recent efforts to distill reasoning capabilities have focused mainly on English, leaving multilingual distillation underexplored. To address this gap, this paper examines existing English reasoning distillation methods that utilize a variety of positive rationales in multilingual settings and proposes d-CoT-nR, a novel approach that incorporates incorrect rationales as additional guidance. Empirical results from multilingual high-school examinations show that d-CoT-nR significantly surpasses the baseline, improving accuracy in unseen languages and correctness in step-by-step reasoning.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models?

Pinzhen Chen, Simon Yu, Zhicheng Guo, Barry Haddow

Multilingual large language models are designed, claimed, and expected to cater to speakers of varied languages. We hypothesise that the current practices of fine-tuning and evaluating these models may not perfectly align with this objective owing to a heavy reliance on translation, which cannot cover language-specific knowledge but can introduce translation defects. It remains unknown whether the nature of the instruction data has an impact on the model output; conversely, it is questionable whether translated test sets can capture such nuances. Due to the often coupled practices of using translated data in both stages, such imperfections could have been overlooked. This work investigates these issues using controlled native or translated data during the instruction tuning and evaluation stages. We show that native or generation benchmarks reveal a notable difference between native and translated instruction data especially when model performance is high, whereas other types of test sets cannot. The comparison between round-trip and single-pass translations reflects the importance of knowledge from language-native resources. Finally, we demonstrate that regularization is beneficial to bridging this gap on structured but not generative tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Unraveling Babel: Exploring Multilingual Activation Patterns of LLMs and Their Applications

Weize Liu, Yinyang Xu, Hongxia Xu, Jintai Chen, Xuming Hu, Jian Wu

Recently, large language models (LLMs) have achieved tremendous breakthroughs in the field of NLP, but still lack understanding of their internal neuron activities when processing different languages. We designed a method to convert dense LLMs into fine-grained MoE architectures, and then visually studied the multilingual activation patterns of LLMs through expert activation frequency heatmaps. Through comprehensive experiments on different model families, different model sizes, and different variants, we analyzed the similarities and differences in the internal neuron activation patterns of LLMs when processing different languages. Specifically, we investigated the distribution of high-frequency activated experts, multilingual shared experts, whether multilingual activation patterns are related to language families, and the impact of instruction tuning on activation patterns. We further explored leveraging the discovered differences in expert activation frequencies to guide sparse activation and pruning. Experimental results demonstrated that our method significantly outperformed random expert pruning and even exceeded the performance of unpruned models in some languages. Additionally, we found that configuring different pruning rates for different layers based on activation level differences could achieve better results. Our findings reveal the multilingual processing mechanisms within LLMs and utilize these insights to offer new perspectives for applications such as sparse activation and model pruning.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm

Aakanksha, Arash Ahmadian, Beyza Ermiş, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker

A key concern with the concept of “alignment” is the implicit question of “alignment to what?”. AI systems are increasingly used across the world, yet safety alignment is often focused on homogeneous monolingual settings. Additionally, preference training and safety measures often overfit to harms common in Western-centric datasets. Here, we explore the viability of different alignment approaches when balancing dual objectives: addressing and optimizing for a non-homogeneous set of languages and cultural preferences while minimizing both global and local harms. We collect the first human annotated red teaming prompts in different languages, distinguishing between global and local harm, which serve as a laboratory to understand the reliability of alignment techniques when faced with preference distributions that are non-stationary across geographies and languages. While this setting is seldom covered by the literature to date, which primarily centers on English harm mitigation, it captures real-world interactions with AI systems around the world. We establish a new precedent for state-of-the-art alignment techniques across 6 languages with minimal degradation in general performance. Our work provides important insights into cross-lingual transfer and novel optimization approaches to safeguard AI systems designed to serve global populations.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

TL-CL: Task And Language Incremental Continual Learning

Shrey Satapara, P. K. SriJith

This paper introduces and investigates the problem of Task and Language Incremental Continual Learning (TLCL), wherein a multilingual model is systematically updated to accommodate new tasks in previously learned languages or new languages for established tasks. This significant yet previously unexplored area holds substantial practical relevance as it mirrors the dynamic requirements of real-world applications. We benchmark a representative set of continual learning (CL) algorithms for TLCL. Furthermore, we propose Task and Language-Specific Adapters (TLSA), an adapter-based parameter-efficient fine-tuning strategy. TLSA facilitates cross-lingual and cross-task transfer and outperforms other parameter-efficient fine-tuning techniques. Crucially, TLSA reduces parameter growth stemming from saving adapters to linear complexity from polynomial complexity as it was with parameter isolation-based adapter tuning. We conducted experiments on several NLP tasks arising across several languages. We observed that TLSA outperforms all other parameter-efficient approaches without requiring access to historical data for replay.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Empowering Multi-step Reasoning across Languages via Program-Aided Language Models

Leonardo Ranaldi, Giulia Pucci

In-context learning methods are popular inference strategies where Large Language Models (LLMs) are elicited to solve a task using provided demonstrations without parameter updates. Among these approaches are the reasoning methods, best exemplified by Chain-of-Thought (CoT) and Program-Aided Language Models (PAL), which elicit LLMs to generate reasoning paths, thus promoting accuracy and attracting increasing attention. However, despite the success of these methods, the ability to deliver multi-step reasoning remains limited to a single language, making it challenging to generalize to other languages and hindering global development. In this work, we propose Cross-lingual Program-Aided Language Models (CrossPAL), a method for aligning reasoning programs across languages. In particular, our method delivers programs as intermediate reasoning steps in different languages through a double-step cross-lingual prompting mechanism inspired by the Program-Aided approach. In addition, we introduce Self-consistent CrossPAL (SCrossPAL) to ensemble different reasoning paths across languages. Our experimental evaluations show that our method significantly outperforms existing prompting methods, reducing the number of interactions and achieving state-of-the-art performance.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text

Michael Ginn, Linda Tjouta, Taigi He, Enora Rice, Graham Neubig, Alexis Palmer, Lori Levin

Language documentation projects often involve the creation of annotated text in a format such as interlinear glossed text (IGT), which captures fine-grained morphosyntactic analyses in a morpheme-by-morpheme format. However, there are few existing resources providing large amounts of standardized, easily accessible IGT data, limiting their applicability to linguistic research, and making it difficult to use such data in NLP modeling. We compile the largest existing corpus of IGT data from a variety of sources, covering over 450k examples across 1.8k languages, to enable research on crosslingual transfer and IGT generation. We normalize much of our data to follow a standard set of labels across languages. Furthermore, we explore the task of automatically generating IGT in order to aid documentation projects. As many languages lack sufficient monolingual data, we pretrain a large multilingual model on our corpus. We demonstrate the utility of this model by finetuning it on monolingual corpora, outperforming SOTA models by up to 6.6%. Our pretrained model and dataset are available on Hugging Face: <https://huggingface.co/collections/lecslab/glosslm-66da150854209e910113dd87>

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

1+2>2: Can Large Language Models Serve as Cross-Lingual Knowledge Aggregators?

Yue Huang, Cherrui Fan, Yuan Li, Siyan Wu, Tianyi Zhou, Xiangliang Zhang, Lichao Sun

Large Language Models (LLMs) have garnered significant attention due to their remarkable ability to process information across various languages. Despite their capabilities, they exhibit inconsistencies in handling identical queries in different languages, presenting challenges for further advancement. This paper introduces a method to enhance the multilingual performance of LLMs by aggregating knowledge from diverse languages. This approach incorporates a low-resource knowledge detector specific to a language, a strategic language selection process, and mechanisms for answer replacement and integration. Our extensive experiments demonstrate notable performance improvements, particularly in reducing the performance disparity across languages. An ablation study confirms that each component of our method significantly contributes to these enhancements. This research highlights the inherent potential of LLMs to harmonize multilingual capabilities and offers valuable insights for further exploration.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Cross-lingual Back-Parsing: Utterance Synthesis from Meaning Representation for Zero-Resource Semantic Parsing

Deokhyung Kang, Seonjeong Hwang, Yunsu Kim, Gary Lee

Recent efforts have aimed to utilize multilingual pretrained language models (mPLMs) to extend semantic parsing (SP) across multiple languages without requiring extensive annotations. However, achieving zero-shot cross-lingual transfer for SP remains challenging, leading to a performance gap between source and target languages. In this study, we propose Cross-Lingual Back-Parsing (CBP), a novel data augmentation methodology designed to enhance cross-lingual transfer for SP. Leveraging the representation geometry of the mPLMs, CBP synthesizes target language utterances from source meaning representations. Our methodology effectively performs cross-lingual data augmentation in challenging zero-resource settings, by utilizing only labeled data in the source language and monolingual corpora. Extensive experiments on two cross-language SP benchmarks (Mschema2QA and Xspider) demonstrate that CBP brings substantial gains in the target language. Further analysis of the synthesized utterances shows that our method successfully generates target language utterances with high slot value alignment rates while preserving semantic integrity. Our codes and data are publicly available at <https://github.com/deokhk/CBP>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

A Comparison of Language Modeling and Translation as Multilingual Pretraining Objectives

Zhihao Li, Shaoxiong Ji, Timothee Miclus, Vincent Segonne, Jörg Tiedemann

Pretrained language models (PLMs) display impressive performances and have captured the attention of the NLP community. Establishing best practices in pretraining has, therefore, become a major focus of NLP research, especially since insights gained from monolingual English models may not necessarily apply to more complex multilingual models. One significant caveat of the current state of the art is that different works are rarely comparable: they often discuss different parameter counts, training data, and evaluation methodology. This paper proposes a comparison of multilingual pretraining objectives in a controlled methodological environment. We ensure that training data and model architectures are comparable, and discuss the downstream performances across 6 languages that we observe in probing and fine-tuning scenarios. We make two key observations: (1) the architecture dictates which pretraining objective is optimal; (2) multilingual translation is a very effective pretraining objective under the right conditions. We make our code, data, and model weights available at <https://github.com/Helsinki-NLP/Im-vs-mt>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Evaluating Large Language Models along Dimensions of Language Variation: A Systematic Investigation of Cross-lingual Generalization

Niyati Bafna, Kenton Murray, David Yarowsky

While large language models exhibit certain cross-lingual generalization capabilities, they suffer from performance degradation (PD) on unseen closely-related languages (CRLs) and dialects relative to their high-resource language neighbour (HRLN). However, we currently lack a fundamental understanding of what kinds of linguistic distances contribute to PD, and to what extent. Furthermore, studies of cross-lingual generalization are confounded by unknown quantities of CRL language traces in the training data, and by the frequent lack of availability of evaluation data in lower-resource related languages and dialects. To address these issues, we model phonological, morphological, and lexical distance as Bayesian noise processes to synthesize artificial languages that are controllably distant from the HRLN. We analyse PD as a function of underlying noise parameters, offering insights on model robustness to isolated and composed linguistic phenomena, and the impact of task and HRL characteristics on PD. We calculate parameter posteriors on real CRL-HRLN pair data and show that they follow computed trends of artificial languages, demonstrating the viability of our noisers. Our framework offers a cheap solution for estimating task performance on an unseen CRL given HRLN performance using its posteriors, as well as for diagnosing observed PD on a CRL in terms of its linguistic distances from its HRLN, and opens doors to principled methods of mitigating performance degradation.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Re-Evaluating Evaluation for Multilingual Summarization

Jessica Zosa Forde, Ruochen Zhang, Lintang Sutawika, Alham Fikri Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Bidenko, Ellie Pavlick

Automatic evaluation approaches (ROUGE, BERTScore, LLM-based evaluators) have been widely used to evaluate summarization tasks. Despite the complexities of script differences and tokenization, these approaches have been indiscriminately applied to summarization across multiple languages. While previous works have argued that these approaches correlate strongly with human ratings in English, it remains unclear whether the conclusion holds for other languages. To answer this question, we construct a small-scale pilot dataset containing article-summary pairs and human ratings in English, Chinese and Indonesian. To measure the strength of summaries, our ratings are measured as head-to-head comparisons with resulting Elo scores across four dimensions. Our analysis reveals that standard metrics are unreliable measures of quality, and that these problems are exacerbated in Chinese and Indonesian. We advocate for more nuanced and careful considerations in designing a robust evaluation framework for multiple languages.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Unsupervised Discrete Representations of American Sign Language

Artem Abzaliev, Rada Mihalcea

Many modalities are naturally represented as continuous signals, making it difficult to use them with models that expect discrete units, such as LLMs. In this paper, we explore the use of audio compression techniques for the discrete representation of the gestures used in sign language. We train a tokenizer for American Sign Language (ASL) fingerspelling, which discretizes sequences of fingerspelling signs into tokens. We also propose a loss function to improve the interpretability of these tokens such that they preserve both the semantic and the visual information of the signal. We show that the proposed method improves the performance of the discretized sequence on downstream tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Multilingual Topic Classification in X: Dataset and Analysis

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Jose Camacho-Collados

In the dynamic realm of social media, diverse topics are discussed daily, transcending linguistic boundaries. However, the complexities of understanding and categorising this content across various languages remain an important challenge with traditional techniques like topic modelling often struggling to accommodate this multilingual diversity. In this paper, we introduce X-Topic, a multilingual dataset featuring content in four distinct languages (English, Spanish, Japanese, and Greek), crafted for the purpose of tweet topic classification. Our dataset includes a wide range of topics, tailored for social media content, making it a valuable resource for scientists and professionals working on cross-linguistic analysis, the development of robust multilingual models, and computational scientists studying online dialogue. Finally, we leverage X-Topic to perform a comprehensive cross-linguistic and multilingual analysis, and compare the capabilities of current general- and domain-specific language models.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions?

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, Mehdi Ali

The adaption of multilingual pre-trained LLMs into eloquent and helpful assistants is essential to facilitate their use across different language regions. In that spirit, we are the first to conduct an extensive study of the performance of multilingual models instruction-tuned on different language compositions on parallel instruction-tuning benchmarks across a selection of the most spoken Indo-European languages. We systematically examine the effects of language and instruction dataset size on a mid-sized and a large, multilingual LLMs by instruction-tuning them on parallel instruction-tuning datasets. Our results demonstrate that instruction-tuning on parallel instead of monolingual corpora benefits cross-lingual instruction following capabilities by up to 9.9%. Furthermore, we show that the Superficial Alignment Hypothesis does not hold in general, as the investigated multilingual 7B parameter model presents a counter-example requiring large-scale instruction-tuning datasets. Finally, we conduct a human annotation study to understand the alignment between human-based and GPT-4-based evaluation within multilingual chat scenarios.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Self-Distillation for Model Stacking Unlocks Cross-Lingual NLU in 200+ Languages

Fabian David Schmidt, Philipp Borchert, Ivan Vulic, Goran Glava

LLMs have become a go-to solution not just for text generation, but also for natural language understanding (NLU) tasks. Acquiring extensive knowledge through language modeling on web-scale corpora, they excel on English NLU, yet struggle to extend their NLU capabilities to underrepresented languages. In contrast, machine translation models (MT) produce excellent multilingual representations, resulting in strong translation performance even for low-resource languages. MT encoders, however, lack the knowledge necessary for comprehensive NLU that LLMs obtain through language modeling training on immense corpora. In this work, we get the best words by integrating MT encoders directly into LLM backbones via sample-efficient self-distillation. The resulting MT-LLMs preserve the inherent multilingual representational alignment from the MT encoder, allowing lower-resource languages to tap into the rich knowledge embedded in English-centric LLMs. Merging the MT encoder and LLM in a single model, we mitigate the propagation of translation errors and inference overhead of MT decoding inherent to discrete translation-based cross-lingual transfer (e.g., translate-test). Evaluation spanning three prominent NLU tasks and 127 predominantly low-resource languages renders MT-LLMs highly effective in cross-lingual transfer. MT-LLMs substantially and consistently outperform translation-test based on the same MT model, showing that we truly unlock multilingual language understanding for LLMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

An Empirical Study on Cross-lingual Vocabulary Adaptation for Efficient Language Model Inference

Atsuki Yamaguchi, Alíne Villavicencio, Nikolaos Aletras

The development of state-of-the-art generative large language models (LLMs) disproportionately relies on English-centric tokenizers, vocabulary and pre-training data. Despite the fact that some LLMs have multilingual capabilities, recent studies have shown that their inference efficiency deteriorates when generating text in languages other than English. This results in increased inference time and costs. Cross-lingual vocabulary adaptation (CVA) methods have been proposed for adapting models to a target language aiming to improve downstream performance. However, the effectiveness of these methods on increasing inference efficiency of generative LLMs has yet to be explored. In this paper, we perform an empirical study of five CVA methods on four generative LLMs (including monolingual and multilingual models) across four typologically-diverse languages and four natural language understanding tasks. We find that CVA substantially contributes to LLM inference speedups of up to 271.5%. We also show that adapting LLMs that have been pre-trained on more balanced multilingual data results in downstream performance comparable to the original models.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, Eduardo Sánchez
Recent advancements in massively multilingual machine translation systems have significantly enhanced translation accuracy; however, even the best performing systems still generate hallucinations, severely impacting user trust. Detecting hallucinations in Machine Translation (MT) remains a critical challenge, particularly since existing methods excel with High-Resource Languages (HRLs) but exhibit substantial limitations when applied to Low-Resource Languages (LRLs). This paper evaluates sentence-level hallucination detection approaches using Large Language Models (LLMs) and semantic similarity within massively multilingual embeddings. Our study spans 16 language directions, covering HRLs, LRLs, with diverse scripts. We find that the choice of model is essential for performance. On average, for HRLs, Llama3-70B outperforms the previous state of the art by as much as 0.16 MCC (Matthews Correlation Coefficient). However, for LRLs we observe that Claude Sonnet outperforms other LLMs on average by 0.03 MCC. The key takeaway from our study is that LLMs can achieve performance comparable or even better than previously proposed models, despite not being explicitly trained for any machine translation task. However, their advantage is less significant for LRLs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Pruning Multilingual Large Language Models for Multilingual Inference

Hwchan Kim, Jun Suzuki, Toshi Hirasawa, Mamoru Komachi

Multilingual large language models (MLLMs), trained on multilingual balanced data, demonstrate better zero-shot learning performance in non-English languages compared to large language models trained on English-dominant data. However, the disparity in performance between English and non-English languages remains a challenge yet to be fully addressed. This study introduces a promising direction for enhancing non-English performance through a specialized pruning approach. Specifically, we prune MLLMs using bilingual sentence pairs from English and other languages and empirically demonstrate that this pruning strategy can enhance the MLLMs' performance in non-English language.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Word-Conditioned 3D American Sign Language Motion Generation

Lu Dong, Xiao Wang, Ifeoma Nwogu

Sign words are the building blocks of any sign language. In this work, we present wSignGen, a word-conditioned 3D American Sign Language (ASL) generation model dedicated to synthesizing realistic and grammatically accurate motion sequences for sign words. Our approach leverages a transformer-based diffusion model, trained on a curated dataset of 3D motion meshes from word-level ASL videos. By integrating CLIP, wSignGen offers two advantages: image-based generation, which is particularly useful for children learning sign language but not yet able to read, and the ability to generalize to unseen synonyms. Experiments demonstrate that wSignGen significantly outperforms the baseline model in the task of sign word generation. Moreover, human evaluation experiments show that wSignGen can generate high-quality, grammatically correct ASL signs effectively conveyed through 3D avatars.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Exploring Design Choices for Building Language-Specific LLMs

Artula Tejaswi, Nitesh Gupta, Eunsol Choi

Despite rapid progress in large language models (LLMs), their performance on a vast majority of languages remains unsatisfactory. In this paper, we study building language-specific LLMs by adapting monolingual and multilingual LLMs. We conduct systematic experiments on how design choices (base model selection, vocabulary extension, and continued pretraining) impact the adapted LLM, both in terms of efficiency (how many tokens are needed to encode the same amount of information) and end task performance. We find that (1) the initial performance of LLM does not always correlate with the final performance after the adaptation. Adapting an English-centric models can yield better results than adapting multilingual models despite their worse initial performance on low-resource languages. (2) Efficiency can easily improved with simple vocabulary extension and continued pretraining in most LLMs we study, and (3) The optimal adaptation method (choice of the base model, new vocabulary size, training data, initialization strategy) is highly language-dependent, and the simplest embedding initialization works well across various experimental settings. Together, our work lays foundations on efficiently building language-specific LLMs by adapting existing LLMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment

Oregel Xhelili, Yihong Liu, Hinrich Schütze

Multilingual pre-trained models (mPLMs) have shown impressive performance on cross-lingual transfer tasks. However, the transfer performance is often hindered when a low-resource target language is written in a different script than the high-resource source language, even though the two languages may be related or share parts of their vocabularies. Inspired by recent work that uses transliteration to address this problem, our paper proposes a transliteration-based post-pretraining alignment (PPA) method aiming to improve the cross-lingual alignment between languages using diverse scripts. We select two areal language groups, **Mediterranean-Amharic-Farsi** and **South+East Asian Languages**, wherein the languages are mutually influenced but use different scripts. We apply our method to these language groups and conduct extensive experiments on a spectrum of downstream tasks. The results show that after PPA, models consistently outperform the original model (up to 50% for some tasks) in English-centric transfer. In addition, when we use languages other than English as sources in transfer, our method obtains even larger improvements. We will make our code and models publicly available at <https://github.com/cisnlp/ComTrans-Transliteration-PPA>.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Cross-Lingual Multi-Hop Knowledge Editing – Benchmarks, Analysis and a Simple Contrastive Learning based Approach

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, Kaixiong Zhou

Large language models (LLMs) are often expected to be constantly adapted to new sources of knowledge and knowledge editing techniques aim to efficiently patch the outdated model knowledge, with minimal modification. Most prior works focus on monolingual knowledge editing in English, even though new information can emerge in any language from any part of the world. We propose the Cross-Lingual Multi-Hop Knowledge Editing paradigm, for measuring and analyzing the performance of various SoTA knowledge editing techniques in a cross-lingual setup. Specifically, we create a parallel cross-lingual benchmark, CroLin-MQuAKE for measuring the knowledge editing capabilities. Our extensive analysis over various knowledge editing techniques uncover significant gaps in performance between the cross-lingual and English-centric setting. Following this, we propose a significantly improved system for cross-lingual multi-hop knowledge editing, CLeVer-CKE. CLeVer-CKE is based on a retrieve, verify and generate knowledge editing framework, where a retriever is formulated to recall edited facts and support an LLM to adhere to knowledge edits. We develop language-aware and hard-negative based contrastive losses for improving the cross-lingual and fine-grained fact retrieval and verification process used within this framework. Extensive experiments across three LLMs, eight languages, and two datasets show the CLeVer-CKE's significant gains of up to 30% over prior methods.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed?

Tannor Kew, Florian Schottmann, Rico Sennrich

The vast majority of today's large language models (LLMs) are English-centric, having been pretrained predominantly on English text. Yet,

in order to meet user expectations, models need to be able to respond appropriately in multiple languages once deployed in downstream applications. This requires strong cross-lingual transfer abilities. In this work, we investigate the minimal amount of multilinguality required during finetuning to elicit cross-lingual generalisation in English-centric LLMs. In experiments across four LLMs, we find that multilingual instruction tuning with as few as two to three languages is both necessary and sufficient to elicit effective cross-lingual generalisation, with the limiting factor being the degree to which a target language is seen during pretraining. Evaluations on five different tasks further reveal that multilingual instruction tuning is most beneficial for generative tasks that assume input/output language agreement, such as in chat settings, while being of less importance for highly structured classification-style tasks.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Representational Isomorphism and Alignment of Multilingual Large Language Models

Di Wu, Yibin Lei, Andrew Yates, Christof Monz

In this paper, we investigate the capability of Large Language Models (LLMs) to represent texts in multilingual contexts. Our findings show that sentence representations derived from LLMs exhibit a high degree of isomorphism across languages. This existing isomorphism can facilitate representational alignments in zero-shot and few-shot settings. Specifically, by applying a contrastive objective at the representation level with only a small number of translation pairs (e.g., 100), we substantially improve models' performance on Semantic Textual Similarity (STS) tasks across languages. This representation-level approach proves to be more efficient and effective for semantic alignment than continued pretraining or instruction tuning. Interestingly, we also observe substantial STS improvements within individual languages, even without a monolingual objective specifically designed for this purpose.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

SSP: Self-Supervised Prompting for Cross-Lingual Transfer to Low-Resource Languages using Large Language Models

Vipul Kumar Rathore, Aniruddha Deb, Ankish Kumar Chandresh, Parag Singla, Mausam

Recently, very large language models (LLMs) have shown exceptional performance on several English NLP tasks with just in-context learning (ICL), but their utility in other languages is still underexplored. We investigate their effectiveness for NLP tasks in low-resource languages (LRLs), especially in the setting of zero-labelled cross-lingual transfer (0-CLT), where no labelled training data for the target language is available – however training data from one or more related medium-resource languages (MRLs) is utilized, alongside the available unlabeled test data for a target language. We introduce Self-Supervised Prompting (SSP), a novel ICL approach tailored for the 0-CLT setting. SSP is based on the key observation that LLMs output more accurate labels if in-context exemplars are from the target language (even if their labels are slightly noisy). To operationalize this, since target language training data is not available in 0-CLT, SSP operates in two stages. In Stage I, using source MRL training data, target language's test data is noisily labeled. In Stage II, these noisy test data points are used as exemplars in ICL for further improved labelling. Additionally, our implementation of SSP uses a novel Integer Linear Programming (ILP)-based exemplar selection that balances similarity, prediction confidence (when available) and label coverage. Experiments on three tasks and eleven LRLs (from three regions) demonstrate that SSP strongly outperforms existing SOTA fine-tuned and prompting-based baselines in 0-CLT setup.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Targeted Multilingual Adaptation for Low-resource Language Families

C. M. Downey, Terra Blevins, Dhwanvi Serai, Dwija Parikh, Shane Steinert-Threlkeld

Massively multilingual models are known to have limited utility in any one language, and to perform particularly poorly on low-resource languages. By contrast, targeted multilinguality has been shown to benefit low-resource languages. To test this approach more rigorously, we systematically study best practices for adapting a pre-trained model to a language family. Focusing on the Uralic family as a test case, we adapt XLM-R under various configurations to model 15 languages; we then evaluate the performance of each experimental setting on two downstream tasks and 11 evaluation languages. Our adapted models significantly outperform mono- and multilingual baselines. A regression analysis reveals that adapted vocabulary size is relatively unimportant for low-resource languages, and that low-resource languages can be aggressively up-sampled during training at little detriment to performance in high-resource languages. These results introduce new best practices for performing language adaptation in a targeted setting.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Mitigating Catastrophic Forgetting in Language Transfer via Model Merging

Anton Alexandrov, Veselin Raychev, Mark Niklas Mueller, Ce Zhang, Martin Vechev, Kristina Toutanova

As open-weight large language models (LLMs) achieve ever more impressive performance across a wide range of tasks in English, practitioners aim to adapt these models to different languages. However, such language adaptation is often accompanied by catastrophic forgetting of the base models' capabilities, severely limiting the usefulness of the resulting model. We address this issue by proposing Branch-and-Merge (BaM), a new adaptation method based on iteratively merging multiple models, fine-tuned on a subset of the available training data. BaM is based on the insight that this yields lower magnitude but higher quality weight changes, reducing forgetting of the source domain while maintaining learning on the target domain. We demonstrate in an extensive empirical study on Bulgarian and German that BaM can significantly reduce forgetting while matching or even improving target domain performance compared to both standard continued pretraining and instruction finetuning across different model architectures.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Improving Multilingual Instruction Finetuning via Linguistically Natural and Diverse Datasets

Sathish Reddy Indurthi, Wenzuan Zhou, Shamil Chollampatt, Ravi Agrawal, Kaiqiang Song, Lingxiao Zhao, Chenguang Zhu

Advancements in Large Language Models (LLMs) have significantly enhanced instruction-following capabilities. However, most Instruction-Fine-Tuning (IFT) datasets are predominantly in English, limiting model performance in other languages. Traditional methods for creating multilingual IFT datasets such as translating existing English IFT datasets or converting existing NLP datasets into IFT datasets by templating struggle to capture linguistic nuances and ensure prompt (instruction) diversity. To address this issue, we propose a novel method for collecting multilingual IFT datasets that preserves linguistic naturalness and ensures prompt diversity. This approach leverages English-focused LLMs, monolingual corpora, and a scoring function to create high-quality, diversified IFT datasets in multiple languages. Experiments demonstrate that LLMs finetuned using these IFT datasets show notable improvements in both generative and discriminative tasks, indicating enhanced language comprehension by LLMs in non-English contexts. Specifically, on the multilingual summarization task, LLMs using our IFT dataset achieved 17.57% and 15.23% improvements over LLMs fine-tuned with translation-based and template-based datasets, respectively.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Are modern neural ASR architectures robust for polysynthetic languages?

Eric Le Ferrand, Zoey Liu, Antti Arppe, Emily Prud'hommeaux

Automatic speech recognition (ASR) technology is frequently proposed as a means of preservation and documentation of endangered languages, with promising results thus far. Among the endangered languages spoken today, a significant number exhibit complex morphology. The models employed in contemporary language documentation pipelines that utilize ASR, however, are predominantly based on isolating or inflectional languages, often from the Indo-European family. This raises a critical concern: building models exclusively on such languages

may introduce a bias, resulting in better performance with simpler morphological structures. In this paper, we investigate the performance of modern ASR architectures on morphologically complex languages. Results indicate that modern ASR architectures appear less robust in managing high OOV rates for morphologically complex languages in terms of word error rate, while character error rates are consistently higher for isolating languages.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

LinguAlchemy: Fusing Typological and Geographical Elements for Unseen Language Generalization

Muhammad Farid Adilzuarda, Samuel Cahyawijaya, Genta Indra Winata, Ayu Purwarianti, Alham Fikri Ajji

Pretrained language models (PLMs) have shown remarkable generalization toward multiple tasks and languages. Nonetheless, the generalization of PLMs towards unseen languages is poor, resulting in significantly worse language performance, or even generating nonsensical responses that are comparable to a random baseline. This limitation has been a longstanding problem of PLMs raising the problem of diversity and equal access to language modeling technology. In this work, we solve this limitation by introducing LinguAlchemy, a regularization technique that incorporates various aspects of languages covering typological, geographical, and phylogenetic constraining the resulting representation of PLMs to better characterize the corresponding linguistics constraints. LinguAlchemy significantly improves the accuracy performance of mBERT and XLM-R on unseen languages by 18% and 2%, respectively compared to fully finetuned models and displaying a high degree of unseen language generalization. We further introduce AlchemyScale and AlchemyTune, extension of LinguAlchemy which adjusts the linguistic regularization weights automatically, alleviating the need for hyperparameter search. LinguAlchemy enables better cross-lingual generalization to unseen languages which is vital for better inclusivity and accessibility of PLMs.

Nov 12 (Tue) 14:00-15:30 - Riverfront Hall

Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, Luke Zettlemoyer

Despite their popularity in non-English NLP, multilingual language models often underperform monolingual ones due to inter-language competition for model parameters. We propose Cross-lingual Expert Language Models (X-ELM), which mitigate this competition by independently training language models on subsets of the multilingual corpus. This process specializes X-ELMs to different languages while remaining effective as a multilingual ensemble. Our experiments show that when given the same compute budget, X-ELM outperforms jointly trained multilingual models across all 16 considered languages and that these gains transfer to downstream tasks. X-ELM provides additional benefits over performance improvements: new experts can be iteratively added, adapting X-ELM to new languages without catastrophic forgetting. Furthermore, training is asynchronous, reducing the hardware requirements for multilingual training and democratizing multilingual modeling.

Session 04 - Nov 12 (Tue) 16:00-17:30

Computational Social Science and Cultural Analytics 2

Nov 12 (Tue) 16:00-17:30 - Room: Riverfront Hall

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Integrating Plutchik's Theory with Mixture of Experts for Enhancing Emotion Classification

Dongjun LIM, Yun-Cyoung Cheong

Emotion significantly influences human behavior and decision-making processes. We propose a labeling methodology grounded in Plutchik's Wheel of Emotions theory for emotion classification. Furthermore, we employ a Mixture of Experts (MoE) architecture to evaluate the efficacy of this labeling approach, by identifying the specific emotions that each expert learns to classify. Experimental results reveal that our methodology improves the performance of emotion classification.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Systematic Biases in LLM Simulations of Debates

Amir Taubenfeld, Yaniv Dover, Roi Reichart, Ariel Goldstein

The emergence of Large Language Models (LLMs), has opened exciting possibilities for constructing computational simulations designed to replicate human behavior accurately. Current research suggests that LLM-based agents become increasingly human-like in their performance, sparking interest in using these AI agents as substitutes for human participants in behavioral studies. However, LLMs are complex statistical learners without straightforward deductive rules, making them prone to unexpected behaviors. Hence, it is crucial to study and pinpoint the key behavioral distinctions between humans and LLM-based agents. In this study, we highlight the limitations of LLMs in simulating human interactions, particularly focusing on LLMs' ability to simulate political debates on topics that are important aspects of people's day-to-day lives and decision-making processes. Our findings indicate a tendency for LLM agents to conform to the model's inherent social biases despite being directed to debate from certain political perspectives. This tendency results in behavioral patterns that seem to deviate from well-established social dynamics among humans. We reinforce these observations using an automatic self-fine-tuning method, which enables us to manipulate the biases within the LLM and demonstrate that agents subsequently align with the altered biases. These results underscore the need for further research to develop methods that help agents overcome these biases, a critical step toward creating more realistic simulations.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Tracking the perspectives of interacting language models

Hayden Helm, Brandon Duderstadt, Youngser Park, Carey Priebe

Large language models (LLMs) are capable of producing high quality information at unprecedented rates. As these models continue to entrench themselves in society, the content they produce will become increasingly pervasive in databases that are, in turn, incorporated into the pre-training data, fine-tuning data, retrieval data, etc. of other language models. In this paper we formalize the idea of a communication network of LLMs and introduce a method for representing the perspective of individual models within a collection of LLMs. Given these tools we systematically study information diffusion in the communication network of LLMs in various simulated settings.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DGLF: A Dual Graph-based Learning Framework for Multi-modal Sarcasm Detection

Zhihong Zhu, Kefan Shen, Zhaorun Chen, Yunyan Zhang, Yuyan Chen, Xiaoqi Jiao, Zhongwei Wan, Wei Liu, Xian Wu, Shaorong Xie, Yefeng Zheng

Capturing inter-modal incongruities within the text-image pair is a critical challenge in multi-modal sarcasm detection (MSD). Fortunately, graph neural networks (GNNs) have made promising advancements in MSD, which show advantages in explicitly capturing data relationships. Nevertheless, current GNN-based MSD methods do not effectively address some of the inherent limitations of GNNs, which include: 1) neglecting high-order relationships, and 2) underestimating high-frequency messages. In this paper, we propose a **Dual Graph-based Learning Framework (DGLF)** to address the above two issues. Specifically, we construct a hypergraph to perform high-order aware propagation and a vanilla graph to perform high-frequency enhanced propagation, respectively. We empower GNNs to 1) better capture the inherent and complicated relationships based on the hypergraph and 2) deliver sufficient modeling through high-frequency enhanced messages on the vanilla graph. Besides, we introduce multi-modal fusion information bottleneck to effectively fuse the two learned graph features. Experimental results on two benchmark datasets show that the proposed model outperforms previous state-of-the-art methods.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

The Computational Anatomy of Humility: Modeling Intellectual Humility in Online Public Discourse

Xiaobo Guo, Neil Potnis, Melody Yu, Nabeel Gillani, Soroush Vosoughi

The ability for individuals to constructively engage with one another across lines of difference is a critical feature of a healthy pluralistic society. This is also true in online discussion spaces like social media platforms. To date, much social media research has focused on preventing ills—like political polarization and the spread of misinformation. While this is important, enhancing the quality of online public discourse requires not just reducing ills, but also, promoting informational human virtues. In this study, we focus on one particular virtue: “intellectual humility” (IH), or acknowledging the potential limitations in one’s own beliefs. Specifically, we explore the development of computational methods for measuring IH at scale. We manually curate and validate an IH codebook on 350 posts about religion drawn from subreddits and use them to develop LLM-based models for automating this measurement. Our best model achieves a Macro-F1 score of 0.64 across labels (and 0.70 when predicting IH/A/Neutral at the coarse level), higher than an expected naive baseline of 0.51 (0.32 for IH/A/Neutral) but lower than a human annotator-informed upper bound of 0.85 (0.83 for IH/A/Neutral). Our results both highlight the challenging nature of detecting IH online—opening the door to new directions in NLP research—and also lay a foundation for computational social science researchers interested in analyzing and fostering more IH in online public discourse.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Analyzing Key Factors Influencing Emotion Prediction Performance of VLLMs in Conversational Contexts

Jaewook Lee, Yejin Jang, Hongjin Kim, Woojin Lee, Harksoo Kim

Emotional intelligence (EI) in artificial intelligence (AI), which refers to the ability of an AI to understand and respond appropriately to human emotions, has emerged as a crucial research topic. Recent studies have shown that large language models (LLMs) and vision large language models (VLLMs) possess EI and the ability to understand emotional stimuli in the form of text and images, respectively. However, factors influencing the emotion prediction performance of VLLMs in real-world conversational contexts have not been sufficiently explored. This study aims to analyze the key elements affecting the emotion prediction performance of VLLMs in conversational contexts systematically. To achieve this, we reconstructed the MELD dataset, which is based on the popular TV series Friends, and conducted experiments through three sub-tasks: overall emotion tone prediction, character emotion prediction, and contextually appropriate emotion expression selection. We evaluated the performance differences based on various model architectures (e.g., image encoders, modality alignment, and LLMs) and image scopes (e.g., entire scene, person, and facial expression). In addition, we investigated the impact of providing persona information on the emotion prediction performance of the models and analyzed how personality traits and speaking styles influenced the emotion prediction process. We conducted an in-depth analysis of the impact of various other factors, such as gender and regional biases, on the emotion prediction performance of VLLMs. The results revealed that these factors significantly influenced the model performance.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations

Yunze Xiao, Yujia Hu, Kenny Tsui Wei Choo, Roy Ka-Wei Lee

Detecting hate speech and offensive language is essential for maintaining a safe and respectful digital environment. This study examines the limitations of state-of-the-art large language models (LLMs) in identifying offensive content within systematically perturbed data, with a focus on Chinese, a language particularly susceptible to such perturbations. We introduce ToxiCloakCN, an enhanced dataset derived from ToxiCN, augmented with homophonic substitutions and emoji transformations, to test the robustness of LLMs against these cloaking perturbations. Our findings reveal that existing models significantly underperform in detecting offensive content when these perturbations are applied. We provide an in-depth analysis of how different types of offensive content are affected by these perturbations and explore the alignment between human and model explanations of offensiveness. Our work highlights the urgent need for more advanced techniques in offensive language detection to combat the evolving tactics used to evade detection mechanisms.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Locating Information Gaps and Narrative Inconsistencies Across Languages: A Case Study of LGBT People Portrayals on Wikipedia

Farhan Samir, Chan Young Park, Vered Shwartz, Anjalie Field, Yulita Tsvetkov

To explain social phenomena and identify systematic biases, much research in computational social science focuses on comparative text analyses. These studies often rely on coarse corpus-level statistics or local word-level analyses, mainly in English. We introduce the InfoGap method—an efficient and reliable approach to locating information gaps and inconsistencies in articles at the fact level, across languages. We evaluate InfoGap by analyzing LGBT people’s portrayals, across 2.7K biography pages on English, Russian, and French Wikipedias. We find large discrepancies in factual coverage across the languages. Moreover, our analysis reveals that biographical facts carrying negative connotations are more likely to be highlighted in Russian Wikipedia. Crucially, InfoGap both facilitates large scale analyses, and pinpoints local document- and fact-level information gaps, laying a new foundation for targeted and nuanced comparative language analysis at scale.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models

Mehar Bhateria, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, Vered Shwartz

Despite recent advancements in vision-language models, their performance remains suboptimal on images from non-western cultures due to underrepresentation in training datasets. Various benchmarks have been proposed to test models’ cultural inclusivity. Still, they have limited coverage of cultures and do not adequately assess cultural diversity across universal and culture-specific local concepts. To address these limitations, we introduce the GlobalRG benchmark, comprising two challenging tasks: retrieval across universals and cultural visual grounding. The former task entails retrieving culturally diverse images for universal concepts from 50 countries, while the latter aims at grounding culture-specific concepts within images from 15 countries. Our evaluation across a wide range of models reveals that the performance varies significantly across cultures – underscoring the necessity for enhancing multicultural understanding in vision-language models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Bridging Modalities: Enhancing Cross-Modality Hate Speech Detection with Few-Shot In-Context Learning

Ming Shan Hee, Aditi Kumaresan, Roy Ka-Wei Lee

The widespread presence of hate speech on the internet, including formats such as text-based tweets and multimodal memes, poses a signif-

icant challenge to digital platform safety. Recent research has developed detection models tailored to specific modalities; however, there is a notable gap in transferring detection capabilities across different formats. This study conducts extensive experiments using few-shot in-context learning with large language models to explore the transferability of hate speech detection between modalities. Our findings demonstrate that text-based hate speech examples can significantly enhance the classification accuracy of vision-language hate speech. Moreover, text-based demonstrations outperform vision-language demonstrations in few-shot learning settings. These results highlight the effectiveness of cross-modality knowledge transfer and offer valuable insights for improving hate speech detection systems.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Image, Tell me your story! Predicting the original meta-context of visual misinformation

Jonathan Tongler, Marie-Francine Moens, Iryna Gurevych

To assist human fact-checkers, researchers have developed automated approaches for visual misinformation detection. These methods assign veracity scores by identifying inconsistencies between the image and its caption, or by detecting forgeries in the image. However, they neglect a crucial point of the human fact-checking process: identifying the original meta-context of the image. By explaining what is actually true about the image, fact-checkers can better detect misinformation, focus their efforts on check-worthy visual content, engage in counter-messaging before misinformation spreads widely, and make their explanation more convincing. Here, we fill this gap by introducing the task of automated image contextualization. We create 5PiLs, a dataset of 1,676 fact-checked images with question-answer pairs about their original meta-context. Annotations are based on the 5 Pillars fact-checking framework. We implement a first baseline that grounds the image in its original meta-context using the content of the image and textual evidence retrieved from the open web. Our experiments show promising results while highlighting several open challenges in retrieval and reasoning.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Generative AI in the Era of "Alternative Facts"

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, Asuman E. Ozdaglar

The spread of misinformation on social media platforms threatens democratic processes, contributes to massive economic losses, and endangers public health. Many efforts to address misinformation focus on a knowledge deficit model and propose interventions for improving users critical thinking through access to facts. Such efforts are often hampered by challenges with scalability, and by platform users personal biases. The emergence of generative AI presents promising opportunities for countering misinformation at scale across ideological barriers. In this paper, we introduce a framework (MisinfoEval) for generating and comprehensively evaluating large language model (LLM) based misinformation interventions. We present (1) an experiment with a simulated social media environment to measure effectiveness of misinformation interventions, and (2) a second experiment with personalized explanations tailored to the demographics and beliefs of users with the goal of countering misinformation by appealing to their pre-existing values. Our findings confirm that LLM-based interventions are highly effective at correcting user behavior (improving overall user accuracy at reliability labeling by up to 41.72%). Furthermore, we find that users favor more personalized interventions when making decisions about news reliability and users shown personalized interventions have significantly higher accuracy at identifying misinformation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Words Matter: Reducing Stigma in Online Conversations about Substance Use with Large Language Models

Layla Bouzoubaa, Elham Aghakhani, Shadi Rezapour

Stigma is a barrier to treatment for individuals struggling with substance use disorders (SUD), which leads to significantly lower treatment engagement rates. With only 7% of those affected receiving any form of help, societal stigma not only discourages individuals with SUD from seeking help but isolates them, hindering their recovery journey and perpetuating a cycle of shame and self-doubt. This study investigates how stigma manifests on social media, particularly Reddit, where anonymity can exacerbate discriminatory behaviors. We analyzed over 1.2 million posts, identifying 3,207 that exhibited stigmatizing language related to people who use substances (PWUS). Of these, 1,649 posts were classified as containing directed stigma towards PWUS, which became the focus of our de-stigmatization efforts. Using Informed and Stylized LLMs, we developed a model to transform these instances into more empathetic language. Our paper contributes to the field by proposing a computational framework for analyzing stigma and de-stigmatizing online content, and delving into the linguistic features that propagate stigma towards PWUS. Our work not only enhances understanding of stigma's manifestations online but also provides practical tools for fostering a more supportive environment for those affected by SUD.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Story Morals: Surfacing value-driven narrative schemas using large language models

David G Hobson, Haiqi Zhou, Derek Ruths, Andrew Piper

Stories are not only designed to entertain but encode lessons reflecting their authors' beliefs about the world. In this paper, we propose a new task of narrative schema labelling based on the concept of "story morals" to identify the values and lessons conveyed in stories. Using large language models (LLMs) such as GPT-4, we develop methods to automatically extract and validate story morals across a diverse set of narrative genres, including folktales, novels, movies and TV, personal stories from social media and the news. Our approach involves a multi-step prompting sequence to derive morals and validate them through both automated metrics and human assessments. The findings suggest that LLMs can effectively approximate human story moral interpretations and offer a new avenue for computational narrative understanding. By clustering the extracted morals on a sample dataset of folktales from around the world, we highlight the commonalities and distinctiveness of narrative values, providing preliminary insights into the distribution of values across cultures. This work opens up new possibilities for studying narrative schemas and their role in shaping human beliefs and behaviors.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Improving Logical Fallacy Reasoning with Logical Structure Tree

Yuanyu Lei, Ruihong Huang

Logical fallacy uses invalid or faulty reasoning in the construction of a statement. Despite the prevalence and harmfulness of logical fallacies, detecting and classifying logical fallacies still remains a challenging task. We observe that logical fallacies often use connective words to indicate an intended logical relation between two arguments, while the argument semantics does not actually support the logical relation. Inspired by this observation, we propose to build a logical structure tree to explicitly represent and track the hierarchical logic flow among relation connectives and their arguments in a statement. Specifically, this logical structure tree is constructed in an unsupervised manner guided by the constituency tree and a taxonomy of connectives for ten common logical relations, with relation connectives as non-terminal nodes and textual arguments as terminal nodes, and the latter are mostly elementary discourse units. We further develop two strategies to incorporate the logical structure tree into LLMs for fallacy reasoning. Firstly, we transform the tree into natural language descriptions and feed the textualized tree into LLMs as a part of the hard text prompt. Secondly, we derive a relation-aware tree embedding and insert the tree embedding into LLMs as a soft prompt. Experiments on benchmark datasets demonstrate that our approach based on logical structure tree significantly improves precision and recall for both fallacy detection and fallacy classification.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

A Closer Look at Multidimensional Online Political Incivility

Sagi Pendzel, Nir Lotan, Alon Zoisner, Einat Minkov

Toxic online political discourse has become prevalent, where scholars debate about its impact to Democratic processes. This work presents a large-scale study of political incivility on Twitter. In line with theories of political communication, we differentiate between harsh ‘impolite’ style and intolerant substance. We present a dataset of 13K political tweets in the U.S. context, which we collected and labeled by those categories using crowd sourcing. Our dataset and results shed light on hostile political discourse focused on partisan conflicts in the U.S. The evaluation of state-of-the-art classifiers illustrates the challenges involved in political incivility detection, which often requires high-level semantic and social understanding. Nevertheless, performing incivility detection at scale, we are able to characterise its distribution across individual users and geopolitical regions, where our findings align and extend existing theories of political communication. In particular, we find that roughly 80% of the uncivil tweets are authored by 20% of the users, where users who are politically engaged are more inclined to use uncivil language. We further find that political incivility exhibits network homophily, and that incivility is more prominent in highly competitive geopolitical regions. Our results apply to both uncivil style and substance.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach

Yanchen Liu, Mingyu Derek Ma, Wennia Qin, Azure Zhou, Jiaao Chen, Weiyian Shi, Wei Wang, Diyi Yang

Susceptibility to misinformation describes the degree of belief in unverifiable claims, a latent aspect of individuals' mental processes that is not observable. Existing susceptibility studies heavily rely on self-reported beliefs, which can be subject to bias, expensive to collect, and challenging to scale for downstream applications. To address these limitations, in this work, we propose a computational approach to efficiently model users' latent susceptibility levels. As shown in previous work, susceptibility is influenced by various factors (e.g., demographic factors, political ideology), and directly influences people's reposting behavior on social media. To represent the underlying mental process, our susceptibility modeling incorporates these factors as inputs, guided by the supervision of people's sharing behavior. Using COVID-19 as a testbed, our experiments demonstrate a significant alignment between the susceptibility scores estimated by our computational modeling and human judgments, confirming the effectiveness of this latent modeling approach. Furthermore, we apply our model to annotate susceptibility scores on a large-scale dataset and analyze the relationships between susceptibility with various factors. Our analysis reveals that political leanings and other psychological factors exhibit varying degrees of association with susceptibility to COVID-19 misinformation, and shows that susceptibility is unevenly distributed across different professional and geographical backgrounds.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Adaptive Axes: A Pipeline for In-domain Social Stereotype Analysis

Qingcheng Zeng, Mingyu Jin, Rob Voigt

Prior work has explored the possibility of using the semantic information obtained from embedding representations to quantify social stereotypes, leveraging techniques such as word embeddings combined with a list of traits (Garg et al., 2018; Charlesworth et al., 2022) or semantic axes (An et al., 2018; Lucy et al., 2022). However, these approaches have struggled to fully capture the variability in stereotypes across different conceptual domains for the same social group (e.g., black in science, health, and art), in part because the identity of a word and the associations formed during pre-training can dominate its contextual representation (Field and Tsvetkov, 2019). This study explores the ability to recover stereotypes from the contexts surrounding targeted entities by utilizing state-of-the-art text embedding models and adaptive semantic axes enhanced by large language models (LLMs). Our results indicate that the proposed pipeline not only surpasses token-based methods in capturing in-domain framing but also effectively tracks stereotypes over time and along domain-specific semantic axes for in-domain texts. Our research highlights the potential of employing text embedding models to achieve a deeper understanding of nuanced social stereotypes.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Towards Measuring and Modeling Culture* in LLMs: A Survey

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivedutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, Monojit Choudhury

We present a survey of more than 90 recent papers that aim to study cultural representation and inclusion in large language models (LLMs). We observe that none of the studies explicitly define “culture,” which is a complex, multifaceted concept; instead, they probe the models on some specially designed datasets which represent certain aspects of “culture.” We call these aspects the proxies of culture, and organize them across two dimensions of demographic and semantic proxies. We also categorize the probing methods employed. Our analysis indicates that only certain aspects of “culture,” such as values and objectives, have been studied, leaving several other interesting and important facets, especially the multitude of semantic domains (Thompson et al., 2020) and aboutness (Herscovitch et al., 2022), unexplored. Two other crucial gaps are the lack of robustness of probing techniques and situated studies on the impact of cultural mis- and under-representation in LLM-based applications.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Cultural Conditioning or Placebo? On the Effectiveness of Socio-Demographic Prompting

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, Monojit Choudhury

Socio-demographic prompting is a commonly employed approach to study cultural biases in LLMs as well as for aligning models to certain cultures. In this paper, we systematically probe four LLMs (Llama 3, Mistral v0.2, GPT-3.5 Turbo and GPT4) with prompts that are conditioned on culturally sensitive and non-sensitive cues, on datasets that are supposed to be culturally sensitive (EtiCor and CALI) or neutral (MMLU and ETHICS). We observe that all models except GPT4 show significant variations in their responses on both kinds of datasets for both kinds of prompts, casting doubt on the robustness of the culturally-conditioned prompting as a method for eliciting cultural bias in models that are not sufficiently stable with respect to arbitrary prompting cues. Further, we also show that some of the supposedly culturally neutral datasets have a non-trivial fraction of culturally sensitive questions/tasks.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

How Susceptible are Large Language Models to Ideological Manipulation?

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, Kristina Lerman

Large Language Models (LLMs) possess the potential to exert substantial influence on public perceptions and interactions with information. This raises concerns about the societal impact that could arise if the ideologies within these models can be easily manipulated. In this work, we investigate how effectively LLMs can learn and generalize ideological biases from their instruction-tuning data. Our findings reveal a concerning vulnerability: exposure to only a small amount of ideologically driven samples significantly alters the ideology of LLMs. Notably, LLMs demonstrate a startling ability to absorb ideology from one topic and generalize it to even unrelated ones. The ease with which LLMs ideologies can be skewed underscores the risks associated with intentionally poisoned training data by malicious actors or inadvertently introduced biases by data annotators. It also emphasizes the imperative for robust safeguards to mitigate the influence of ideological manipulations on LLMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Media Attitude Detection via Framing Analysis with Events and their Relations

Jin Zhao, Jingxuan Tu, Han Du, Nianwen Xue

Framing is used to present some selective aspects of an issue and making them more salient, which aims to promote certain values, interpretations, or solutions (Entman, 1993). This study investigates the nuances of media framing on public perception and understanding by examining how events are presented within news articles. Unlike previous research that primarily focused on word choice as a framing device, this work explores the comprehensive narrative construction through events and their relations. Our method integrates event extraction, cross-document event coreference, and causal relationship mapping among events to extract framing devices employed by media to assess their role in framing the narrative. We evaluate our approach with a media attitude detection task and show that the use of event mentions, event cluster descriptors, and their causal relations effectively captures the subtle nuances of framing, thereby providing deeper insights into the attitudes conveyed by news articles. The experimental results show the framing device models surpass the baseline models and offers a more detailed and explainable analysis of media framing effects. We make the source code and dataset publicly available.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, Vinodkumar Prabhakaran

While human annotations play a crucial role in language technologies, annotator subjectivity has long been overlooked in data collection. Recent studies that critically examine this issue are often focused on Western contexts, and solely document differences across age, gender, or racial groups. Consequently, NLP research on subjectivity have failed to consider that individuals within demographic groups may hold diverse values, which influence their perceptions beyond group norms. To effectively incorporate these considerations into NLP pipelines, we need datasets with extensive parallel annotations from a variety of social and cultural groups. In this paper we introduce the D3CODE dataset: a large-scale cross-cultural dataset of parallel annotations for offensive language in over 4.5K English sentences annotated by a pool of more than 4k annotators, balanced across gender and age, from across 21 countries, representing eight geo-cultural regions. The dataset captures annotators' moral values along six moral foundations: care, equality, proportionality, authority, loyalty, and purity. Our analyses reveal substantial regional variations in annotators' perceptions that are shaped by individual moral values, providing crucial insights for developing pluralistic, culturally sensitive NLP models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Virtual Personas for Language Models via an Anthology of Backstories

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widjadewi Soedarmadij, Eran Kohen Behar, David Chan

Large language models (LLMs) are trained from vast repositories of text authored by millions of distinct authors, reflecting an enormous diversity of human traits. While these models bear the potential to be used as approximations of human subjects in behavioral studies, prior efforts have been limited in steering model responses to match individual human users. In this work, we introduce Anthology, a method for conditioning LLMs to particular virtual personas by harnessing open-ended life narratives, which we refer to as backstories. We show that our methodology enhances the consistency and reliability of experimental outcomes while ensuring better representation of diverse subpopulations. Across three nationally representative human surveys conducted as part of Pew Research Center's American Trends Panel (ATP), we demonstrate that Anthology achieves up to 18% improvement in matching the response distributions of human respondents and 27% improvement in consistency metrics.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Promoting Constructive Deliberation: Reframing for Receptiveness

Gauri Kambhatla, Matthew Lease, Ashwin Rajadesigan

To promote constructive discussion of controversial topics online, we propose automatic reframing of disagreeing responses to signal receptiveness to a preceding comment. Drawing on research from psychology, communications, and linguistics, we identify six strategies for reframing. We automatically reframe replies to comments according to each strategy, using a Reddit dataset. Through human-centered experiments, we find that the replies generated with our framework are perceived to be significantly more receptive than the original replies and a generic receptiveness baseline. We illustrate how transforming receptiveness, a particular social science construct, into a computational framework, can make LLM generations more aligned with human perceptions. We analyze and discuss the implications of our results, and highlight how a tool based on our framework might be used for more teachable and creative content moderation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Using RL to Identify Divisive Perspectives Improves LLMs Abilities to Identify Communities on Social Media

Nikhil Mehta, Dan Goldwasser

The large scale usage of social media, combined with its significant impact, has made it increasingly important to understand it. In particular, identifying user communities, can be helpful for many downstream tasks. However, particularly when models are trained on past data and tested on future, doing this is difficult. In this paper, we hypothesize to take advantage of Large Language Models (LLMs), to better identify user communities. Due to the fact that many LLMs, such as ChatGPT, are fixed and must be treated as black-boxes, we propose an approach to better prompt them, by training a smaller LLM to do this. We devise strategies to train this smaller model, showing how it can improve the larger LLMs ability to detect communities. Experimental results show improvements on Reddit and Twitter data, and the tasks of community detection, bot detection, and news media profiling.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Enabling Cross-Platform Comparison of Online Communities Using Content and Opinion Similarity

Prasanna Lakkur Subramanyam, Jeng-Yu Chou, Kevin K. Nam, Brian Levine

With the continuous growth of online communities, understanding their similarities and dissimilarities is more crucial than ever for enhancing digital interactions, maintaining healthy interactions, and improving content recommendation and moderation systems. In this work, we present two novel techniques: BOTS for finding similarity between online communities based on their opinion, and Emb-PSR for finding similarity in the content they post. To facilitate finding the similarity based on opinion, we model the opinions on online communities using upvotes and downvotes as an indicator for community approval. Our results demonstrate that BOTS and Emb-PSR outperform existing techniques at their individual tasks while also being flexible enough to allow for cross-platform comparison of online communities. We demonstrate this novel cross-platform capability by comparing GAB with various subreddits.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models

Yin Jou Huang, Rafik Hadfi

Psychological evidence reveals the influence of personality traits on decision-making. For instance, agreeableness is generally associated with positive outcomes in negotiations, whereas neuroticism is often linked to less favorable outcomes. This paper introduces a simulation framework centered on large language model (LLM) agents endowed with synthesized personality traits. The agents negotiate within bargaining domains and possess customizable personalities and objectives. The experimental results show that the behavioral tendencies of LLM-based simulations can reproduce behavioral patterns observed in human negotiations. The contribution is twofold. First, we propose a simulation methodology that investigates the alignment between the linguistic and economic capabilities of LLM agents. Secondly, we offer empirical insights into the strategic impacts of Big Five personality traits on the outcomes of bilateral negotiations. We also provide an in-depth analysis

based on simulated bargaining dialogues to reveal intriguing behaviors, including deceitful and compromising behaviors.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Mental Disorder Classification via Temporal Representation of Text

Raja Kumar, Kishan Maharaj, Ashita Saxena, Pushpak Bhattacharyya

Mental disorders pose a global challenge, aggravated by the shortage of qualified mental health professionals. Mental disorder prediction from social media posts by current LLMs is challenging due to the complexities of sequential text data and the limited context length of language models. Current language model-based approaches split a single data instance into multiple chunks to compensate for limited context size. The predictive model is then applied to each chunk individually, and the most voted output is selected as the final prediction. This results in the loss of inter-post dependencies and important time variant information, leading to poor performance. We propose a novel framework which first compresses the large sequence of chronologically ordered social media posts into a series of numbers. We then use this time variant representation for mental disorder classification. We demonstrate the generalization capabilities of our framework by outperforming the current SOTA in three different mental conditions: depression, self-harm, and anorexia, by an absolute improvement of 5% in the F1 score. We also investigate the situation when current data instances fall within the context length of language models and present empirical results highlighting the importance of temporal properties of textual data. Furthermore, we utilize the proposed framework for a cross-domain study, exploring commonalities across disorders and the possibility of inter-domain data usage.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Explaining Mixtures of Sources in News Articles

Alexander Spangher, James Youn, Matt DeButts, Nanyun Peng, Jonathan May

Human writers plan, _then_ write. For large language models (LLMs) to play a role in longer-form article generation, we must understand the planning steps humans make before writing. We explore one kind of planning, source-selection in news, as a case-study for evaluating plans in long-form generation. We ask: why do _specific_ stories call for _specific_ kinds of sources? We imagine a generative process for story writing where a source-selection schema is first selected by a journalist, and then sources are chosen based on categories in that schema. Learning the article's _plan_ means predicting the schema initially chosen by the journalist. Working with professional journalists, we adapt five existing schemata and introduce three new ones to describe journalistic plans for the inclusion of sources in documents. Then, inspired by Bayesian latent-variable modeling, we develop metrics to select the most likely plan, or schema, underlying a story, which we use to compare schemata. We find that two schemata: `_stance_` and `_social affiliation_` best explain source plans in most documents. However, other schemata like `_textual entailment_` explain source plans in factually rich topics like "Science". Finally, we find we can predict the most suitable schema given just the article's headline with reasonable accuracy. We see this as an important case-study for human planning, and provides a framework and approach for evaluating other kinds of plans, like discourse or plot-oriented plans. We release a corpora, `_News-Sources_`, with annotations for 4M articles, for further study.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Reap the Wild Wind: Detecting Media Storms in Large-Scale News Corpora

Dror Kris Markus, Effi Levi, Tamir Shefer, Shaul Rafael Shenhar

Media storms, dramatic outbursts of attention to a story, are central components of media dynamics and the attention landscape. Despite their importance, there has been little systematic and empirical research on this concept due to issues of measurement and operationalization. We introduce an iterative human-in-the-loop method to identify media storms in a large-scale corpus of news articles. The text is first transformed into signals of dispersion based on several textual characteristics. In each iteration, we apply unsupervised anomaly detection to these signals; each anomaly is then validated by an expert to confirm the presence of a storm, and those results are then used to tune the anomaly detection in the next iteration. We make available the resulting media storm dataset. Both the method and dataset provide a basis for comprehensive empirical study of media storms.

Demo

Nov 12 (Tue) 16:00-17:30 - Room: Riverfront Hall

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

RAGViz: Diagnose and Visualize Retrieval-Augmented Generation

Chenyan Xiong, Jingyuan He, Tevin Wang

Retrieval-augmented generation (RAG) combines knowledge from domain-specific sources into large language models to ground answer generation. Current RAG systems lack customizable visibility on the context documents and the model's attentiveness towards such documents. We propose RAGViz, a RAG diagnosis tool that visualizes the attentiveness of the generated tokens in retrieved documents. With a built-in user interface, retrieval index, and Large Language Model (LLM) backbone, RAGViz provides two main functionalities: (1) token and document-level attention visualization and (2) generation comparison when adding and removing documents. We host a system demonstration of RAGViz configured with AnchorDR as the embedding model, Pile-CC training split as the data store, and LLaMa2-7B as the LLM backbone, with each of the above settings being customizable. The open source codebase is available at [https://github.com/Tevin-Wang/ragviz]([https://github.com/TevinWang/ragviz]). The demo video is shown at [https://youtu.be/cTAbuTu6ur4]([https://youtu.be/cTAbuTu6ur4]). A live demonstration is available at [https://boston.liu.cs.cmu.edu/tevinw/ragviz/ui/]([https://boston.liu.cs.cmu.edu/tevinw/ragviz/ui/]). To access it, please use the API key *r0GkTCjVRDkc083IKQzklOIluXB51EMV8*. RAGViz is performant, as the total query time has a median of around 5 seconds for small queries with an average token length of 11.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Lighthouse: A User-Friendly Library for Reproducible Video Moment Retrieval and Highlight Detection

Hokuto Munakata, Shota Nakada, Taichi Nishimura, Tatsuya Komatsu

We propose Lighthouse, a user-friendly library for reproducible video moment retrieval and highlight detection (MR-HD). Although researchers proposed various MR-HD approaches, the research community holds two main issues. The first is a lack of comprehensive and reproducible experiments across various methods, datasets, and video-text features. This is because no unified training and evaluation codebase covers multiple settings. The second is user-unfriendly design. Because previous works use different libraries, researchers set up individual environments. In addition, most works release only the training codes, requiring users to implement the whole inference process of MR-HD. Lighthouse addresses these issues by implementing a unified reproducible codebase that includes six models, three features, and five datasets. In addition, it provides an inference API and web demo to make these methods easily accessible for researchers and developers. Our experiments demonstrate that Lighthouse generally reproduces the reported scores in the reference papers. The code is available at https://github.com/line/lighthouse.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

BattleAgent: Multi-modal Dynamic Emulation on Historical Battles to Complement Historical Analysis

Che-Jui Chang, Hang Hua, Lingya Li, Lizhou Fan, Mingyu Jin, Shuhang Lin, Wenyue Hua, Yongfeng Zhang, Jianchao Ji, Jiebo Luo

This paper presents **BattleAgent**, a detailed emulation demonstration system that combines the Large Vision-Language Model (VLM) and Multi-Agent System (MAS). This novel system aims to emulate complex dynamic interactions among multiple agents, as well as between agents and their environments, over a period of time. The emulation showcases the current capabilities of agents, featuring fine-grained multi-modal interactions between agents and landscapes. It develops customizable agent structures to meet specific situational requirements, for example, a variety of battle-related activities like scouting and trench digging. These components collaborate to recreate historical events in a lively and comprehensive manner. This methodology holds the potential to substantially improve visualization of historical events and deepen our understanding of historical events especially from the perspective of decision making. The data and code for this project are accessible at <https://github.com/agiresearch/battleagent> and the demo is accessible at <https://drive.google.com/file/d/115B3KWiYCSSPluMiPGNnXlTmild-MzRJ/view?usp=sharing>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

ULLME: A Unified Framework for Large Language Model Embeddings with Generation-Augmented Learning

Franck Dernoncourt, Hieu Man, Nghia Trung Ngo, Thien Huu Nguyen

Large Language Models (LLMs) excel in various natural language processing tasks, but leveraging them for dense passage embedding remains challenging. This is due to their causal attention mechanism and the misalignment between their pre-training objectives and the text ranking tasks. Despite some recent efforts to address these issues, existing frameworks for LLM-based text embeddings have been limited by their support for only a limited range of LLM architectures and fine-tuning strategies, limiting their practical application and versatility. In this work, we introduce the Unified framework for Large Language Model Embedding (ULLME), a flexible, plug-and-play implementation that enables bidirectional attention across various LLMs and supports a range of fine-tuning strategies. We also propose Generation-augmented Representation Learning (GRL), a novel fine-tuning method to boost LLMs for text embedding tasks. GRL enforces consistency between representation-based and generation-based relevance scores, leveraging LLMs' powerful generative abilities for learning passage embeddings. To showcase our framework's flexibility and effectiveness, we release three pre-trained models from ULLME with different backbone architectures, ranging from 1.5B to 8B parameters, all of which demonstrate strong performance on the Massive Text Embedding Benchmark. Our framework is publicly available at: <https://github.com/nlp-uoregon/ullme>. A demo video for ULLME can also be found at <https://rb.gy/ws11e>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

To the Globe (TTG): Towards Language-Driven Guaranteed Travel Planning

Aaron Foss, Andrew Cohen, Arman Zharmagambetov, Brandon Amos, Da Ju, Justine T Kao, Maryam Fazel-Zarandi, Sasha Mitts, Song Jiang, Xian Li, Yuandong Tian

Travel planning is a challenging and time-consuming task that aims to find an itinerary which satisfies multiple, interdependent constraints regarding flights, accommodations, attractions, and other travel arrangements. In this paper, we propose To the Globe (TTG), a real-time demo system that takes natural language requests from users, translates it to symbolic form via a fine-tuned Large Language Model, and produces optimal travel itineraries with Mixed Integer Linear Programming solvers. The overall system takes 5 seconds to reply to the user request with guaranteed itineraries. To train TTG, we develop a synthetic data pipeline that generates user requests, flight and hotel information in symbolic form without human annotations, based on the statistics of real-world datasets, and fine-tune an LLM to translate NL user requests to their symbolic form, which is sent to the symbolic solver to compute optimal itineraries. Our NL-symbolic translation achieves 91% exact match in a backtranslation metric (i.e., whether the estimated symbolic form of generated natural language matches the groundtruth), and its returned itineraries have a ratio of 0.979 compared to the optimal cost of the ground truth user request. When evaluated by users, TTG achieves consistently high Net Promoter Scores (NPS) of 35-40% on generated itinerary.

Machine Translation 2

Nov 12 (Tue) 16:00-17:30 - Room: Riverfront Hall

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice?

Dawei Zhu, Pinchen Chen, Miaoan Zhang, Barry Haddow, Xiaoyu Shen, Dietrich Klakow

Traditionally, success in multilingual machine translation can be attributed to three key factors in training data: large volume, diverse translation directions, and high quality. In the current practice of fine-tuning large language models (LLMs) for translation, we revisit the importance of these factors. We find that LLMs display strong translation capability after being fine-tuned on as few as 32 parallel sentences and that fine-tuning on a single translation direction enables translation in multiple directions. However, the choice of direction is critical: fine-tuning LLMs with only English on the target side can lead to task misinterpretation, which hinders translation into non-English languages. Problems also arise when noisy synthetic data is placed on the target side, especially when the target language is well-represented in LLM pre-training. Yet interestingly, synthesized data in an under-represented language has a less pronounced effect. Our findings suggest that when adapting LLMs to translation, the requirement on data quantity can be eased but careful considerations are still crucial to prevent an LLM from exploiting unintended data biases.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Aligning Translation-Specific Understanding to General Understanding in Large Language Models

Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, Bing Qin

Large Language models (LLMs) have exhibited remarkable abilities in understanding complex texts, offering a promising path towards human-like translation performance. However, this study reveals the misalignment between the translation-specific understanding and the general understanding inside LLMs. This understanding misalignment leads to LLMs mistakenly or literally translating some complicated concepts that they accurately comprehend in the general scenarios (e.g., QA). To align the translation-specific understanding to the general one, we propose a novel translation process, DUAT (Difficult words Understanding Aligned Translation), explicitly incorporating the general understanding on the complicated content incurring inconsistent understandings to guide the translation. Specifically, DUAT performs cross-lingual interpretation for the difficult-to-translate words and enhances the translation with the generated interpretations. Furthermore, we reframe the external tools to improve DUAT in detecting difficult words and generating helpful interpretations. We conduct experiments on the self-constructed benchmark Challenge-WMT, consisting of samples that are prone to mistranslation. Human evaluation results on high-resource and low-resource language pairs indicate that DUAT significantly facilitates the understanding alignment, which improves the translation quality (up to +3.85 COMET) and reduces translation literalness by -25% -51%.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

MMTE: Corpus and Metrics for Evaluating Machine Translation Quality of Metaphorical Language

Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, Chenghua Lin

Machine Translation (MT) has developed rapidly since the release of Large Language Models and current MT evaluation is performed through comparison with reference human translations or by predicting quality scores from human-labeled data. However, these mainstream evaluation methods mainly focus on fluency and factual reliability, whilst paying little attention to figurative quality. In this paper, we investigate the figurative quality of MT and propose a set of human evaluation metrics focused on the translation of figurative language. We additionally present a multilingual parallel metaphor corpus generated by post-editing. Our evaluation protocol is designed to estimate four aspects of MT: Metaphorical Equivalence, Emotion, Authenticity, and Quality. In doing so, we observe that translations of figurative expressions display different traits from literal ones.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Domain adapted machine translation: What does catastrophic forgetting forget and why?

Danielle Saunders, Steve DeNefele

Neural Machine Translation (NMT) models can be specialized by domain adaptation, often involving fine-tuning on a dataset of interest. This process risks catastrophic forgetting: rapid loss of generic translation quality. Forgetting has been widely observed, with many mitigation methods proposed. However, the causes of forgetting and the relationship between forgetting and adaptation data are underexplored. This paper takes a novel approach to understanding catastrophic forgetting during NMT adaptation by investigating the impact of the data. We provide a first investigation of what is forgotten, and why. We examine the relationship between forgetting and the in-domain data, and show that the amount and type of forgetting is linked to that data's target vocabulary coverage. Our findings pave the way toward better informed NMT domain adaptation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Can Automatic Metrics Assess High-Quality Translations?

Sweta Agrawal, António Farinhas, Ricardo Rei, André Martins

Automatic metrics for evaluating translation quality are typically validated by measuring how well they correlate with human assessments. However, correlation methods tend to capture only the ability of metrics to differentiate between good and bad source-translation pairs, overlooking their reliability in distinguishing alternative translations for the same source. In this paper, we confirm that this is indeed the case by showing that current metrics are insensitive to nuanced differences in translation quality. This effect is most pronounced when the quality is high and the variance among alternatives is low. Given this finding, we shift towards detecting high-quality correct translations, an important problem in practical decision-making scenarios where a binary check of correctness is prioritized over a nuanced evaluation of quality. Using the MQM framework as the gold standard, we systematically stress-test the ability of current metrics to identify translations with no errors as marked by humans. Our findings reveal that current metrics often over- or underestimate translation quality, indicating significant room for improvement in machine translation evaluation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Modeling User Preferences with Automatic Metrics: Creating a High-Quality Preference Dataset for Machine Translation

Bowen Xing, Lizi Liao, Minlie Huang, Ivor Tsang

Alignment with human preferences is an important step in developing accurate and safe large language models. This is no exception in machine translation (MT), where better handling of language nuances and context-specific variations leads to improved quality. However, preference data based on human feedback can be very expensive to obtain and curate at a large scale. Automatic metrics, on the other hand, can induce preferences, but they might not match human expectations perfectly. In this paper, we propose an approach that leverages the best of both worlds. We first collect sentence-level quality assessments from professional linguists on translations generated by multiple high-quality MT systems and evaluate the ability of current automatic metrics to recover these preferences. We then use this analysis to curate a new dataset, MT-Pref (metric induced translation preference) dataset, which comprises 18k instances covering 18 language directions, using texts sourced from multiple domains post-2022. We show that aligning TOWER models on MT-Pref significantly improves translation quality on WMT23 and FLORES benchmarks.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Ladder: A Model-Agnostic Framework Boosting LLM-based Machine Translation to the Next Level

Zhaopeng Feng, Ruijie Chen, Yan Zhang, Zijie Meng, Zuozhu Liu

General-purpose Large Language Models (LLMs) like GPT-4 have achieved remarkable advancements in machine translation (MT) by leveraging extensive web content. On the other hand, translation-specific LLMs are built by pre-training on domain-specific monolingual corpora and fine-tuning with human-annotated translation data. Despite the superior performance, these methods either demand an unpreceded scale of computing and data or substantial human editing and annotation efforts. In this paper, we develop MT-Ladder, a novel model-agnostic and cost-effective tool to refine the performance of general LLMs for MT. MT-Ladder is trained on pseudo-refinement triplets which can be easily obtained from existing LLMs without additional human cost. During training, we propose a hierarchical fine-tuning strategy with an easy-to-hard schema, improving MT-Ladder's refining performance progressively. The trained MT-Ladder can be seamlessly integrated with any general-purpose LLMs to boost their translation performance. By utilizing Gemma-2B/7B as the backbone, MT-Ladder-2B can elevate raw translations to the level of top-tier open-source models (e.g., refining BigTranslate-13B with +6.91 BLEU and +3.52 COMET for XXEn), and MT-Ladder-7B can further enhance model performance to be on par with the state-of-the-art GPT-4. Extensive ablation and analysis corroborate the effectiveness of MT-Ladder in diverse settings.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Mitigating the Language Mismatch and Repetition Issues in LLM-based Machine Translation via Model Editing

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Lingqi Song, Ying Wei

Large Language Models (LLMs) have recently revolutionized the NLP field, while they still fall short in some specific down-stream tasks. In the work, we focus on utilizing LLMs to perform machine translation, where we observe that two patterns of errors frequently occur and drastically affect the translation quality: language mismatch and repetition. The work sets out to explore the potential for mitigating these two issues by leveraging model editing methods, e.g., by locating Feed-Forward Network (FFN) neurons or something that are responsible for the errors and deactivating them in the inference time. We find that directly applying such methods either limited effect on the targeted errors or has significant negative side-effect on the general translation quality, indicating that the located components may also be crucial for ensuring machine translation with LLMs on the rails. To this end, we propose to refine the located components by fetching the intersection of the locating results under different language settings, filtering out the aforementioned information that is irrelevant to targeted errors. The experiment results empirically demonstrate that our methods can effectively reduce the language mismatch and repetition ratios and meanwhile enhance or keep the general translation quality in most cases.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Simultaneous Masking, Not Prompting Optimization: A Paradigm Shift in Fine-tuning LLMs for Simultaneous Translation

Matthew Raffel, Victor Agostinelli, Lihong Chen

Large language models (LLMs) have achieved state-of-the-art performance in various language processing tasks, motivating their adoption in simultaneous translation. Current fine-tuning methods to adapt LLMs for simultaneous translation focus on prompting optimization strategies using either data augmentation or prompt structure modifications. However, these methods suffer from several issues, such as unnecessarily expanded training sets, computational inefficiency from dumping the key and value cache, increased prompt sizes, or restriction to a single decision policy. To eliminate these issues, in this work, we propose SimulMask, a new paradigm for fine-tuning LLMs for simultaneous translation. It utilizes a novel attention mask approach that models simultaneous translation during fine-tuning by masking attention for a desired decision policy. Applying the proposed SimulMask on a Falcon LLM for the IWSLT 2017 dataset, we have observed a significant translation quality improvement compared to state-of-the-art prompting optimization strategies on five language pairs while reducing the computational cost.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Enhanced Hallucination Detection in Neural Machine Translation through Simple Detector Aggregation

Anas Himm, Guillaume Staerman, Marine Picot, Pierre Colombo, Nuno M Guerreiro

Hallucinated translations pose significant threats and safety concerns when it comes to practical deployment of machine translation systems. Previous research works have identified that detectors exhibit complementary performance — different detectors excel at detecting different types of hallucinations. In this paper, we propose to address the limitations of individual detectors by combining them and introducing a straightforward method for aggregating multiple detectors. Our results demonstrate the efficacy of our aggregated detector, providing a promising step towards evermore reliable machine translation systems.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Error Analysis of Multilingual Language Models in Machine Translation for Low-resource Languages: A Case Study of Amharic to English Bi-directional Machine Translation

Hizkel Mitku Alemayehu, Hamada M Zahera, Axel-Cyrille Ngonga Ngomo

Multilingual large language models (mLLMs) have significantly advanced machine translation, yet challenges remain for low-resource languages like Amharic. This study evaluates the performance of state-of-the-art mLLMs, specifically NLLB-200 (NLLB3.3, NLLB1.3 Distilled1.3, NLB600) and M2M (M2M1.2B, M2M418), in English-Amharic bidirectional translation using the Lesan AI dataset. We employed both automatic and human evaluation methods to analyze translation errors. Automatic evaluation used BLEU, METEOR, chRF, and TER metrics, while human evaluation assessed translation quality at both word and sentence levels. Sentence-level accuracy was rated by annotators on a scale from 0 to 5, and word-level quality was evaluated using Multidimensional Quality Metrics. Our findings indicate that the NLLB3.3B model consistently outperformed other mLLMs across all evaluation methods. Common error included mistranslation, omission, untranslated segments, and additions, with mistranslation being particularly common. Punctuation and spelling errors were rare in our experiment.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Back to School: Translation Using Grammar Books

Jonathan Hus, Antonios Anastasopoulos

Machine translation systems for high resource languages perform exceptionally well and produce high quality translations. Unfortunately, the vast majority of languages are not considered high resource and lack the quantity of parallel sentences needed to train such systems. These under-represented languages are not without resources, however, and bilingual dictionaries and grammar books are available as linguistic reference material. With current large language models (LLMs) supporting near book-length contexts, we can begin to use the available material to ensure advancements are shared among all of the world's languages. In this paper, we demonstrate incorporating grammar books in the prompt of GPT-4 to improve machine translation and evaluate the performance on 16 topologically diverse low-resource languages, using a combination of reference material to show that the machine translation performance of LLMs can be improved using this method.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

How Good is my MT Metric? A Framework for the Interpretation of Metric Assessments

Stefano Perella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, Roberto Navigli

Machine Translation (MT) evaluation metrics assess translation quality automatically. Recently, researchers have employed MT metrics for various new use cases, such as data filtering and translation re-ranking. However, most MT metrics return assessments as scalar scores that are difficult to interpret, posing a challenge to making informed design choices. Moreover, MT metrics' capabilities have historically been evaluated using correlation with human judgment, which, despite its efficacy, falls short of providing intuitive insights into metric performance, especially in terms of new metric use cases. To address these issues, we introduce an interpretable evaluation framework for MT metrics. Within this framework, we evaluate metrics in two scenarios that serve as proxies for the data filtering and translation re-ranking use cases. Furthermore, by measuring the performance of MT metrics using Precision, Recall, and F-score, we offer clearer insights into their capabilities than correlation with human judgments. Finally, we raise concerns regarding the reliability of manually curated data following the Direct Assessments+Scalar Quality Metrics (DA+SQM) guidelines, reporting a notably low agreement with Multidimensional Quality Metrics (MQM) annotations.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Exploring Intrinsic Language-specific Subspaces in Fine-tuning Multilingual Neural Machine Translation

Zhe Cao, Zhi Qu, Hirotaka Kamigaito, Taro Watanabe

Multilingual neural machine translation models support fine-tuning hundreds of languages simultaneously. However, fine-tuning on full parameters solely is inefficient potentially leading to negative interactions among languages. In this work, we demonstrate that the fine-tuning for a language occurs in its intrinsic language-specific subspace with a tiny fraction of entire parameters. Thus, we propose language-specific LoRA to isolate intrinsic language-specific subspaces. Furthermore, we propose architecture learning techniques and introduce a gradual pruning schedule during fine-tuning to exhaustively explore the optimal setting and the minimal intrinsic subspaces for each language, resulting in a lightweight yet effective fine-tuning procedure. The experimental results on a 12-language subset and a 30-language subset of FLORES-101 show that our methods not only outperform full-parameter fine-tuning up to 2.25 spBLEU scores but also reduce trainable parameters to 0.4% for high and medium-resource languages and 1.6% for low-resource ones.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SpeechQE: Estimating the Quality of Direct Speech Translation

Hyojung Han, Kevin Duh, Marine Carpuat

Recent advances in automatic quality estimation for machine translation have exclusively focused on written language, leaving the speech modality underexplored. In this work, we formulate the task of quality estimation for speech translation (SpeechQE), construct a benchmark, and evaluate a family of systems based on cascaded and end-to-end architectures. In this process, we introduce a novel end-to-end system leveraging pre-trained text LLMs. Results suggest that end-to-end approaches are better suited to estimating the quality of direct speech translation than using quality estimation systems designed for text in cascaded systems. More broadly, we argue that quality estimation of speech translation needs to be studied as a separate problem from that of text, and release our [data and models](<https://github.com/h-j-h/SpeechQE>)

han/SpeechQE) to guide further research in this space.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Simultaneous Interpretation Corpus Construction by Large Language Models in Distant Language Pair

Yusuke Sakai, Mana Makinae, Hidetaka Kamigaito, Taro Watanabe

In Simultaneous Machine Translation (SiMT), training with a simultaneous interpretation (SI) corpus is an effective method for achieving high-quality yet low-latency. However, constructing such a corpus is challenging due to high costs, and limitations in annotator capabilities, and as a result, existing SI corpora are limited. Therefore, we propose a method to convert existing speech translation (ST) corpora into interpretation-style corpora, maintaining the original word order and preserving the entire source content using Large Language Models (LLM-SI-Corpus). We demonstrate that fine-tuning SiMT models using the LLM-SI-Corpus reduces latency while achieving better quality compared to models fine-tuned with other corpora in both speech-to-text and text-to-text settings. The LLM-SI-Corpus is available at <https://github.com/yusuke1997/LLM-SI-Corpus>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

A Deep Analysis of the Impact of Multiword Expressions and Named Entities on Chinese-English Machine Translations

Huacheng Song, Hongchi Xu

In this paper, we present a study on the impact of so-called multiword expressions (MWEs) and multiword named entities (NEs) on the performance of Chinese-English machine translation (MT) systems. Built on an extended version of the data from the WMT22 Metrics Shared Task (with extra labels of 9 types of Chinese MWEs, and 19 types of Chinese multiword NEs) which includes scores and error annotations provided by human experts, we make further extraction of MWE- and NE-related translation errors. By investigating the evaluation scores and the error rates on each category of MWEs and NEs, we find that: 1) MT systems tend to perform significantly worse on Chinese sentences with most kinds of MWEs and NEs; 2) MWEs and NEs which make up of about twenty percent of tokens, i.e. characters in Chinese, result in one-third of translation errors; 3) for 13 categories of MWEs and NEs, the error rates exceed 50% with the highest to be 84.8%. Based on the results, we emphasize that MWEs and NEs are still a bottleneck issue for MT and special attention to MWEs and NEs should be paid to further improving the performance of MT systems.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Cross-lingual Contextualized Phrase Retrieval

Huayang Li, Deng Cai, Zhi Qu, Qu Cui, Hidetaka Kamigaito, Lemao Liu, Taro Watanabe

Phrase-level dense retrieval has shown many appealing characteristics in downstream NLP tasks by leveraging the fine-grained information that phrases offer. In our work, we propose a new task formulation of dense retrieval, cross-lingual contextualized phrase retrieval, which aims to augment cross-lingual applications by addressing polysemy using context information. However, the lack of specific training data and models are the primary challenges to achieve our goal. As a result, we extract pairs of cross-lingual phrases using word alignment information automatically induced from parallel sentences. Subsequently, we train our Cross-lingual Contextualized Phrase Retriever (CCPR) using contrastive learning, which encourages the hidden representations of phrases with similar contexts and semantics to align closely. Comprehensive experiments on both the cross-lingual phrase retrieval task and a downstream task, i.e., machine translation, demonstrate the effectiveness of CCPR. On the phrase retrieval task, CCPR surpasses baselines by a significant margin, achieving a top-1 accuracy that is at least 13 points higher. When utilizing CCPR to augment the large-language-model-based translator, it achieves average gains of 0.7 and 1.5 in BERTScore for translations from X=>En and vice versa, respectively, on WMT16 dataset. We will release our code and data.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Low-Resource Machine Translation through the Lens of Personalized Federated Learning

Viktor Moskvetzskii, Nazarii Tuptitsa, Chris Biemann, Samuel Horváth, Eduard Gorbanov, Irina Nikishina

We present a new approach called MeritOpt based on the Personalized Federated Learning algorithm MeritFed that can be applied to Natural Language Tasks with heterogeneous data. We evaluate it on the Low-Resource Machine Translation task, using the datasets of South East Asian and Fino-Ugric languages. In addition to its effectiveness, MeritOpt is also highly interpretable, as it can be applied to track the impact of each language used for training. Our analysis reveals that target dataset size affects weight distribution across auxiliary languages, that unrelated languages do not interfere with the training, and auxiliary optimizer parameters have minimal impact. Our approach is easy to apply with a few lines of code, and we provide scripts for reproducing the experiments (<https://github.com/VityaVitalich/MeritOpt>).

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Creative and Context-Aware Translation of East Asian Idioms with GPT-4

Kenan Tang, Peiyang Song, Yao Qin, Xifeng Yan

As a type of figurative language, an East Asian idiom condenses rich cultural background into only a few characters. Translating such idioms is challenging for human translators, who often resort to choosing a context-aware translation from an existing list of candidates. However, compiling a dictionary of candidate translations demands much time and creativity even for expert translators. To alleviate such burden, we evaluate if GPT-4 can help generate high-quality translations. Based on automatic evaluations of faithfulness and creativity, we first identify Pareto-optimal prompting strategies that can outperform translation engines from Google and DeepL. Then, at a low cost, our context-aware translations can achieve far more high-quality translations per idiom than the human baseline. We open-source all code and data to facilitate further research.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Introducing Compiler Semantics into Large Language Models as Programming Language Translators: A Case Study of C to x86 Assembly

Shuoming Zhang, Jiacheng Zhao, Chunwei Xia, Zheng Wang, Yunji Chen, Huimin Cui

Compilers are complex software containing millions of lines of code, taking years to develop. This paper investigates to what extent Large Language Models (LLMs) can replace hand-crafted compilers in translating high-level programming languages to machine instructions, using C to x86 assembly as a case study. We identify two challenges of using LLMs for code translation and introduce two novel data pre-processing techniques to address the challenges: numerical value conversion and training data resampling. While only using a 13B model, our approach achieves a behavioral accuracy of over 91%, outperforming the much larger GPT-4 Turbo model by over 50%. Our results are encouraging, showing that LLMs have the potential to transform how compilation tools are constructed.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Together We Can: Multilingual Automatic Post-Editing for Low-Resource Languages

Sourabh Dattatray Deoghare, Diplesh Kanjola, Pushpak Bhattacharyya

This exploratory study investigates the potential of multilingual Automatic Post-Editing (APE) systems to enhance the quality of machine translations for low-resource Indo-Aryan languages. Focusing on two closely related language pairs, English-Marathi and English-Hindi, we exploit the linguistic similarities to develop a robust multilingual APE model. To facilitate cross-linguistic transfer, we generate synthetic Hindi-Marathi and Marathi-Hindi APE triplets. Additionally, we incorporate a Quality Estimation (QE)-APE multi-task learning frame-

work. While the experimental results underline the complementary nature of APE and QE, we also observe that QE-APE multitask learning facilitates effective domain adaptation. Our experiments demonstrate that the multilingual APE models outperform their corresponding English-Hindi and English-Marathi single-pair models by 2.5 and 2.39 TER points, respectively, with further notable improvements over the multilingual APE model observed through multi-task learning (+1.29 and +1.44 TER points), data augmentation (+0.53 and +0.45 TER points) and domain adaptation (+0.35 and +0.45 TER points). We release the synthetic data, code, and models accrued during this study publicly for further research.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Benchmarking Machine Translation with Cultural Awareness

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, Junjie Hu

Translating culture-related content is vital for effective cross-cultural communication. However, many culture-specific items (CSIs) often lack literal translation across languages, making it challenging to collect high-quality, diverse parallel corpora with CSI annotations. This difficulty hinders the analysis of cultural awareness of machine translation (MT) systems, including traditional neural MT and the emerging MT paradigm using large language models (LLM). To address this gap, we introduce a novel parallel corpus, enriched with CSI annotations in 6 language pairs for investigating Cultural-Aware Machine Translation—CAMT. Furthermore, we design two evaluation metrics to assess CSI translations, focusing on their pragmatic translation quality. Our findings show the superior ability of LLMs over neural MTs in leveraging external cultural knowledge for translating CSIs, especially those lacking translations in the target culture.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Does Context Help Mitigate Gender Bias in Neural Machine Translation?

Harrixa Gete, Thierry Eichegoyen

Neural Machine Translation models tend to perpetuate gender bias present in their training data distribution. Context-aware models have been previously suggested as a means to mitigate this type of bias. In this work, we examine this claim by analysing in detail the translation of stereotypical professions in English to German, and translation with non-informative context in Basque to Spanish. Our results show that, although context-aware models can significantly enhance translation accuracy for feminine terms, they can still maintain or even amplify gender bias. These results highlight the need for more fine-grained approaches to bias mitigation in Neural Machine Translation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Analyzing Context Contributions in LLM-based Machine Translation

Emmanouil Zarassis, Nuno M Guerreiro, Andre Martins

Large language models (LLMs) have achieved state-of-the-art performance in machine translation (MT) and demonstrated the ability to leverage in-context learning through few-shot examples. However, the mechanisms by which LLMs use different parts of the input context remain largely unexplored. In this work, we provide a comprehensive analysis of context utilization in MT, studying how LLMs use various context parts, such as few-shot examples and the source text, when generating translations. We highlight several key findings: (1) the source part of few-shot examples appears to contribute more than its corresponding targets, irrespective of translation direction; (2) finetuning LLMs with parallel data alters the contribution patterns of different context parts; and (3) there is a positional bias where earlier few-shot examples have higher contributions to the translated sequence. Finally, we demonstrate that inspecting anomalous context contributions can potentially uncover pathological translations, such as hallucinations. Our findings shed light on the internal workings of LLM-based MT which go beyond those known for standard encoder-decoder MT models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

xTower: A Multilingual LLM for Explaining and Correcting Translation Errors

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, Andre Martins

While machine translation (MT) systems are achieving increasingly strong performance on benchmarks, they often produce translations with errors and anomalies. Understanding these errors can potentially help improve the translation quality and user experience. This paper introduces xTower, an open large language model (LLM) built on top of TowerBase designed to provide free-text explanations for translation errors in order to guide the generation of a corrected translation. The quality of the generated explanations by xTower are assessed via both intrinsic and extrinsic evaluation. We ask expert translators to evaluate the quality of the explanations across two dimensions: relatedness towards the error span being explained and helpfulness in error understanding and improving translation quality. Extrinsicly, we test xTower across various experimental setups in generating translation corrections, demonstrating significant improvements in translation quality. Our findings highlight xTower's potential towards not only producing plausible and helpful explanations of automatic translations, but also leveraging them to suggest corrected translations.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation

David Dale, Marta R. Costa-jussà

We present BLASER 2.0, an automatic metric of machine translation quality which supports both speech and text modalities. Compared to its predecessor BLASER (Chen et al., 2023), BLASER 2.0 is based on better underlying text and speech representations that cover 202 text languages and 57 speech ones and extends the training data. BLASER 2.0 comes in two varieties: a reference-based and a reference-free (quality estimation) model. We demonstrate that the reference-free version is applicable not only at the dataset level, for evaluating the overall model performance, but also at the sentence level, for scoring individual translations. In particular, we show its applicability for detecting translation hallucinations and filtering training datasets to obtain more reliable translation models. The BLASER 2.0 models are publicly available at <https://github.com/facebookresearch/sonar>.

Question Answering 1

Nov 12 (Tue) 16:00-17:30 - Room: Riverfront Hall

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Table Question Answering for Low-resourced Indic Languages

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, Maarten de Rijke

TableQA is the task of answering questions over tables of structured information, returning individual cells or tables as output. TableQA research has focused primarily on high-resource languages, leaving medium- and low-resource languages with little progress due to scarcity of annotated data and neural models. We address this gap by introducing a fully automatic large-scale tableQA data generation process for low-resource languages with limited budget. We incorporate our data generation method on two Indic languages, Bengali and Hindi, which

have no tableQA datasets or models. TableQA models trained on our large-scale datasets outperform state-of-the-art LLMs. We further study the trained models on different aspects, including mathematical reasoning capabilities and zero-shot cross-lingual transfer. Our work is the first on low-resource tableQA focusing on scalable data generation and evaluation procedures. Our proposed data generation method can be applied to any low-resource language with a web presence. We release datasets, models, and code (<https://github.com/kolk/Low-Resource-TableQA-Indic-languages>).

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Learning Planning-based Reasoning by Trajectories Collection and Process Reward Synthesizing

Fangkai Jiao, Chengwei Qin, Zhengyuan Liu, Nancy F. Chen, Shafiq Joty

Large Language Models (LLMs) have demonstrated significant potential in handling complex reasoning tasks through step-by-step rationale generation. However, recent studies have raised concerns regarding the hallucination and flaws in their reasoning process. Substantial efforts are being made to improve the reliability and faithfulness of the generated rationales. Some approaches model reasoning as planning, while others focus on annotating for process supervision. Nevertheless, the planning-based search process often results in high latency due to the frequent assessment of intermediate reasoning states and the extensive exploration space. Additionally, supervising the reasoning process with human annotation is costly and challenging to scale for LLM training. To address these issues, in this paper, we propose a framework to learn planning-based reasoning through Direct Preference Optimization (DPO) on collected trajectories, which are ranked according to synthesized process rewards. Our results on challenging logical reasoning benchmarks demonstrate the effectiveness of our learning framework, showing that our 7B model can surpass the strong counterparts like GPT-3.5-Turbo.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models

Hongbang Yuan, Pengfei Cao, Zhioran Jin, Yubo Chen, Daojian Zeng, Kang Liu, Jun Zhao

Large Language Models (LLMs) have shown impressive capabilities but still suffer from the issue of hallucinations. A significant type of this issue is the false premise hallucination, which we define as the phenomenon when LLMs generate hallucinated text when confronted with false premise questions. In this paper, we perform a comprehensive analysis of the false premise hallucination and elucidate its internal working mechanism: a small subset of attention heads (which we designate as false premise heads) disturb the knowledge extraction process, leading to the occurrence of false premise hallucination. Based on our analysis, we propose FAITH (False premise Attention head constraining for mITigating Hallucinations), a novel and effective method to mitigate false premise hallucinations. It constrains the false premise attention heads during the model inference process. Impressively, extensive experiments demonstrate that constraining only approximately 1% of the attention heads in the model yields a notable increase of nearly 20% of model performance.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

CoTKR: Chain-of-Thought Enhanced Knowledge Rewriting for Complex Knowledge Graph Question Answering

Yike Wu, Yi Huang, Nan Hu, YUNCHENG HUA, Guolin Qi, Jiayuan Chen, Jeff Z. Pan

Recent studies have explored the use of Large Language Models (LLMs) with Retrieval Augmented Generation (RAG) for Knowledge Graph Question Answering (KGQA). They typically require rewriting retrieved subgraphs into natural language formats comprehensible to LLMs. However, when tackling complex questions, the knowledge rewritten by existing methods may include irrelevant information, omit crucial details, or fail to align with the question's semantics. To address them, we propose a novel rewriting method CoTKR, Chain-of-Thought Enhanced Knowledge Rewriting, for generating reasoning traces and corresponding knowledge in an interleaved manner, thereby mitigating the limitations of single-step knowledge rewriting. Additionally, to bridge the preference gap between the knowledge rewriter and the question answering (QA) model, we propose a training strategy PAQAF, Preference Alignment from Question Answering Feedback, for leveraging feedback from the QA model to further optimize the knowledge rewriter. We conduct experiments using various LLMs across several KGQA benchmarks. Experimental results demonstrate that, compared with previous knowledge rewriting methods, CoTKR generates the most beneficial knowledge representation for QA models, which significantly improves the performance of LLMs in KGQA.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Teaching LLMs to Abstain across Languages via Multilingual Feedback

Shangbin Feng, Weijia Shi, Yike Wang, Wenzuan Ding, Orenvaghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, Yulia Tsvetkov

Multilingual LLMs often have knowledge disparities across languages, with larger gaps in under-resourced languages. Teaching LLMs to abstain in the face of knowledge gaps is thus a promising strategy to mitigate hallucinations in multilingual settings. However, previous studies on LLM abstention primarily focus on English; we find that directly applying existing solutions beyond English results in up to 20.5% performance gaps between high and low-resource languages, potentially due to LLMs' drop in calibration and reasoning beyond a few resource-rich languages. To this end, we propose strategies to enhance LLM abstention by learning from multilingual feedback, where LLMs self-reflect on proposed answers in one language by generating multiple feedback items in related languages: we show that this helps identifying the knowledge gaps across diverse languages, cultures, and communities. Extensive experiments demonstrate that our multilingual feedback approach outperforms various strong baselines, achieving up to 9.2% improvement for low-resource languages across three black-box and open models on three datasets, featuring open-book, closed-book, and commonsense QA. Further analysis reveals that multilingual feedback is both an effective and a more equitable abstain strategy to serve diverse language speakers, and cultural factors have great impact on language selection and LLM abstention behavior, highlighting future directions for multilingual and multi-cultural reliable language modeling.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

RAG-QA Arena: Evaluating Domain Robustness for Long-form Retrieval Augmented Question Answering

Ruijun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jinyuan Wang, Lan Liu, William Yang Wang, Bonan Min, Vittorio Castelli

Question answering based on retrieval augmented generation (RAG-QA) is an important research topic in NLP and has a wide range of real-world applications. However, most existing datasets for this task are either constructed using a single source corpus or consist of short extractive answers, which fall short of evaluating large language model (LLM) based RAG-QA systems on cross-domain generalization. To address these limitations, we create Long-form RobustQA (LFRQA), a new dataset comprising human-written long-form answers that integrate short extractive answers from multiple documents into a single, coherent narrative, covering 26K queries and large corpora across seven different domains. We further propose RAG-QA Arena by directly comparing model-generated answers against LFRQA's answers using LLMs as evaluators. We show via extensive experiments that RAG-QA Arena and human judgments on answer quality are highly correlated. Moreover, only 41.3% of the most competitive LLM's answers are preferred to LFRQA's answers, demonstrating RAG-QA Arena as a challenging evaluation platform for future research.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering

Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, Ji-Rong Wen

Considering the limited internal parametric knowledge, retrieval-augmented generation (RAG) has been widely used to extend the knowledge scope of large language models (LLMs). Despite the extensive efforts on RAG research, in existing methods, LLMs cannot precisely assess

the relevance of retrieved documents, thus likely leading to misleading or even incorrect utilization of external knowledge (i.e., retrieved documents). To address this issue, in this paper, we propose REAR, a Relevance-Aware Retrieval-augmented approach for open-domain question answering (QA). As the key motivation, we aim to enhance the self-awareness regarding the reliability of external knowledge for LLMs, so as to adaptively utilize external knowledge in RAG systems. Specially, we develop a novel architecture for LLM based RAG system, by incorporating a specially designed assessment module that precisely assesses the relevance of retrieved documents. Furthermore, we propose an improved training method based on bi-granularity relevance fusion and noise-resistant training. By combining the improvements in both architecture and training, our proposed REAR can better utilize external knowledge by effectively perceiving the relevance of retrieved documents. Experiments on four open-domain QA tasks show that REAR significantly outperforms previous a number of competitive RAG approaches. Our codes can be accessed at <https://github.com/RUCAIBox/REAR>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering

Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, Scott Sanner

Knowledge Graph Question Answering (KGQA) methods seek to answer Natural Language questions using the relational information stored in Knowledge Graphs (KGs). With the recent advancements of Large Language Models (LLMs) and their remarkable reasoning abilities, there is a growing trend to leverage them for KGQA. However, existing methodologies have only focused on answering factual questions, e.g., “*In which city was Silvio Berlusconi’s first wife born?”*. leaving questions involving commonsense reasoning that real-world users may pose more often, e.g., “*Do I need separate visas to see the Venus of Willendorf and attend the Olympics this summer?*”*. unaddressed. In this work, we first observe that existing LLM-based methods for KGQA struggle with hallucination on such questions, especially on queries targeting long-tail entities (e.g., non-mainstream and recent entities), thus hindering their applicability in real-world applications especially since their reasoning processes are not easily verifiable. In response, we propose Right for Right Reasons (R^3), a commonsense KGQA methodology that allows for a verifiable reasoning procedure by axiomatically surfacing intrinsic commonsense knowledge of LLMs and grounding every factual reasoning step on KG triples. Through experimental evaluations across three different tasksquestion answering, claim verification, and preference matchingour findings showcase R^3 as a superior approach, outperforming existing methodologies and notably reducing instances of hallucination and reasoning errors.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

From RAG to Riches: Retrieval Interleaved with Sequence Generation

Palak Jain, Livio Baldini Soares, Tom Kwiatkowski

We present RICHES, a novel approach that interleaves retrieval with sequence generation tasks. RICHES offers an alternative to conventional RAG systems by eliminating the need for separate retriever and generator. It retrieves documents by directly decoding their contents, constrained on the corpus. Unifying retrieval with generation allows us to adapt to diverse new tasks via prompting alone. RICHES can work with any Instruction-tuned model, without additional training. It provides attributed evidence, supports multi-hop retrievals and interleaves thoughts to plan on what to retrieve next, all within a single decoding pass of the LLM. We demonstrate the strong performance of RICHES across ODQA tasks including attributed and multi-hop QA.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Unlocking Markets: A Multilingual Benchmark to Cross-Market Question Answering

Yifei Yuan, Yang Deng, Anders Søgaard, Mohammad Aliannejadi

Users post numerous product-related questions on e-commerce platforms, affecting their purchase decisions. Product-related question answering (PQA) entails utilizing product-related resources to provide precise responses to users. We propose a novel task of Multilingual Cross-market Product-based Question Answering (MCPQA) and define the task as providing answers to product-related questions in a main marketplace by utilizing information from another resource-rich auxiliary marketplace in a multilingual context. We introduce a large-scale dataset comprising over 7 million questions from 17 marketplaces across 11 languages. We then perform automatic translation on the Electronics category of our dataset, naming it as McMarket. We focus on two subtasks: review-based answer generation and product-related question ranking. For each subtask, we label a subset of McMarket using an LLM and further evaluate the quality of the annotations via human assessment. We then conduct experiments to benchmark our dataset, using models ranging from traditional lexical models to LLMs in both single-market and cross-market scenarios across McMarket and the corresponding LLM subset. Results show that incorporating cross-market information significantly enhances performance in both tasks.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

TheoremLlama: Transforming General-Purpose LLMs into Lean4 Experts

Ruida WANG, Jipeng Zhang, Yizhen Jia, Rui Pan, Shizhe Diao, Renjie Pi, Tong Zhang

Proving mathematical theorems using computer-verifiable formal languages like Lean significantly impacts mathematical reasoning. One approach to formal theorem proving involves generating complete proofs using Large Language Models (LLMs) based on Natural Language (NL) proofs. However, due to the scarcity of aligned NL and Formal Language (FL) theorem-proving data most modern LLMs exhibit sub-optimal performance. This scarcity results in a paucity of methodologies for training LLMs and techniques to fully utilize their capabilities in composing formal proofs. To address these challenges, this paper proposes **TheoremLlama***, an end-to-end framework that trains a general-purpose LLM to be a Lean4 expert. **TheoremLlama*** includes NL-FL dataset generation and bootstrapping method to obtain aligned dataset, curriculum learning and training techniques to train the model, and iterative proof writing method to write Lean4 proofs that work together synergistically. Using the dataset generation method in **TheoremLlama***, we provide **Open Bootstrapped Theorems** (GBT), an NL-FL aligned and bootstrapped dataset. Our novel NL-FL bootstrapping method, where NL proofs are integrated into Lean4 code for training datasets, leverages the NL reasoning ability of LLMs for formal reasoning. The **TheoremLlama*** framework achieves cumulative accuracies of 36.48% and 33.61% on MiniF2F-Valid and Test datasets respectively, surpassing the GPT-4 baseline of 22.95% and 25.41%. Our code, model checkpoints, and the generated dataset is published in GitHub

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

ControlMath: Controllable Data Generation Promotes Math Generalist Models

Nuo Chen, Ning Wu, Jianhui Chang, MING GONG, Linjun Shou, Dongmei Zhang, Jia Li

Utilizing large language models (LLMs) for data augmentation has yielded encouraging results in mathematical reasoning. However, these approaches face constraints in problem diversity, potentially restricting them to in-domain/distribution data generation. To this end, we propose **ControlMath***, an iterative method involving an equation-generator module and two LLM-based agents. The module creates diverse equations, which the Problem-Crafter agent then transforms into math word problems. The Reverse-Agent filters and selects high-quality data, adhering to the ‘less is more’ principle. This approach enables the generation of diverse math problems, not limited to specific domains or distributions. As a result, we collect ControlMathQA, which involves 190k math word problems. Extensive results prove that combining our dataset with in-domain datasets like GSM8K can help improve the model’s mathematical ability to generalize, leading to improved performance both within and beyond specific domains.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Gotcha! Don't trick me with unanswerable questions! Self-aligning Large Language Models for Proactively Responding to Unknown Questions

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, Tat-Seng Chua

Despite the remarkable abilities of Large Language Models (LLMs) to answer questions, they often display a considerable level of overconfidence even when the question does not have a definitive answer. To avoid providing hallucinated answers to these unknown questions, existing studies typically investigate approaches to refusing to answer these questions. In this work, we propose a novel and scalable self-alignment method to utilize the LLM itself to enhance its response-ability to different types of unknown questions, being capable of not just refusing to answer but further proactively providing explanations to the unanswerability of unknown questions. Specifically, the Self-Align method first employ a two-stage class-aware self-augmentation approach to generate a large amount of unknown question-response data. Then we conduct disparity-driven self-curation to select qualified data for fine-tuning the LLM itself for aligning the responses to unknown questions as desired. Experimental results on two datasets across four types of unknown questions validate the superiority of the Self-Aligned method over existing baselines in terms of three types of task formulation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

KnowTuning: Knowledge-aware Fine-tuning for Large Language Models

Youngang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, Zhaochun Ren

Despite their success at many natural language processing (NLP) tasks, large language models still struggle to effectively leverage knowledge for knowledge-intensive tasks, manifesting limitations such as generating incomplete, non-factual, or illogical answers. These limitations stem from inadequate knowledge awareness of LLMs during vanilla fine-tuning. To address these problems, we propose a knowledge-aware fine-tuning (KnowTuning) method to improve fine-grained and coarse-grained knowledge awareness of LLMs. We devise a fine-grained knowledge augmentation stage to train LLMs to identify difficult fine-grained knowledge in answers. We also propose a coarse-grained knowledge comparison stage to train LLMs to distinguish between reliable and unreliable knowledge, in three aspects: completeness, factuality, and logicality. Extensive experiments on both generic and medical question answering (QA) datasets confirm the effectiveness of KnowTuning, through automatic and human evaluations, across various sizes of LLMs. We further verify that KnowTuning generates more facts with less factual error rate under fine-grained facts evaluation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models

Wenhai Yu, Hongming Zhang, Xiaoman Pan, peixin cao, Kaixin Ma, Jian Li, Hongwei Wang, Dong Yu

Retrieval-augmented language model (RALM) represents a significant advancement in mitigating factual hallucination by leveraging external knowledge sources. However, the reliability of the retrieved information is not always guaranteed, and the retrieval of irrelevant data can mislead the response generation. Moreover, standard RALMs frequently neglect their intrinsic knowledge due to the interference from retrieved information. In instances where the retrieved information is irrelevant, RALMs should ideally utilize their intrinsic knowledge or, in the absence of both intrinsic and retrieved knowledge, opt to respond with "unknown" to avoid hallucination. In this paper, we introduces Chain-of-Note (CoN), a novel approach to improve robustness of RALMs in facing noisy, irrelevant documents and in handling unknown scenarios. The core idea of CoN is to generate sequential reading notes for each retrieved document, enabling a thorough evaluation of their relevance to the given question and integrating this information to formulate the final answer. Our experimental results show that GPT-4, when equipped with CoN, outperforms the Chain-of-Thought approach. Besides, we utilized GPT-4 to create 10K CoN data, subsequently trained on smaller models like OPT and LLaMa-2. Our experiments across four open-domain QA benchmarks show that fine-tuned RALMs equipped with CoN significantly outperform standard fine-tuned RALMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Learn Beyond The Answer: Training Language Models with Reflection for Mathematical Reasoning

Zhihan Zhang, Tao Ge, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, Meng Jiang

Supervised fine-tuning enhances the problem-solving abilities of language models across various mathematical reasoning tasks. To maximize such benefits, existing research focuses on "broadening" the training set with various data augmentation techniques, which is effective for standard single-round question-answering settings. Our work introduces a novel technique aimed at cultivating a "deeper" understanding of the training problems at hand, enhancing performance not only in standard settings but also in more complex scenarios that require reflective thinking. Specifically, we propose **reflective augmentation***, a method that embeds problem reflection into each training instance. It trains the model to consider alternative perspectives and engage with abstractions and analogies, thereby fostering a thorough comprehension through reflective reasoning. Extensive experiments validate the achievement of our aim, underscoring the unique advantages of our method and its complementary nature relative to existing augmentation techniques.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Do great minds think alike? Investigating Human-AI Complementarity for Question Answering

Maharshi Gor, Hal Daumé III, Tianyu Zhou, Jordan Lee Boyd-Graber

Recent advancements of large language models (LLMs) have led to claims of AI surpassing humans in natural language processing NLP tasks such as textual understanding and reasoning.%This work investigates these assertions by introducing CAIMIRA, a novel framework rooted in item response theory IRT that enables quantitative assessment and comparison of problem-solving abilities in question-answering QA agents.%Through analysis of over 300,000 responses from ~ 70 AI systems and 155 humans across thousands of quiz questions, CAIMIRA uncovers distinct proficiency patterns in knowledge domains and reasoning skills. %Humans outperform AI systems in knowledge-grounded abductive and conceptual reasoning, while state-of-the-art LLMs like GPT-4 Turbo and Llama-3-70B demonstrate superior performance on-targeted information retrieval and fact-based reasoning, particularly when information gaps are well-defined and addressable through pattern matching or data retrieval.%These findings identify key areas for future QA tasks and model development, highlighting the critical need for questions that not only challenge higher-order reasoning and scientific thinking, but also demand nuanced linguistic and cross-contextual application.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Datasets for Multilingual Answer Sentence Selection

Matteo Gabbiro, Stefano Campese, Federico Agostini, Alessandro Moschitti

Answer Sentence Selection (AS2) is a critical task for designing effective retrieval-based Question Answering (QA) systems. Most advancements in AS2 focus on English due to the scarcity of annotated datasets for other languages. This lack of resources prevents the training of effective AS2 models in different languages, creating a performance gap between QA systems in English and other locales. In this paper, we introduce new high-quality datasets for AS2 in five European languages (French, German, Italian, Portuguese, and Spanish), obtained through supervised Automatic Machine Translation (AMT) of existing English AS2 datasets such as ASNQ, WikiQA, and TREC-QA using a Large Language Model (LLM). We evaluated our approach and the quality of the translated datasets through multiple experiments with different Transformer architectures. The results indicate that our datasets are pivotal in producing robust and powerful multilingual AS2 models, significantly contributing to closing the performance gap between English and other languages.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

MDCR: A Dataset for Multi-Document Conditional Reasoning

Peter Baile Chen, Yi Zhang, Chunwei Liu, Sejal Gupta, Yoon Kim, Mike Cafarella

The same real-life questions posed to different individuals may lead to different answers based on their unique situations. For instance, whether a student is eligible for a scholarship depends on eligibility conditions, such as major or degree required. ConditionalQA was proposed to evaluate models' capability of reading a document and answering eligibility questions, considering "unmentioned" conditions. However, it is limited to questions on single documents, neglecting harder cases that may require "cross-document reasoning" and "optimization", for example, "What is the maximum number of scholarships attainable?" Such questions over multiple documents are not only more challenging due to more context to understand, but also because the model has to (1) explore all possible combinations of unmentioned conditions and (2) understand the relationship between conditions across documents, to reason about the optimal outcome. To evaluate models' capability of answering such questions, we propose a new dataset MDCR, which can reflect real-world challenges and serve as a new test bed for complex conditional reasoning that requires optimization. We evaluate this dataset using the most recent LLMs and demonstrate their limitations in solving this task. We believe this dataset will facilitate future research in answering optimization questions with unknown conditions.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

CoXQL: A Dataset for Parsing Explanation Requests in Conversational XAI Systems

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller

Conversational explainable artificial intelligence (ConvXAI) systems based on large language models (LLMs) have garnered significant interest from the research community in natural language processing (NLP) and human-computer interaction (HCI). Such systems can provide answers to user questions about explanations in dialogues, have the potential to enhance users' comprehension and offer more information about the decision-making and generation processes of LLMs. Currently available ConvXAI systems are based on intent recognition rather than free chat, as this has been found to be more precise and reliable in identifying users' intentions. However, the recognition of intents still presents a challenge in the case of ConvXAI, since little training data exist and the domain is highly specific, as there is a broad range of XAI methods to map requests onto. In order to bridge this gap, we present CoXQL, the first dataset in the NLP domain for user intent recognition in ConvXAI, covering 31 intents, seven of which require filling multiple slots. Subsequently, we enhance an existing parsing approach by incorporating template validations, and conduct an evaluation of several LLMs on CoXQL using different parsing strategies. We conclude that the improved parsing approach (MP+) surpasses the performance of previous approaches. We also discover that intents with multiple slots remain highly challenging for LLMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

MedLogic-AQA: Enhancing Medicare Question Answering with Abstractive Models Focusing on Logical Structures

Aizan Zafar, Kshitij Mishra, Asif Ekbal

In Medicare question-answering (QA) tasks, the need for effective systems is pivotal in delivering accurate responses to intricate medical queries. However, existing approaches often struggle to grasp the intricate logical structures and relationships inherent in medical contexts, thus limiting their capacity to furnish precise and nuanced answers. In this work, we address this gap by proposing a novel Abstractive QA system MedLogic-AQA that harnesses first-order logic-based rules extracted from both context and questions to generate well-grounded answers. Through initial experimentation, we identified six pertinent first-order logical rules, which were then used to train a Logic-Understanding (LU) model capable of generating logical triples for a given context, question, and answer. These logic triples are then integrated into the training of MedLogic-AQA, enabling reasoned and coherent reasoning during answer generation. This distinctive fusion of logical reasoning with abstractive question answering equips our system to produce answers that are logically sound, relevant, and engaging. Evaluation with respect to both automated and human-based demonstrates the robustness of MedLogic-AQA against strong baselines. Through empirical assessments and case studies, we validate the efficacy of MedLogic-AQA in elevating the quality and comprehensiveness of answers in terms of reasoning as well as informativeness.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

AHP-Powered LLM Reasoning for Multi-Criteria Evaluation of Open-Ended Responses

Xiaotian Lu, Jiyi Li, Koh Takeuchi, Hisashi Kashima

Question answering (QA) tasks have been extensively studied in the field of natural language processing (NLP). Answers to open-ended questions are highly diverse and difficult to quantify, and cannot be simply evaluated as correct or incorrect, unlike close-ended questions with definitive answers. While large language models (LLMs) have demonstrated strong capabilities across various tasks, they exhibit relatively weaker performance in evaluating answers to open-ended questions. In this study, we propose a method that leverages LLMs and the analytic hierarchy process (AHP) to assess answers to open-ended questions. We utilized LLMs to generate multiple evaluation criteria for a question. Subsequently, answers were subjected to pairwise comparisons under each criterion with LLMs, and scores for each answer were calculated in the AHP. We conducted experiments on four datasets using both ChatGPT-3.5-turbo and GPT-4. Our results indicate that our approach more closely aligns with human judgment compared to the four baselines. Additionally, we explored the impact of the number of criteria, variations in models, and differences in datasets on the results.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SyntTQA: Synergistic Table-based Question Answering via Mixture of Text-to-SQL and E2E TQA

Siyue Zhang, Anh Tuan Luu, Chen Zhao

Text-to-SQL parsing and end-to-end question answering (E2E TQA) are two main approaches for Table-based Question Answering task. Despite success on multiple benchmarks, they have yet to be compared and their synergy remains unexplored. In this paper, we identify different strengths and weaknesses through evaluating state-of-the-art models on benchmark datasets: Text-to-SQL demonstrates superiority in handling questions involving arithmetic operations and long tables; E2E TQA excels in addressing ambiguous questions, non-standard table schema, and complex table contents. To combine both strengths, we propose a Synergistic Table-based Question Answering approach that integrates different models via answer selection, which is agnostic to any model types. Further experiments validate that ensembling models by either feature-based or LLM-based answer selector significantly improves the performance over individual models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Typos that Broke the RAGs Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, Jong C. Park

The robustness of recent Large Language Models (LLMs) has become increasingly crucial as their applicability expands across various domains and real-world applications. Retrieval-Augmented Generation (RAG) is a promising solution for addressing the limitations of LLMs, yet existing studies on the robustness of RAG often overlook the interconnected relationships between RAG components or the potential threats prevalent in real-world databases, such as minor textual errors. In this work, we investigate two underexplored aspects when assessing the robustness of RAG: 1) vulnerability to noisy documents through low-level perturbations and 2) a holistic evaluation of RAG robustness. Furthermore, we introduce a novel attack method, the Genetic Attack on RAG (GARAG), which targets these aspects. Specifically, GARAG is designed to reveal vulnerabilities within each component and test the overall system functionality against noisy documents. We validate

RAG robustness by applying our GARAG to standard QA datasets, incorporating diverse retrievers and LLMs. The experimental results show that GARAG consistently achieves high attack success rates. Also, it significantly devastates the performance of each component and their synergy, highlighting the substantial risk that minor textual inaccuracies pose in disrupting RAG systems in the real world. Code is available at <https://github.com/zomss/GARAG>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

A Notion of Complexity for Theory of Mind via Discrete World Models

X. Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Aspert, Anthony G. Cohn, Michael J. Wooldridge

Theory of Mind (ToM) can be used to assess the capabilities of Large Language Models (LLMs) in complex scenarios where social reasoning is required. While the research community has proposed many ToM benchmarks, their hardness varies greatly, and their complexity is not well defined. This work proposes a framework inspired by cognitive load theory to measure the complexity of ToM tasks. We quantify a problem's complexity as the number of states necessary to solve it correctly. Our complexity measure also accounts for spurious states of a ToM problem designed to make it apparently harder. We use our method to assess the complexity of five widely adopted ToM benchmarks. On top of this framework, we design a prompting technique that augments the information available to a model with a description of how the environment changes with the agents' interactions. We name this technique Discrete World Models (DWM) and show how it elicits superior performance on ToM tasks.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

KB-Plugin: A Plug-and-play Framework for Large Language Models to Induce Programs over Low-resourced Knowledge Bases

Jiajie Zhang, Shulin Cao, Linmet Hu, Ling Feng, Lei Hou, Juanzi Li

Program induction (PI) has become a promising paradigm for using knowledge bases (KBs) to help large language models (LLMs) answer complex knowledge-intensive questions. Nonetheless, PI typically relies on a large number of parallel question-program pairs to make the LLM aware of the schema of a given KB, and is thus challenging for many low-resourced KBs that lack annotated data. To this end, we propose KB-Plugin, a plug-and-play framework that enables LLMs to induce programs over any low-resourced KB. Firstly, KB-Plugin adopts self-supervised learning to encode the detailed schema information of a given KB into a pluggable module, namely schema plugin. Secondly, KB-Plugin utilizes abundant annotated data from a rich-resourced KB to train another pluggable module, namely PI plugin, which can help the LLM extract question-relevant schema information from the schema plugin of any KB and utilize the information to induce programs over this KB. Experiments show that KB-Plugin outperforms SOTA low-resourced PI methods with 25x smaller backbone LLM on both large-scale and domain-specific KBs, and even approaches the performance of supervised methods.

Resources and Evaluation 2

Nov 12 (Tue) 16:00-17:30 - Room: Riverfront Hall

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

CLongEval: A Chinese Benchmark for Evaluating Long-Context Large Language Models

Zexuan Qiu, Jingjing Li, Shijue Huang, Xiaiqi Jiao, Wanjun Zhong, Irwin King

Developing Large Language Models (LLMs) with robust long-context capabilities has been the recent research focus, resulting in the emergence of long-context LLMs proficient in Chinese. However, the evaluation of these models remains underdeveloped due to a lack of benchmarks. To address this gap, we present CLongEval, a comprehensive Chinese benchmark for evaluating long-context LLMs. CLongEval is characterized by three key features: (1) Sufficient data volume, comprising 7 distinct tasks and 7,267 examples; (2) Broad applicability, accommodating to models with context windows size from 1K to 100K; (3) High quality, with over 2,000 manually annotated question-answer pairs in addition to the automatically constructed labels. With CLongEval, we undertake a comprehensive assessment of 6 open-source long-context LLMs and 2 leading commercial counterparts that feature both long-context abilities and proficiency in Chinese. We also provide in-depth analysis based on the empirical results, trying to shed light on the critical capabilities that present challenges in long-context settings. The dataset, evaluation scripts, and model outputs will be released.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Can Large Language Models Always Solve Easy Problems if They Can Solve Harder Ones?

Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, Zifang Sui

Large language models (LLMs) have demonstrated impressive capabilities, but still suffer from inconsistency issues (e.g., LLMs can react differently to disturbances like rephrasing or inconsequential order change). In addition to these inconsistencies, we also observe that LLMs, while capable of solving hard problems, can paradoxically fail at easier ones. To evaluate this hard-to-easy inconsistency, we develop the ConsisEval benchmark, where each entry comprises a pair of questions with a strict order of difficulty. Furthermore, we introduce the concept of consistency score to quantitatively measure this inconsistency and analyze the potential for improvement in consistency by relative consistency score. Based on comprehensive experiments across a variety of existing models, we find: (1) GPT-4 achieves the highest consistency score of 92.2% but is still inconsistent to specific questions due to distraction by redundant information, misinterpretation of questions, etc.; (2) models with stronger capabilities typically exhibit higher consistency, but exceptions also exist; (3) hard data enhances consistency for both fine-tuning and in-context learning. Our data and code will be publicly available on GitHub.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

VIVA: A Benchmark for Vision-Grounded Decision-Making with Human Values

Zhe Hu, Yixiao Ren, Jing Li, Yu Yin

This paper introduces VIVA, a benchmark for VIision-grounded decision-making driven by human VA. While most large vision-language models (VLMs) focus on physical-level skills, our work is the first to examine their multimodal capabilities in leveraging human values to make decisions under a vision-depicted situation. VIVA contains 1,062 images depicting diverse real-world situations and the manually annotated decisions grounded in them. Given an image there, the model should select the most appropriate action to address the situation and provide the relevant human values and reason underlying the decision. Extensive experiments based on VIVA show the limitation of VLMs in using human values to make multimodal decisions. Further analyses indicate the potential benefits of exploiting action consequences and predicted human values.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Evaluating Large Language Models via Linguistic Profiling

Alessio Miaschi, Felice Dell'Orletta, Giulia Venturi

Large Language Models (LLMs) undergo extensive evaluation against various benchmarks collected in established leaderboards to assess their performance across multiple tasks. However, to the best of our knowledge, there is a lack of comprehensive studies evaluating these

models' linguistic abilities independent of specific tasks. In this paper, we introduce a novel evaluation methodology designed to test LLMs' sentence generation abilities under specific linguistic constraints. Drawing on the 'linguistic profiling' approach, we rigorously investigate the extent to which five LLMs of varying sizes, tested in both zero- and few-shot scenarios, effectively adhere to (morpho)syntactic constraints. Our findings shed light on the linguistic proficiency of LLMs, revealing both their capabilities and limitations in generating linguistically-constrained sentences.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

ScalingFilter: Assessing Data Quality through Inverse Utilization of Scaling Laws

Ruihang Li, Yixuan Wei, Miaozen Zhang, Nenghai Yu, Han Hu, Houwen Peng

High-quality data is crucial for the pre-training performance of large language models. Unfortunately, existing quality filtering methods rely on a known high-quality dataset as reference, which can introduce potential bias and compromise diversity. In this paper, we propose ScalingFilter, a novel approach that evaluates text quality based on the perplexity difference between two language models trained on the same data, thereby eliminating the influence of the reference dataset in the filtering process. An theoretical analysis shows that ScalingFilter is equivalent to an inverse utilization of scaling laws. Through training models with 1.3B parameters on the same data source processed by various quality filters, we find ScalingFilter can improve zero-shot performance of pre-trained models in downstream tasks. To assess the bias introduced by quality filtering, we introduce semantic diversity, a metric of utilizing text embedding models for semantic representations. Extensive experiments reveal that semantic diversity is a reliable indicator of dataset diversity, and ScalingFilter achieves an optimal balance between downstream performance and semantic diversity.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

I Could've Asked That: Reformulating Unanswerable Questions

Wenting Zhao, Ge Gao, Claire Cardie, Alexander M Rush

When seeking information from unfamiliar documents, users frequently pose questions that cannot be answered by the documents. While existing large language models (LLMs) identify these unanswerable questions, they do not assist users in reformulating their questions, thereby reducing their overall utility. We curate CouldAsk, an evaluation benchmark composed of existing and new datasets for document-grounded question answering, specifically designed to study reformulating unanswerable questions. We evaluate state-of-the-art open-source and proprietary LLMs on CouldAsk. The results demonstrate the limited capabilities of these models in reformulating questions. Specifically, GPT-4 and Llama2-7B successfully reformulate questions only 26% and 12% of the time, respectively. Error analysis shows that 62% of the unsuccessful reformulations stem from the models merely rephrasing the questions or even generating identical questions. We publicly release the benchmark and the code to reproduce the experiments.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, Minjoon Seo

Proprietary LMs such as GPT-4 are often employed to assess the quality of responses from various LMs. However, concerns including transparency, controllability, and affordability strongly motivate the development of open-source LMs specialized in evaluations. On the other hand, existing open evaluator LMs exhibit critical shortcomings: 1) they issue scores that significantly diverge from those assigned by humans, and 2) they lack the flexibility to perform both direct assessment and pairwise ranking, the two most prevalent forms of assessment. Additionally, they do not possess the ability to evaluate based on custom evaluation criteria, focusing instead on general attributes like helpfulness and harmlessness. To address these issues, we introduce Prometheus 2, a more powerful evaluator LM than its predecessor that closely mirrors human and GPT-4 judgements. Moreover, it is capable of processing both direct assessment and pair-wise ranking formats grouped with a user-defined evaluation criteria. On four direct assessment benchmarks and four pairwise ranking benchmarks, Prometheus 2 scores the highest correlation and agreement with humans and proprietary LM judges among all tested open evaluator LMs. Our models, code, and data are all publicly available.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners

Bowen Jiang, Yangxinyu Xie, Zhuoquin Hao, Xiaomeng Wang, Tamvi Mallick, Weijie J Su, Camillo Jose Taylor, Dan Roth

This study introduces a hypothesis-testing framework to assess whether large language models (LLMs) possess genuine reasoning abilities or primarily depend on token bias. We go beyond evaluating LLMs on accuracy; rather, we aim to investigate their token bias in solving logical reasoning tasks. Specifically, we develop carefully controlled synthetic datasets, featuring conjunction fallacy and syllogistic problems. Our framework outlines a list of hypotheses where token biases are readily identifiable, with all null hypotheses assuming genuine reasoning capabilities of LLMs. The findings in this study suggest, with statistical guarantee, that most LLMs still struggle with logical reasoning. While they may perform well on classic problems, their success largely depends on recognizing superficial patterns with strong token bias, thereby raising concerns about their actual reasoning and generalization abilities.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Quality Matters: Evaluating Synthetic Data for Tool-Using LLMs

Shadi Iskander, Sofia Tolnach, Ori Shapira, Nachshon Cohen, Zohar Karnin

Training large language models (LLMs) for external tool usage is a rapidly expanding field, with recent research focusing on generating synthetic data to address the shortage of available data. However, the absence of systematic data quality checks poses complications for properly training and testing models. To that end, we propose two approaches for assessing the reliability of data for training LLMs to use external tools. The first approach uses intuitive, human-defined correctness criteria. The second approach uses a model-driven assessment with in-context evaluation. We conduct a thorough evaluation of data quality on two popular benchmarks, followed by an extrinsic evaluation that showcases the impact of data quality on model performance. Our results demonstrate that models trained on high-quality data outperform those trained on unvalidated data, even when trained with a smaller quantity of data. These findings empirically support the significance of assessing and ensuring the reliability of training data for tool-using LLMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

NLEBench+NorGLM: A Comprehensive Empirical Analysis and Benchmark Dataset for Generative Language Models in Norwegian

Peng Liu, Lemei Zhang, Terje Farup, Even W. Lauvås, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, Zhirong Yang

Norwegian, spoken by only 5 million population, is under-representative within the most impressive breakthroughs in NLP tasks. To the best of our knowledge, there has not yet been a comprehensive evaluation of the existing language models (LMs) on Norwegian generation tasks during the article writing process. To fill this gap, we 1) compiled the existing Norwegian dataset and pre-trained 4 Norwegian Open Language Models varied from parameter scales and architectures, collectively called NorGLM; 2) introduced a comprehensive benchmark, NLEBench, for evaluating natural language generation capabilities in Norwegian, encompassing translation and human annotation. Based on the investigation, we find that: 1) the mainstream, English-dominated LM GPT-3.5 has limited capability in understanding the Norwegian context; 2) the increase in model parameter scales demonstrates limited impact on the performance of downstream tasks when the pre-training

dataset is constrained in size; 3) smaller models also demonstrate the reasoning capability through Chain-of-Thought; 4) a multi-task dataset that includes synergy tasks can be used to verify the generalizability of LLMs on natural language understanding and, meanwhile, test the interconnectedness of these NLP tasks. We share our resources and code for reproducibility under a CC BY-NC 4.0 license.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Boosting Scientific Concepts Understanding: Can Analogies from Teacher Models Empower Student Models?

Siyu Yuan, Cheng Jiayang, Lin Qiu, Dqing Yang

Analogical reasoning plays a critical role in human cognition, enabling us to understand new concepts by associating them with familiar ones. Previous research in the AI community has mainly focused on identifying and generating analogies and then examining their quality under human evaluation, which overlooks the practical application of these analogies in real-world settings. Inspired by the human education process, in this paper, we propose to investigate how analogies created by teacher language models (LMs) can assist student LMs in understanding scientific concepts, thereby aligning more closely with practical scenarios. Our results suggest that free-form analogies can indeed aid LMs in understanding concepts. Additionally, analogies generated by student LMs can improve their own performance on scientific question answering, demonstrating their capability to use analogies for self-learning new knowledge. Resources are available at <https://github.com/siyuyuan/SCUA>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

InterIntent: Investigating Social Intelligence of LLMs via Intention Understanding in an Interactive Game Context

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, Jieyu Zhao

Large language models (LLMs) have demonstrated the potential to mimic human social intelligence. However, most studies focus on simplistic and static self-report or performance-based tests, which limits the depth and validity of the analysis. In this paper, we developed a novel framework, InterIntent, to assess LLMs' social intelligence by mapping their ability to understand and manage intentions in a game setting. We focus on four dimensions of social intelligence: situational awareness, self-regulation, self-awareness, and theory of mind. Each dimension is linked to a specific game task: intention selection, intention following, intention summarization, and intention guessing. Our findings indicate that while LLMs exhibit high proficiency in selecting intentions, achieving an accuracy of 88%, their ability to infer the intentions of others is significantly weaker, trailing human performance by 20%. Additionally, game performance correlates with intention understanding, highlighting the importance of the four components towards success in this game. These findings underline the crucial role of intention understanding in evaluating LLMs' social intelligence and highlight the potential of using social deduction games as a complex testbed to enhance LLM evaluation. InterIntent contributes a structured approach to bridging the evaluation gap in social intelligence within multiplayer LLM-based games.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

A Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data

Ishaan Wattis, Varun Gunna, Aditya Yadavalli, Vivek Seshadri, Manohar Swamiathan, Sunayana Sitaram

Evaluation of multilingual Large Language Models (LLMs) is challenging due to a variety of factors – the lack of benchmarks with sufficient linguistic diversity, contamination of popular benchmarks into LLM pre-training data and the lack of local, cultural nuances in translated benchmarks. In this work, we study human and LLM-based evaluation in a multilingual, multi-cultural setting. We evaluate 30 models across 10 Indic languages by conducting 90K human evaluations and 30K LLM-based evaluations and find that models such as GPT-4o and Llama-3 70B consistently perform best for most Indic languages. We build leaderboards for two evaluation settings - pairwise comparison and direct assessment and analyse the agreement between humans and LLMs. We find that humans and LLMs agree fairly well in the pairwise setting but the agreement drops for direct assessment evaluation especially for languages such as Bengali and Odia. We also check for various biases in human and LLM-based evaluation and find evidence of self-bias in the GPT-based evaluator. Our work presents a significant step towards scaling up multilingual evaluation of LLMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works

Xinfeng Yuan, Siyu Yuan, Yuhuan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, Dqing Yang

Large language models (LLMs) have demonstrated impressive performance and spurred numerous AI applications, in which role-playing agents (RPAs) are particularly popular, especially for fictional characters. The prerequisite for these RPAs lies in the capability of LLMs to understand characters from fictional works. Previous efforts have evaluated this capability via basic classification tasks or characteristic imitation, failing to capture the nuanced character understanding with LLMs. In this paper, we propose evaluating LLMs' character understanding capability via the character profiling task, i.e., summarizing character profiles from corresponding materials, a widely adopted yet understudied practice for RPA development. Specifically, we construct the CROSS dataset from literature experts and assess the generated profiles by comparing them with ground truth references and evaluating their applicability in downstream tasks. Our experiments, which cover various summarization methods and LLMs, have yielded promising results. These results strongly validate the character understanding capability of LLMs. Resources are available at https://github.com/Joanna0123/character_profiling.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Precise Model Benchmarking with Only a Few Observations

Riccardo Fogliato, Pratik Patil, Nil-Jana Akpinar, Mathew Montfort

How can we precisely estimate a large language model's (LLM) accuracy on questions belonging to a specific topic within a larger question-answering dataset? The standard direct estimator, which averages the model's accuracy on the questions in each subgroup, may exhibit high variance for subgroups (topics) with small sample sizes. Synthetic regression modeling, which leverages the model's accuracy on questions about other topics, may yield biased estimates that are too unreliable for large subgroups. We prescribe a simple yet effective solution: an empirical Bayes (EB) estimator that balances direct and regression estimates for each subgroup separately, improving the precision of subgroup-level estimates of model performance. Our experiments on multiple datasets show that this approach consistently provides more precise estimates of the LLM performance compared to the direct and regression approaches, achieving substantial reductions in the mean squared error. Confidence intervals for EB estimates also have near-nominal coverage and are narrower compared to those for the direct estimator. Additional experiments on tabular and vision data validate the benefits of this EB approach.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

"Rows, Columns and Values, Oh My!" Synthesizing Scientific Literature into Tables using Language Models

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliuue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, Kyle Lo

When conducting literature reviews, scientists often create literature review tables/tables whose rows are publications and whose columns constitute a schema, a set of aspects used to compare and contrast the papers. Can we automatically generate these tables using language models (LMs)? In this work, we introduce a framework that leverages LMs to perform this task by decomposing it into separate schema and value generation steps. To enable experimentation, we address two main challenges: First, we overcome a lack of high-quality datasets to benchmark table generation by curating and releasing arxivDIGESTTables, a new dataset of 2,228 literature review tables extracted from ArXiv papers that synthesize a total of 7,542 research papers. Second, to support scalable evaluation of model generations against human-authored

reference tables, we develop DeContextEval, an automatic evaluation method that aligns elements of tables with the same underlying aspects despite differing surface forms. Given these tools, we evaluate LMs abilities to reconstruct reference tables, finding this task benefits from additional context to ground the generation (e.g. table captions, in-text references). Finally, through a human evaluation study we find that even when LMs fail to fully reconstruct a reference table, their generated novel aspects can still be useful.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Reasoning Robustness of LLMs to Adversarial Typographical Errors

Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, Michael Shieh

Large Language Models (LLMs) have demonstrated impressive capabilities in reasoning using Chain-of-Thought (CoT) prompting. However, CoT can be biased by users' instruction. In this work, we study the reasoning robustness of LLMs to typographical errors, which can naturally occur in users' queries. We design an Adversarial Typo Attack (ATA) algorithm that iteratively samples typos for words that are important to the query and selects the edit that is most likely to succeed in attacking. It shows that LLMs are sensitive to minimal adversarial typographical changes. Notably, with 1 character edit, Mistral-7B's accuracy drops from 43.7% to 38.6% on GSM8K, while with 8 character edits the performance further drops to 19.2%. To extend our evaluation to larger and closed-source LLMs, we develop the R²ATA benchmark, which assesses models' Reasoning Robustness to ATA. It includes adversarial typographical questions derived from three widely-used reasoning datasets GSM8K, BBH, and MMLU by applying ATA to open-source LLMs. R²ATA demonstrates remarkable transferability and causes notable performance drops across multiple super large and closed-source LLMs.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

AmbigNLG: Addressing Task Ambiguity in Instruction for NLG

Ayana Niwa, Hayate Iso

We introduce AmbigNLG, a novel task designed to tackle the challenge of task ambiguity in instructions for Natural Language Generation (NLG). Ambiguous instructions often impede the performance of Large Language Models (LLMs), especially in complex NLG tasks. To tackle this issue, we propose an ambiguity taxonomy that categorizes different types of instruction ambiguities and refines initial instructions with clearer specifications. Accompanying this task, we present AmbigSNI_NLG, a dataset comprising 2,500 instances annotated to facilitate research in AmbigNLG. Through comprehensive experiments with state-of-the-art LLMs, we demonstrate that our method significantly enhances the alignment of generated text with user expectations, achieving up to a 15.02-point increase in ROUGE scores. Our findings highlight the critical importance of addressing task ambiguity to fully harness the capabilities of LLMs in NLG tasks. Furthermore, we confirm the effectiveness of our method in practical settings involving interactive ambiguity mitigation with users, underscoring the benefits of leveraging LLMs for interactive clarification.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DataTales: A Benchmark for Real-World Intelligent Data Narration

Yajing Yang, Qian Liu, Min-Yen Kan

We introduce DataTales, a novel benchmark designed to assess the proficiency of language models in data narration, a task crucial for transforming complex tabular data into accessible narratives. Existing benchmarks often fall short in capturing the requisite analytical complexity for practical applications. DataTales addresses this gap by offering 4.9k financial reports paired with corresponding market data, showcasing the demand for models to create clear narratives and analyze large datasets while understanding specialized terminology in the field. Our findings highlights the significant challenge that language models face in achieving the necessary precision and analytical depth for proficient data narration, suggesting promising avenues for future model development and evaluation methodologies.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading

Tu Anh Dinh, Carlos Mullivo, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon ReiSS, Jieun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbaecher, Clemens Böhm, Jan Niehues

With the rapid development of Large Language Models (LLMs), it is crucial to have benchmarks which can evaluate the ability of LLMs on different domains. One common use of LLMs is performing tasks on scientific topics, such as writing algorithms, querying databases or giving mathematical proofs. Inspired by the way university students are evaluated on such tasks, in this paper, we propose SciEx - a benchmark consisting of university computer science exam questions, to evaluate LLMs' ability on solving scientific tasks. SciEx is (1) multilingual, containing both English and German exams, and (2) multi-modal, containing questions that involve images, and (3) contains various types of freeform questions with different difficulty levels, due to the nature of university exams. We evaluate the performance of various state-of-the-art LLMs on our new benchmark. Since SciEx questions are freeform, it is not straightforward to evaluate LLM performance. Therefore, we provide human expert grading of the LLM outputs on SciEx. We show that the free-form exams in SciEx remain challenging for the current LLMs, where the best LLM only achieves 59.4% exam grade on average. We also provide detailed comparisons between LLM performance and student performance on SciEx. To enable future evaluation of new LLMs, we propose using LLM-as-a-judge to grade the LLM answers on SciEx. Our experiments show that, although they do not perform perfectly on solving the exams, LLMs are decent as graders, achieving 0.948 Pearson correlation with expert grading.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

"A good pun is its own reward": Can Large Language Models Understand Puns?

Zhijun Xu, Siyu Yuan, Lingjie Chen, Deqing Yang

Puns play a vital role in academic research due to their distinct structure and clear definition, which aid in the comprehensive analysis of linguistic humor. However, the understanding of puns in large language models (LLMs) has not been thoroughly examined, limiting their use in creative writing and humor creation. In this paper, we leverage three popular tasks, i.e., pun recognition, explanation and generation to systematically evaluate the capabilities of LLMs in pun understanding. In addition to adopting the automated evaluation metrics from prior research, we introduce new evaluation methods and metrics that are better suited to the in-context learning paradigm of LLMs. These new metrics offer a more rigorous assessment of an LLM's ability to understand puns and align more closely with human cognition than previous metrics. Our findings reveal the "lazy pun generation" pattern and identify the primary challenges LLMs encounter in understanding puns.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SUPER: Evaluating Agents on Setting Up and Executing Tasks from Research Repositories

Ben Bogin, Kejiani Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, Tushar Khot

Given that Large Language Models (LLMs) have made significant progress in writing code, can they now be used to autonomously reproduce results from research repositories? Such a capability would be a boon to the research community, helping researchers validate, understand, and extend prior work. To advance towards this goal, we introduce SUPER, the first benchmark designed to evaluate the capability of LLMs in setting up and executing tasks from research repositories. SUPER aims to capture the realistic challenges faced by researchers working with Machine Learning (ML) and Natural Language Processing (NLP) research repositories. Our benchmark comprises three distinct problem sets: 45 end-to-end problems with annotated expert solutions, 152 sub-problems derived from the expert set that focus on specific challenges (e.g.,

configuring a trainer), and 602 automatically generated problems for larger-scale development. We introduce various evaluation measures to assess both task success and progress, utilizing gold solutions when available or approximations otherwise. We show that state-of-the-art approaches struggle to solve these problems with the best model (GPT-4o) solving only 16.3% of the end-to-end set, and 46.1% of the scenarios. This illustrates the challenge of this task, and suggests that SUPER can serve as a valuable resource for the community to make and measure progress.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DA-Code: Agent Data Science Code Generation Benchmark for Large Language Models

Yining Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lijiu Huang, Xiao Liu, Jun Zhao, Kang Liu

We introduce DA-Code, a code generation benchmark specifically designed to assess LLMs on agent-based data science tasks. This benchmark features three core elements: First, the tasks within DA-Code are inherently challenging, setting them apart from traditional code generation tasks and demanding advanced coding skills in grounding and planning. Second, examples in DA-Code are all based on real and diverse data, covering a wide range of complex data wrangling and analytics tasks. Third, to solve the tasks, the models must utilize complex data science programming languages, including Python and SQL, to perform intricate data processing and derive the answers. We set up the benchmark in a controllable and executable environment that aligns with real-world data analysis scenarios and is scalable. The annotators meticulously designed the evaluation suite to ensure the accuracy and robustness of the evaluation. We developed the DA-Agent baseline. Experiments show that although the baseline performs better than other existing frameworks, using the current best LLMs achieves only 30.5% accuracy, leaving ample room for improvement. We release our benchmark at [link](<https://github.com/yiyihum/dabench>)

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DynamicER: Resolving Emerging Mentions to Dynamic Entities for RAG

Jinyoung Kim, Dayeon Ko, Gunhee Kim

In the rapidly evolving landscape of language, resolving new linguistic expressions in continuously updating knowledge bases remains a formidable challenge. This challenge becomes critical in retrieval-augmented generation (RAG) with knowledge bases, as emerging expressions hinder the retrieval of relevant documents, leading to generator hallucinations. To address this issue, we introduce a novel task aimed at resolving emerging mentions to dynamic entities and present DynamicER benchmark. Our benchmark includes dynamic entity mention resolution and entity-centric knowledge-intensive QA task, evaluating entity linking and RAG model's adaptability to new expressions, respectively. We discovered that current entity linking models struggle to link these new expressions to entities. Therefore, we propose a temporal segmented clustering method with continual adaptation, effectively managing the temporal dynamics of evolving entities and emerging mentions. Extensive experiments demonstrate that our method outperforms existing baselines, enhancing RAG model performance on QA task with resolved mentions.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Towards an Open-Source Speech Foundation Model for EU: 950,000 Hours of Open-Source Compliant Speech Data for EU Languages

Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, Matteo Negri
The rise of foundation models (FMs), coupled with regulatory efforts addressing their risks and impacts, has sparked significant interest in open-source models. However, existing speech FMs (SFM) fall short of full compliance with the open-source principles, even if claimed otherwise, as no existing SFM has model weights, code, and training data publicly available under open-source terms. In this work, we take the first step toward filling this gap by focusing on the 24 official languages of the European Union (EU). We collect suitable training data by surveying automatic speech recognition datasets and unlabeled speech corpora under open-source compliant licenses, for a total of 950k hours. Additionally, we release automatic transcripts for 441k hours of unlabeled data under the permissive CC-BY license, thereby facilitating the creation of open-source SFMs for the EU languages.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Revisiting Automated Evaluation for Long-form Table Question Answering in the Era of Large Language Models

Yugi Wang, Lyuhao Chen, Yilun Zhao

In the era of data-driven decision-making, Long-Form Table Question Answering (LFTQA) is essential for integrating structured data with complex reasoning. Despite recent advancements in Large Language Models (LLMs) for LFTQA, evaluating their effectiveness remains a significant challenge. We introduce LFTQA-Eval, a meta-evaluation dataset comprising 2,988 human-annotated examples, to rigorously assess the efficacy of current automated metrics in assessing LLM-based LFTQA systems, with a focus on faithfulness and comprehensiveness. Our findings reveal that existing automatic metrics poorly correlate with human judgments and fail to consistently differentiate between factually accurate responses and those that are coherent but factually incorrect. Additionally, our in-depth examination of the limitations associated with automated evaluation methods provides essential insights for the improvement of LFTQA automated evaluation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

FinDVer: Explainable Claim Verification over Long and Hybrid-content Financial Documents

Yilun Zhao, Yitao Long, Tintin Jiang, Weiyuan Chen, Chengye Wang, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, Arman Cohan
We introduce FinDVer, a comprehensive benchmark specifically designed to evaluate the explainable claim verification capabilities of LLMs in the context of understanding and analyzing long, hybrid-content financial documents. FinDVer contains 4,000 expert-annotated examples across four subsets, each focusing on a type of scenario that frequently arises in real-world financial domains. We assess a broad spectrum of 25 LLMs under long-context and RAG settings. Our results show that even the current best-performing system (i.e., GPT-4o) significantly lags behind human experts. Our detailed findings and insights highlight the strengths and limitations of existing LLMs in this new task. We believe FinDVer can serve as a valuable benchmark for evaluating LLM capabilities in claim verification over complex, expert-domain documents.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

AKEW: Assessing Knowledge Editing in the Wild

Xiaobao Wu, Liangming Pan, William Yang Wang, Anh Tuan Luu

Knowledge editing injects knowledge updates into language models to keep them correct and up-to-date. However, its current evaluations deviate significantly from practice: their knowledge updates solely consist of structured facts derived from meticulously crafted datasets, instead of practical sources—unstructured texts like news articles, and they often overlook practical real-world knowledge updates. To address these issues, in this paper we propose AKEW (Assessing Knowledge Editing in the Wild), a new practical benchmark for knowledge editing. AKEW fully covers three editing settings of knowledge updates: structured facts, unstructured texts as facts, and extracted triplets. It further introduces new datasets featuring both counterfactual and real-world knowledge updates. Through extensive experiments, we demonstrate the considerable gap between state-of-the-art knowledge-editing methods and practical scenarios. Our analyses further highlight key insights to motivate future research for practical knowledge editing.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

WorryWords: Norms of Anxiety Association for 44,450 English Words

Saif M. Mohammad

Anxiety, the anticipatory unease about a potential negative outcome, is a common and beneficial human emotion. However, there is still much that is not known about anxiety, such as how it relates to our body and how it manifests in language; especially pertinent given the increasing impact of related disorders. In this work, we introduce *WorryWords*, the first large-scale repository of manually derived word-anxiety associations for over 44,450 English words. We show that the anxiety associations are highly reliable. We use *WorryWords* to study the relationship between anxiety and other emotion constructs, as well as the rate at which children acquire anxiety words with age. Finally, we show that using *WorryWords* alone, one can accurately track the change of anxiety in streams of text. *WorryWords* enables a wide variety of anxiety-related research in psychology, NLP, public health, and social sciences. *WorryWords* (and its translations to over 100 languages) is freely available. <http://saifmohammad.com/worrywords.html>

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

YesBut: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Milon Madhur Das, Aman Bansal, ANKIT RAJ, Pawan Goyal, Niloy Ganguly

Understanding satire and humor is a challenging task for even current Vision-Language models. In this paper, we propose the challenging tasks of Satirical Image Detection (detecting whether an image is satirical), Understanding (generating the reason behind the image being satirical), and Completion (given one half of the image, selecting the other half from 2 given options, such that the complete image is satirical) and release a high-quality dataset ***YesBut***, consisting of 2547 images, 1084 satirical and 1463 non-satirical, containing different artistic styles, to evaluate those tasks. Each satirical image in the dataset depicts a normal scenario, along with a conflicting scenario which is funny or ironic. Despite the success of current Vision-Language Models on multimodal tasks such as Visual QA and Image Captioning, our benchmarking experiments show that such models perform poorly on the proposed tasks on the ***YesBut*** Dataset in Zero-Shot Settings w.r.t both automated as well as human evaluation. Additionally, we release a dataset of 119 real, satirical photographs for further research.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Mathador-LM: A Dynamic Benchmark for Mathematical Reasoning on Large Language Models

Eldar Kurtic, Amir Moeini, Dan Alistarh

We introduce Mathador-LM, a new benchmark for evaluating the mathematical reasoning on large language models (LLMs), combining ruleset interpretation, planning, and problem-solving. This benchmark is inspired by the Mathador game, where the objective is to reach a target number using basic arithmetic operations on a given set of base numbers, following a simple set of rules. We show that, across leading LLMs, we obtain stable average performance while generating benchmark instances dynamically, following a target difficulty level. Thus, our benchmark alleviates concerns about test-set leakage into training data, an issue that often undermines popular benchmarks. Additionally, we conduct a comprehensive evaluation of both open and closed-source state-of-the-art LLMs on Mathador-LM. Our findings reveal that contemporary models struggle with Mathador-LM, scoring significantly lower than average 3rd graders. This stands in stark contrast to their strong performance on popular mathematical reasoning benchmarks. The implementation of Mathador-LM benchmark is available at <https://github.com/IST-DASLab/Mathador-LM>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

One Thousand and One Pairs: A "novel" challenge for long-context language models

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, Mohit Iyyer

Synthetic long-context LLM benchmarks (e.g., "needle-in-the-haystack") test only surface-level retrieval capabilities; but how well can long-context LLMs retrieve, synthesize, and reason over information across book-length inputs? We address this question by creating NoCha, a dataset of 1,001 minimally different pairs of true and false claims about 67 recently-published English fictional books, written by human readers of those books. In contrast to existing long-context benchmarks, our annotators confirm that the largest share of pairs in NoCha require global reasoning over the entire book to verify. Our experiments show that while human readers easily perform this task, it is enormously challenging for all ten long-context LLMs that we evaluate: no open-weight model performs above random chance (despite their strong performance on synthetic benchmarks), while GPT-4o achieves the highest pair accuracy at 55.8%. Further analysis reveals that (1) on average, models perform much better on pairs that require only sentence-level retrieval vs. global reasoning; (2) model-generated explanations for their decisions are often inaccurate even for correctly-labeled claims; and (3) models perform substantially worse on speculative fiction books that contain extensive world-building. The methodology proposed in NoCha allows for the evolution of the benchmark dataset and the easy analysis of future models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

StorySpark: Expert-Annotated QA Pairs with Real-World Knowledge for Children Storytelling

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, Yuling Sun

Interactive story reading is common in early childhood education, where teachers expect to teach both language skills and real-world knowledge beyond the story. While many story reading systems have been developed for this activity, they often fail to infuse real-world knowledge into the conversation. This limitation can be attributed to the existing question-answering (QA) datasets used for children's education, upon which the systems are built, failing to capture the nuances of how education experts think when conducting interactive story reading activities. To bridge this gap, we design an annotation framework, empowered by existing knowledge graph to capture experts annotations and thinking process, and leverage this framework to construct StorySparkQA dataset, which comprises 5,868 expert-annotated QA pairs with real-world knowledge. We conduct automated and human expert evaluations across various QA pair generation settings to demonstrate that our StorySparkQA can effectively support models in generating QA pairs that target real-world knowledge beyond story content. StorySparkQA is available at <https://huggingface.co/datasets/NEU-HAI/StorySparkQA>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Data Contamination Can Cross Language Barriers

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, Jingbo Shang

The opacity in developing large language models (LLMs) is raising growing concerns about the potential contamination of public benchmarks in the pre-training data. Existing contamination detection methods are typically based on the text overlap between training and evaluation data, which can be too superficial to reflect deeper forms of contamination. In this paper, we first present a cross-lingual form of contamination that inflates LLMs' performance while evading current detection methods: deliberately injected by overfitting LLMs on the translated versions of benchmark test sets. Then, we propose generalization-based approaches to unmask such deeply concealed contamination. Specifically, we examine the LLM's performance change after modifying the original benchmark by replacing the false answer choices with correct ones from other questions. Contaminated models can hardly generalize to such easier situations, where the false choices can be *not even wrong*, as all choices are correct in their memorization. Experimental results demonstrate that cross-lingual contamination can easily fool existing detection methods, but not ours. In addition, we discuss the potential utilization of cross-lingual contamination in interpreting LLMs' working mechanisms and in post-training LLMs for enhanced multilingual capabilities.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

GuardBench: A Large-Scale Benchmark for Guardrail Models

Elias Bassani, Ignacio Sanchez

Generative AI systems powered by Large Language Models have become increasingly popular in recent years. Lately, due to the risk of providing users with unsafe information, the adoption of those systems in safety-critical domains has raised significant concerns. To respond to this situation, input-output filters, commonly called guardrail models, have been proposed to complement other measures, such as model alignment. Unfortunately, the lack of a standard benchmark for guardrail models poses significant evaluation issues and makes it hard to compare results across scientific publications. To fill this gap, we introduce GuardBench, a large-scale benchmark for guardrail models comprising 40 safety evaluation datasets. To facilitate the adoption of GuardBench, we release a Python library providing an automated evaluation pipeline built on top of it. With our benchmark, we also share the first large-scale prompt moderation datasets in German, French, Italian, and Spanish. To assess the current state-of-the-art, we conduct an extensive comparison of recent guardrail models and show that a general-purpose instruction-following model of comparable size achieves competitive results without the need for specific fine-tuning.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DataNarrative: Automated Data-Driven Storytelling with Visualizations and Texts

Mohammed Saiful Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty

Data-driven storytelling is a powerful method for conveying insights by combining narrative techniques with visualizations and text. These stories integrate visual aids, such as highlighted bars and lines in charts, along with textual annotations explaining insights. However, creating such stories requires a deep understanding of the data and meticulous narrative planning, often necessitating human intervention, which can be time-consuming and mentally taxing. While Large Language Models (LLMs) excel in various NLP tasks, their ability to generate coherent and comprehensive data stories remains underexplored. In this work, we introduce a novel task for data story generation and a benchmark containing 1,449 stories from diverse sources. To address the challenges of crafting coherent data stories, we propose a multi-agent framework employing two LLM agents designed to replicate the human storytelling process: one for understanding and describing the data (Reflection), generating the outline, and narration, and another for verification at each intermediary step. While our agentic framework generally outperforms non-agentic counterparts in both model-based and human evaluations, the results also reveal unique challenges in data story generation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation

Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Małgorzata Anna Wilczyńska, Adam Wierzbicki

This study presents a novel corpus of 15,356 Polish web articles, including articles identified as containing disinformation. Our dataset enables a multifaceted understanding of disinformation. We present a distinctive multilayered methodology for annotating disinformation in texts. What sets our corpus apart is its focus on uncovering hidden intent and manipulation in disinformative content. A team of experts annotated each article with multiple labels indicating both disinformation creators' intents and the manipulation techniques employed. Additionally, we set new baselines for binary disinformation detection and two multiclass multilabel classification tasks: manipulation techniques and intention types classification.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese

Rifki Afina Putri, Faiz Ghifari Haznirama, Dea Adhistha, Alice Oh

Large Language Models (LLMs) are increasingly being used to generate synthetic data for training and evaluating models. However, it is unclear whether they can generate a good quality of question answering (QA) dataset that incorporates knowledge and cultural nuance embedded in a language, especially for low-resource languages. In this study, we investigate the effectiveness of using LLMs in generating culturally relevant commonsense QA datasets for Indonesian and Sundanese languages. To do so, we create datasets for these languages using various methods involving both LLMs and human annotators, resulting in 4.5K questions per language (9K in total), making our dataset the largest of its kind. Our experiments show that automatic data adaptation from an existing English dataset is less effective for Sundanese. Interestingly, using the direct generation method on the target language, GPT-4 Turbo can generate questions with adequate general knowledge in both languages, albeit not as culturally 'deep' as humans. We also observe a higher occurrence of fluency errors in the Sundanese dataset, highlighting the discrepancy between medium- and lower-resource languages.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SciDQA: A Deep Reading Comprehension Dataset over Scientific Papers

Shruti Singh, Nandan Sarkar, Arman Cohan

Scientific literature is typically dense, requiring significant background knowledge and deep comprehension for effective engagement. We introduce SciDQA, a new dataset for reading comprehension that challenges language models to deeply understand scientific articles, consisting of 2,937 QA pairs. Unlike other scientific QA datasets, SciDQA sources questions from peer reviews by domain experts and answers by paper authors, ensuring a thorough examination of the literature. We enhance the dataset's quality through a process that carefully decontextualizes the content, tracks the source document across different versions, and incorporates a bibliography for multi-document question-answering. Questions in SciDQA necessitate reasoning across figures, tables, equations, appendices, and supplementary materials, and require multi-document reasoning. We evaluate several open-source and proprietary LLMs across various configurations to explore their capabilities in generating relevant and factual responses, as opposed to simple review memorization. Our comprehensive evaluation, based on metrics for surface-level and semantic similarity, highlights notable performance discrepancies. SciDQA represents a rigorously curated, naturally derived scientific QA dataset, designed to facilitate research on complex reasoning within the domain of question answering for scientific texts.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

ArMeme: Propagandistic Content in Arabic Memes

Firoj Alam, Abul Hasanat, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain

With the rise of digital communication memes have become a significant medium for cultural and political expression that is often used to mislead audience. Identification of such misleading and persuasive multimodal content become more important among various stakeholders, including social media platforms, policymakers, and the broader society as they often cause harm to the individuals, organizations and/or society. While there has been effort to develop AI based automatic system for resource rich languages (e.g., English), it is relatively little to none for medium to low resource languages. In this study, we focused on developing an Arabic memes dataset with manual annotations of propagandistic content. We annotated $\sim 6K$ Arabic memes collected from various social media platforms, which is a first resource for Arabic multimodal research. We provide a comprehensive analysis aiming to develop computational tools for their detection. We made the dataset publicly available for the community.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Evaluating Large Language Models on Time Series Feature Understanding: A Comprehensive Taxonomy and Benchmark

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, Svitlana Vytnrenko

Large Language Models (LLMs) offer the potential for automatic time series analysis and reporting, which is a critical task across many

domains, spanning healthcare, finance, climate, energy, and many more. In this paper, we propose a framework for rigorously evaluating the capabilities of LLMs on time series understanding, encompassing both univariate and multivariate forms. We introduce a comprehensive taxonomy of time series features, a critical framework that delineates various characteristics inherent in time series data. Leveraging this taxonomy, we have systematically designed and synthesized a diverse dataset of time series, embodying the different outlined features, each accompanied by textual descriptions. This dataset acts as a solid foundation for assessing the proficiency of LLMs in comprehending time series. Our experiments shed light on the strengths and limitations of state-of-the-art LLMs in time series understanding, revealing which features these models readily comprehend effectively and where they falter. In addition, we uncover the sensitivity of LLMs to factors including the formatting of the data, the position of points queried within a series and the overall time series length.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Holistic Evaluation for Interleaved Text-and-Image Generation

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, Lifu Huang

Interleaved text-and-image generation has been an intriguing research direction, where the models are required to generate both images and text pieces in an arbitrary order. Despite the emerging advancements in interleaved generation, the progress in its evaluation still significantly lags behind. Existing evaluation benchmarks do not support arbitrarily interleaved images and text for both inputs and outputs, and they only cover a limited number of domains and use cases. Also, current works predominantly use similarity-based metrics which fall short in assessing the quality in open-ended scenarios. To this end, we introduce InterleavedBench, the first benchmark carefully curated for the evaluation of interleaved text-and-image generation. InterleavedBench features a rich array of tasks to cover diverse real-world use cases. In addition, we present InterleavedEval, a strong reference-free metric powered by GPT-4o to deliver accurate and explainable evaluation. We carefully define five essential evaluation aspects for InterleavedEval, including text quality, perceptual quality, image coherence, text-image coherence, and helpfulness, to ensure a comprehensive and fine-grained assessment. Through extensive experiments and rigorous human evaluation, we show that our benchmark and metric can effectively evaluate the existing models with a strong correlation with human judgments surpassing previous reference-based metrics. We also provide substantial findings and insights to foster future research in interleaved generation and its evaluation.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

FOLIO: Natural Language Reasoning with First-Order Logic

SIMENG HAN, Hailey Schoekopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekataterina Zubova, Matthew Burttel, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Limyong Nan, Jungo Kasai, Tai Yu, Rui Zhang, Alexander Fabri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, Dragomir Radev

Large language models (LLMs) have achieved remarkable performance on a variety of natural language understanding tasks. However, existing benchmarks are inadequate in measuring the complex logical reasoning capabilities of a model. We present FOLIO, a human-annotated, logically complex and diverse dataset for reasoning in natural language (NL), equipped with first-order logic (FOL) annotations. FOLIO consists of 1,430 examples (unique conclusions), each paired with one of 487 sets of premises used to deductively reason for the validity of each conclusion. The logical correctness of the premises and conclusions is ensured by their FOL annotations, which are automatically verified by an FOL inference engine. In addition to the main NL reasoning task, NL-FOL pairs in FOLIO constitute a new NL-FOL translation dataset. Our experiments on FOLIO systematically evaluate the FOL reasoning ability of supervised fine-tuning on medium-sized language models. For both NL reasoning and NL-FOL translation, we benchmark multiple state-of-the-art language models. Our results show that a subset of FOLIO remains a challenge for one of the most capable Large Language Model (LLM) publicly available, GPT-4.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

PolyWER: A Holistic Evaluation Framework for Code-Switched Speech Recognition

Karima Kadaoui, Maryam Al Ali, Hawau Olamide Toyin, Ibrahim Mohammed, Hanan Aldarmaki

Code-switching in speech, particularly between languages that use different scripts, can potentially be correctly transcribed in various forms, including different ways of transliteration of the embedded language into the matrix language script. Traditional methods for measuring accuracy, such as Word Error Rate (WER), are too strict to address this challenge. In this paper, we introduce PolyWER, a proposed framework for evaluating speech recognition systems to handle language-mixing. PolyWER accepts transcriptions of code-mixed segments in different forms, including transliterations and translations. We demonstrate the algorithms use cases through detailed examples, and evaluate it against human judgement. To enable the use of this metric, we appended the annotations of a publicly available Arabic-English code-switched dataset with transliterations and translations of code-mixed speech. We also utilize these additional annotations for fine-tuning ASR models and compare their performance using PolyWER. In addition to our main finding on PolyWERs effectiveness, our experiments show that alternative annotations could be more effective for fine-tuning monolingual ASR models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

VeriScore: Evaluating the factuality of verifiable claims in long-form text generation

Xiaoxia Song, Yekyoung Kim, Mohit Iyer

Existing metrics for evaluating the factuality of long-form text, such as FACTSCORE (Min et al., 2023) and SAFE (Wei et al., 2024), decompose an input text into atomic claims and verify each against a knowledge base like Wikipedia. These metrics are not suitable for most generation tasks because they assume that every claim is verifiable (i.e., can plausibly be proven true or false). We address this issue with VERISCORE, a metric for evaluating factuality in diverse long-form generation tasks that contain both verifiable and unverifiable content. VERISCORE can be effectively implemented with either closed or fine-tuned open-weight language models. Human evaluation confirms that VERISCOREs extracted claims are more sensible than those from competing methods across eight different long-form tasks. We use VERISCORE to evaluate generations from 16 different models across multiple long-form tasks and find that while GPT-4o is the best-performing model overall, open-weight models such as Mixtral-8E22 are closing the gap. We show that an LMs VERISCORE on one task (e.g., biography generation) does not necessarily correlate to its VERISCORE on a different task (e.g., long-form QA), highlighting the need for expanding factuality evaluation across tasks with varying fact density.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

OffsetBias: Leveraging Debiased Data for Tuning Evaluators

Junssoo Park, Seungyeon Jwe, REN MEIYING, Daeyoung Kim, Sanghyuk Choi

Employing Large Language Models (LLMs) to assess the quality of generated responses has become a widely adopted evaluation method. Specifically, instruct-tuned models and fine-tuned judge models based on open-source LLMs have been reported. While it is known that judge models are vulnerable to certain biases, such as favoring longer answers regardless of content, the specifics of these biases remain under-explored. In this work, we qualitatively identify six types of biases inherent in various judge models. We propose EvalBiasBench as a meta-evaluation collection of hand-crafted test cases for each bias type. Additionally, we present de-biasing dataset construction methods and the associated preference dataset OffsetBias. Experimental results demonstrate that fine-tuning on our dataset significantly enhances the robustness of judge models against biases and improves performance across most evaluation scenarios. We release our datasets and the fine-tuned judge model to public.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

SAFARI: Cross-lingual Bias and Factuality Detection in News Media and News Articles

Dilshod Azizov, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, Shangsong Liang

In an era where information is quickly shared across many cultural and language contexts, the neutrality and integrity of news media are essential. Ensuring that media content remains unbiased and factual is crucial for maintaining public trust. With this in mind, we introduce SAFARI (Cross-lingual BiAs and Factuality Detection in News Media and News ARTicles), a novel corpus of news media and articles for predicting political bias and the factuality of reporting in a multilingual and cross-lingual setup. To the best of our knowledge, this corpus is unprecedented in its collection and introduces a dataset for political bias and factuality for three tasks: (i) media-level, (ii) article-level, and (iii) joint modeling at the article-level. At the media and article levels, we evaluate the cross-lingual ability of the models; however, in joint modeling, we evaluate on English data. Our frameworks set a new benchmark in the cross-lingual evaluation of political bias and factuality. This is achieved through the use of various Multilingual Pre-trained Language Models (MPLMs) and Large Language Models (LLMs) coupled with ensemble learning methods.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

DAVINCI: Dataset for Detection of Violent Incidents

Hernán Lamba, Anton Abilov, Ke Zhang, Elizabeth M Olson, Henry Kudzanai Dambanemuya, João Cordovil Bárcia, David S. Batista, Christina Wille, Aoife Cahill, Joel R. Tetreault, Alejandro Jaimes

Humanitarian organizations can enhance their effectiveness by analyzing data to discover trends, gather aggregated insights, manage their security risks, support decision-making, and inform advocacy and funding proposals. However, data about violent incidents with direct impact and relevance for humanitarian aid operations is not readily available. An automatic data collection and NLP-backed classification framework aligned with humanitarian perspectives can help bridge this gap. In this paper, we present HumVI a dataset comprising news articles in three languages (English, French, Arabic) containing instances of different types of violent incidents categorized by the humanitarian sector they impact, e.g., aid security, education, food security, health, and protection. Reliable labels were obtained for the dataset by partnering with a data-backed humanitarian organization, Insecurity Insight. We provide multiple benchmarks for the dataset, employing various deep learning architectures and techniques, including data augmentation and mask loss, to address different task-related challenges, e.g., domain expansion. The dataset is publicly available at <https://github.com/dataminer-ai/humvi-dataset>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, OSAMA MOHAMMED AFZAL, Liyangang Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, Preslav Nakov

The increased use of large language models (LLMs) across a variety of real-world applications calls for mechanisms to verify the factual accuracy of their outputs. In this work, we present Factcheck-Bench, a holistic end-to-end framework for annotating and evaluating the factuality of LLM-generated responses, which encompasses a multi-stage annotation scheme designed to yield detailed labels for fact-checking and correcting not just the final prediction, but also the intermediate steps that a fact-checking system might need to take. Based on this framework, we construct an open-domain factuality benchmark in three-levels of granularity: claim, sentence, and document. We further propose a system, Factcheck-GPT, which follows our framework, and we show that it outperforms several popular LLM fact-checkers. We make our annotation tool, annotated data, benchmark, and code available at <https://github.com/yuxiaiw/Factcheck-GPT>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

MORE: Evaluating and Quantifying Unimodal Biases in Multimodal Large Language Models through a Causal Lens

Meiqi Chen, Yixin Cao, Yan Zhang, Chaochao Lu

Recent advancements in Large Language Models (LLMs) have facilitated the development of Multimodal LLMs (MLLMs). Despite their impressive capabilities, MLLMs often suffer from over-reliance on unimodal biases (e.g., language bias and vision bias), leading to incorrect answers in complex multimodal tasks. To investigate this issue, we propose a causal framework to interpret the biases in Visual Question Answering (VQA) problems. Within this framework, we conduct an in-depth causal analysis to assess the causal effect of these biases on MLLM predictions. Based on the analysis, we introduce 1) a novel MORE dataset with 12,000 challenging VQA instances requiring multi-hop reasoning and overcoming unimodal biases. 2) a causality-enhanced agent framework CAVE that guides models to comprehensively integrate information from different modalities and mitigate biases. Our experiments show that MLLMs perform poorly on MORE, indicating strong unimodal biases and limited semantic understanding. However, when integrated with our CAVE, promising improvements in reasoning and bias mitigation can be seen. These findings provide important insights for the development of more robust MLLMs and contribute to the broader goal of advancing multimodal AI systems capable of deeper understanding and reasoning. Our project page is at <https://github.com/OpenCausalLab/MORE>.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

PythonSaga: Redefining the Benchmark to Evaluate Code Generating LLMs

Ankit Yadav, Mayank Singh, Himanshu Beniwal

Driven by the surge in code generation using large language models (LLMs), numerous benchmarks have emerged to evaluate these LLMs capabilities. We conducted a large-scale human evaluation of *HumanEval* and *MBPP*, two popular benchmarks for Python code generation, analyzing their diversity and difficulty. Our findings unveil a critical bias towards a limited set of programming concepts, neglecting most of the other concepts entirely. Furthermore, we uncover a worrying prevalence of easy tasks that can inflate model performance estimations. To address these limitations, we propose a novel benchmark, *PythonSaga*, featuring 185 hand-crafted prompts in a balanced representation of 38 programming concepts across diverse difficulty levels. The robustness of our benchmark is demonstrated by the poor performance of existing Code-LLMs. The code and data set are openly available to the NLP community at [URL](<https://github.com/PythonSaga/Python-Saga>).

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

The SIFO Benchmark: Investigating the Sequential Instruction Following Ability of Large Language Models

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, Maarten de Rijke

Following multiple instructions is a crucial ability for large language models (LLMs). Evaluating this ability comes with significant challenges: (i) limited coherence between multiple instructions, (ii) positional bias where the order of instructions affects model performance, and (iii) a lack of objectively verifiable tasks. To address these issues, we introduce a benchmark designed to evaluate models' abilities to follow multiple instructions through sequential instruction following (SIFO) tasks. In SIFO, the successful completion of multiple instructions is verifiable by examining only the final instruction. Our benchmark evaluates instruction following using four tasks (text modification, question answering, mathematics, and security rule following), each assessing different aspects of sequential instruction following. Our evaluation of popular LLMs, both closed-source and open-source, shows that more recent and larger models significantly outperform their older and smaller counterparts on the SIFO tasks, validating the benchmark's effectiveness. All models struggle with following sequences of instructions, hinting at an important lack of robustness of today's language models.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Efficient Data Generation for Source-grounded Information-seeking Dialogs: A Use Case for Meeting Transcripts

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, Ido Dagan

Automating data generation with Large Language Models (LLMs) has become increasingly popular. In this work, we investigate the feasibility and effectiveness of LLM-based data generation in the challenging setting of source-grounded information-seeking dialogs, with response attribution, over long documents. Our source texts consist of long and noisy meeting transcripts, adding to the task complexity. Since automating attribution remains difficult, we propose a semi-automatic approach: dialog queries and responses are generated with LLMs, followed by human verification and identification of attribution spans. Using this approach, we created MISeD – Meeting Information Seeking Dialogs dataset – a dataset of information-seeking dialogs focused on meeting transcripts. Models finetuned with MISeD demonstrate superior performance compared to off-the-shelf models, even those of larger size. Finetuning on MISeD gives comparable response generation quality to finetuning on fully manual data, while improving attribution quality and reducing time and effort.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

IntentionQA: A Benchmark for Evaluating Purchase Intention Comprehension Abilities of Language Models in E-commerce

Wenxuan Ding, Bingyi Wang, Sze Heng Douglas Kwok, Minghao LIU, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingsyu Yin, Bing Yin, Junxian He, Yangguo Song

Enhancing Language Models' (LMs) ability to understand purchase intentions in E-commerce scenarios is crucial for their effective assistance in various downstream tasks. However, previous approaches that distill intentions from LMs often fail to generate meaningful and human-centric intentions applicable in real-world E-commerce contexts. This raises concerns about the true comprehension and utilization of purchase intentions by LMs. In this paper, we present IntentionQA, a double-task multiple-choice question answering benchmark to evaluate LMs' comprehension of purchase intentions in E-commerce. Specifically, LMs are tasked to infer intentions based on purchased products and utilize them to predict additional purchases. IntentionQA consists of 4,360 carefully curated problems across three difficulty levels, constructed using an automated pipeline to ensure scalability on large E-commerce platforms. Human evaluations demonstrate the high quality and low false-negative rate of our benchmark. Extensive experiments across 19 language models show that they still struggle with certain scenarios, such as understanding products and intentions accurately, jointly reasoning with products and intentions, and more, in which they fall far behind human performances.

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tauseef Nayeem, Enamul Hoque

Natural language is a powerful complementary modality of communication for data visualizations, such as bar and line charts. To facilitate chart-based reasoning using natural language, various downstream tasks have been introduced recently such as chart question answering, chart summarization, and fact-checking with charts. These tasks pose a unique challenge, demanding both vision-language reasoning and a nuanced understanding of chart data tables, visual encodings, and natural language instructions. Despite the recent success of Large Language Models (LLMs) across diverse NLP tasks, their abilities and limitations in the realm of data visualization remain under-explored, possibly due to their lack of multi-modal capabilities. To bridge the gap, this paper presents one of the first comprehensive evaluations of the recently developed large vision language models (LVLMs) for chart understanding and reasoning tasks. Our evaluation includes a comprehensive assessment of both closed and open-sourced LVLMs across five major chart reasoning tasks. Furthermore, we perform a qualitative evaluation of LVLMs' performance on a diverse range of charts, aiming to provide a thorough analysis. Our findings reveal that while LVLMs demonstrate impressive abilities in generating fluent texts covering high-level data insights, they also encounter common problems like hallucinations, factual errors, and data bias. We highlight the key strengths and limitations of LVLMs in chart comprehension tasks, offering insights for future research³

Nov 12 (Tue) 16:00-17:30 - Riverfront Hall

Comparing Edge-based and Node-based Methods on a Citation Prediction Task

Peter Vickers, Kenneth Church

Citation Prediction, estimating whether paper a cites paper b, is particularly interesting in a forecasting setting where the model is trained on papers published before time t, and evaluated on papers published after h, where h is the forecast horizon. Performance improves with t (larger training sets) and degrades with h (longer forecast horizons). The trade-off between edge-based methods and node-based methods depends on t. Because edges grow faster than nodes, larger training sets favor edge-based methods. We introduce a new forecast-based Citation Prediction benchmark of 3 million papers to quantify these trends. Our benchmark shows that desirable policies for combining edge- and node-based methods depend on h and t. We release our benchmark, evaluation scripts, and embeddings.

Sentiment Analysis, Stylistic Analysis, and Argument Mining

Nov 12 (Tue) 16:00-17:30 - Room: Jasmine

Nov 12 (Tue) 16:00-17:30 - Jasmine

Hateful Word in Context Classification

Same Hoeken, Sina ZarrieSS, Özge Alacam

Hate speech detection is a prevalent research field, yet it remains underexplored at the level of word meaning. This is significant, as terms used to convey hate often involve non-standard or novel usages which might be overlooked by commonly leveraged LMs trained on general language use. In this paper, we introduce the Hateful Word in Context Classification (**HateWiC**) task and present a dataset of ∼4000 WiC-instances, each labeled by three annotators. Our analyses and computational exploration focus on the interplay between the subjective nature (context-dependent connotations) and the descriptive nature (as described in dictionary definitions) of hateful word senses. HateWiC annotations confirm that harmfulness of a word in context does not always derive from the sense definition alone. We explore the prediction of both majority and individual annotator labels, and we experiment with modeling context- and sense-based inputs. Our findings indicate that including definitions proves effective overall, yet not in cases where hateful connotations vary. Conversely, including annotator demographics becomes more important for mitigating performance drop in subjective hate prediction.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Eyes Don't Lie: Subjective Hate Annotation and Detection with Gaze

³We will make all our prompts as well as LVLMs' responses open source for future research.

Özge Alacam, Sanne Hoeken, Sina ZarrieSS

Hate speech is a complex and subjective phenomenon. In this paper, we present a dataset (GAZE4HATE) that provides gaze data collected in a hate speech annotation experiment. We study whether the gaze of an annotator provides predictors of their subjective harmfulness rating, and how gaze features can improve Hate Speech Detection (HSD). We conduct experiments on statistical modeling of subjective hate ratings and gaze and analyze to what extent rationales derived from hate speech models correspond to human gaze and explanations in our data. Finally, we introduce MEANION, a first gaze-integrated HSD model. Our experiments show that particular gaze features like dwell time or fixation counts systematically correlate with annotators subjective hate ratings and improve predictions of text-only hate speech models.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Overcome Noise and Bias: Segmentation-Aided Multi-Granularity Denoising and Debiasing for Enhanced Quadruples Extraction in Dialogue

Xianlong Luo, Yihao Wang, Meng Yang

Dialogue Aspect-based Sentiment Quadruple analysis (DiaASQ) extends ABSA to more complex real-world scenarios (i.e., dialogues), which makes existing generation methods encounter heightened noise and order bias challenges, leading to decreased robustness and accuracy. To address these, we propose the Segmentation-Aided multi-grained Denoising and Debiasing (SADD) method. For noise, we propose the Multi-Granularity Denoising Generation model (MGDG), achieving word-level denoising via sequence labeling and utterance-level denoising via topic-aware dialogue segmentation. Denoised Attention in MGDG integrates multi-grained denoising information to help generate denoised output. For order bias, we first theoretically analyze its direct cause as the gap between ideal and actual training objectives and propose a distribution-based solution. Since this solution introduces a one-to-many learning challenge, our proposed Segmentation-aided Order Bias Mitigation (SOBM) method utilizes dialogue segmentation to supplement order diversity, concurrently mitigating this challenge and order bias. Experiments demonstrate SADD's effectiveness, achieving state-of-the-art results with a 6.52% F1 improvement.

Nov 12 (Tue) 16:00-17:30 - Jasmine

MiniConGTS: A Near Ultimate Minimalist Contrastive Grid Tagging Scheme for Aspect Sentiment Triplet Extraction

Qiao Sun, Liuqia Yang, Minghao Ma, Nanyang Ye, Qinying Gu

Aspect Sentiment Triplet Extraction (ASTE) aims to co-extract the sentiment triplets in a given corpus. Existing approaches within the pre-training-finetuning paradigm tend to either meticulously craft complex tagging schemes and classification heads, or incorporate external semantic augmentation to enhance performance. In this study, we, for the first time, re-evaluate the redundancy in tagging schemes and the internal enhancement in pretrained representations. We propose a method to improve and utilize pretrained representations by integrating a minimalist tagging scheme and a novel token-level contrastive learning strategy. The proposed approach demonstrates comparable or superior performance compared to state-of-the-art techniques while featuring a more compact design and reduced computational overhead. Additionally, we are the first to formally evaluate GPT-4's performance in few-shot learning and Chain-of-Thought scenarios for this task. The results demonstrate that the pretraining-finetuning paradigm remains highly effective even in the era of large language models.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Dynamic Multi-Reward Weighting for Multi-Style Controllable Generation

Karin De Langis, Ryan Koo, Dongyeop Kang

Textual style expresses a diverse set of information, including interpersonal dynamics (e.g., formality) and the authors emotions or attitudes (e.g., disgust). An open question is how language models can be explicitly controlled so that they weave together target styles when generating text: for example, to produce text that is both negative and non-toxic. One approach to such controlled generation is multi-objective reinforcement learning (RL), but how to best combine multiple objectives in a reward function is an open question. In this paper, we investigate various formulations of multi-style reward formulations, including calibrated outputs from discriminators and dynamic weighting by discriminator gradient magnitudes. We find that our proposed dynamic weighting outperforms static weighting approaches with respect style control while maintaining linguistic quality, and we explore its effectiveness in 2- and 3-style control.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Diversity Over Size: On the Effect of Sample and Topic Sizes for Topic-Dependent Argument Mining Datasets

Benjamin Schiller, Johannes Daxenberger, Andreas Waldls, Iryna Gurevych

Topic-Dependent Argument Mining (TDAM), that is extracting and classifying argument components for a specific topic from large document sources, is an inherently difficult task for machine learning models and humans alike, as large TDAM datasets are rare and recognition of argument components requires expert knowledge. The task becomes even more difficult if it also involves stance detection of retrieved arguments. In this work, we investigate the effect of TDAM dataset composition in few- and zero-shot settings. Our findings show that, while fine-tuning is mandatory to achieve acceptable model performance, using carefully composed training samples and reducing the training sample size by up to almost 90% can still yield 95% of the maximum performance. This gain is consistent across three TDAM tasks on three different datasets. We also publish a new dataset and code for future benchmarking.

Nov 12 (Tue) 16:00-17:30 - Jasmine

A Bayesian Approach to Harnessing the Power of LLMs in Authorship Attribution

Zhengjian Hu, Tong Zheng, Heng Huang

Authorship attribution aims to identify the origin or author of a document. Traditional approaches have heavily relied on manual features and fail to capture long-range correlations, limiting their effectiveness. Recent advancements leverage text embeddings from pre-trained language models, which require significant fine-tuning on labeled data, posing challenges in data dependency and limited interpretability. Large Language Models (LLMs), with their deep reasoning capabilities and ability to maintain long-range textual associations, offer a promising alternative. This study explores the potential of pre-trained LLMs in one-shot authorship attribution, specifically utilizing Bayesian approaches and probability outputs of LLMs. Our methodology calculates the probability that a text entails previous writings of an author, reflecting a more nuanced understanding of authorship. By utilizing only pre-trained models such as Llama-3-70B, our results on the IMDb and blog datasets show an impressive 85% accuracy in one-shot authorship classification across ten authors. Our findings set new baselines for one-shot authorship analysis using LLMs and expand the application scope of these models in forensic linguistics. This work also includes extensive ablation studies to validate our approach.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Style-Specific Neurons for Steering LLMs in Text Style Transfer

Wen Lai, Viktor Hanya, Alexander Fraser

Text style transfer (TST) aims to modify the style of a text without altering its original meaning. Large language models (LLMs) demonstrate superior performance across multiple tasks, including TST. However, in zero-shot setups, they tend to directly copy a significant portion of the input text to the output without effectively changing its style. To enhance the stylistic variety and fluency of the text, we present sNeuron-TST, a novel approach for steering LLMs using style-specific neurons in TST. Specifically, we identify neurons associated with the source and target styles and deactivate source-style-only neurons to give target-style words a higher probability, aiming to enhance the stylistic diversity of the generated text. However, we find that this deactivation negatively impacts the fluency of the generated text, which we address by proposing

an improved contrastive decoding method that accounts for rapid token probability shifts across layers caused by deactivated source-style neurons. Empirical experiments demonstrate the effectiveness of the proposed method on six benchmarks, encompassing formality, toxicity, politeness, politeness, authorship, and sentiment.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, Rodrigo Agerri

Explaining Artificial Intelligence (AI) decisions is a major challenge nowadays in AI, in particular when applied to sensitive scenarios like medicine and law. However, the need to explain the rationale behind decisions is a main issues also for human-based deliberation as it is important to justify why a certain decision has been taken. Resident medical doctors for instance are required not only to provide a (possibly correct) diagnosis, but also to explain how they reached a certain conclusion. Developing new tools to aid residents to train their explanation skills is therefore a central objective of AI in education. In this paper, we follow this direction, and we present, to the best of our knowledge, the first multilingual dataset for Medical Question Answering where correct and incorrect diagnoses for a clinical case are enriched with a natural language explanation written by doctors. These explanations have been manually annotated with argument components (i.e., premise, claim) and argument relations (i.e., attack, support). The Multilingual CasiMedicos-arg dataset consists of 558 clinical cases (English, Spanish, French, Italian) with explanations, where we annotated 5021 claims, 2313 premises, 2431 support relations, and 1106 attack relations. We conclude by showing how competitive baselines perform over this challenging dataset for the argument mining task.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Flee the Flaw: Annotating the Underlying Logic of Fallacious Arguments Through Templates and Slot-filling

Irfan Robbani, Paul Reisert, Surawat Pothong, Naoya Inoue, Camélia Guerraoui, Wenzhi Wang, Shoichi Naito, Jungmin Choi, Kentaro Inui
Prior research in computational argumentation has mainly focused on scoring the quality of arguments, with less attention on explicating logical errors. In this work, we introduce four sets of explainable templates for common informal logical fallacies designed to explicate a fallacy's implicit logic. Using our templates, we conduct an annotation study on top of 400 fallacious arguments taken from LOGIC dataset and achieve a high agreement score (Krippendorff's α of 0.54) and reasonable coverage 83%. Finally, we conduct an experiment for detecting the structure of fallacies and discover that state-of-the-art language models struggle with detecting fallacy templates (0.47 accuracy). To facilitate research on fallacies, we make our dataset and guidelines publicly available.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Lets Discuss! Quality Dimensions and Annotated Datasets for Computational Argument Quality

Rositsa V Ivanova, Thomas Huber, Christina Niklaus

Research in the computational assessment of Argumentation Quality has gained popularity over the last ten years. Various quality dimensions have been explored through the creation of domain-specific datasets and assessment methods. We survey the related literature (211 publications and 32 datasets), while addressing potential overlaps and blurry boundaries to related domains. This paper provides a representative overview of the state of the art in Computational Argument Quality Assessment with a focus on quality dimensions and annotated datasets. The aim of the survey is to identify research gaps and to aid future discussions and work in the domain.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Contrastive Classification via Linear Layer Extrapolation

Mayukh Sharma, Sean O'Brien, Julian McAuley

Certain abilities of Transformer-based language models consistently emerge in their later layers. Previous research has leveraged this phenomenon to improve factual accuracy through self-contrast, penalizing early-exit predictions based on the premise that later-layer updates are more factually reliable than earlier-layer associations. We observe a similar pattern for fine-grained emotion classification in text, demonstrating that self-contrast can enhance encoder-based text classifiers. Additionally, we reinterpret self-contrast as a form of linear extrapolation, which motivates a refined approach that dynamically adjusts the contrastive strength based on the selected intermediate layer. Experiments across multiple models and emotion classification datasets show that our method outperforms standard classification techniques in fine-grained emotion classification tasks.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

UniMEEC: Towards Unified Multimodal Emotion Recognition and Emotion Cause

Guimin Hu, Zihlong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, Jiayuan Xie

Multimodal emotion recognition in conversation (MERC) and multimodal emotion-cause pair extraction (MECPE) have recently garnered significant attention. Emotions are the expression of affect or feelings; responses to specific events, or situations – known as emotion causes. Both collectively explain the causality between human emotion and intents. However, existing works treat emotion recognition and emotion cause extraction as two individual problems, ignoring their natural causality. In this paper, we propose a Unified Multimodal Emotion recognition and Emotion-Cause analysis framework (UniMEEC) to explore the causality between emotion and emotion cause. Concretely, UniMEEC reformulates the MERC and MECPE tasks as mask prediction problems and unifies them with a causal prompt template. To differentiate the modal effects, UniMEEC proposes a multimodal causal prompt to probe the pre-trained knowledge specified to modality and implements cross-task and cross-modality interactions under task-oriented settings. Experiment results on four public benchmark datasets verify the model performance on MERC and MECPE tasks and achieve consistent improvements compared with the previous state-of-the-art methods.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models

Huy Nghiêm, Hal Daumé III

The widespread use of social media necessitates reliable and efficient detection of offensive content to mitigate harmful effects. Although sophisticated models perform well on individual datasets, they often fail to generalize due to varying definitions and labeling of "offensive content." In this paper, we introduce HateCOT, an English dataset with over 52,000 samples from diverse sources, featuring explanations generated by GPT-3.5Turbo and curated by humans. We demonstrate that pretraining on HateCOT significantly enhances the performance of open-source Large Language Models on three benchmark datasets for offensive content detection in both zero-shot and few-shot settings, despite differences in domain and task. Additionally, HateCOT facilitates effective K-shot fine-tuning of LLMs with limited data and improves the quality of their explanations, as confirmed by our human evaluation.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

On the Causal Nature of Sentiment Analysis

Zhiheng Lyu, Zhijing Jin, Fernando Gonzalez Adauto, Rada Mihalcea, Bernhard Schölkopf, Mrinmaya Sachan

Sentiment analysis (SA) aims to identify the sentiment expressed in a piece of text, often in the form of a review. Assuming a review and the sentiment associated with it, in this paper we formulate SA as a combination of two tasks: (1) a causal discovery task that distinguishes whether a review primes the sentiment (Causal Hypothesis C1), or the sentiment primes the review (Causal Hypothesis C2); and (2) the

traditional prediction task to model the sentiment using the review as input. Using the peak-end rule in psychology, we classify a sample as C1 if its overall sentiment score approximates an average of all the sentence-level sentiments in the review, and as C2 if the overall sentiment score approximates an average of the peak and end sentiments. For the prediction task, we use the discovered causal mechanisms behind the samples to improve the performance of LLMs by proposing causal prompts that give the models an inductive bias of the underlying causal graph, leading to substantial improvements by up to 32.13 F1 points on zero-shot five-class SA.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Make Compound Sentences Simple to Analyze: Learning to Split Sentences for Aspect-based Sentiment Analysis

Yongsik Seo, Sungwon Song, Ryang Heo, Jieyoung Kim, Dongha Lee

In the domain of Aspect-Based Sentiment Analysis (ABSA), generative methods have shown promising results and achieved substantial advancements. However, despite these advancements, the tasks of extracting sentiment quadruplets, which capture the nuanced sentiment expressions within a sentence, remain significant challenges. In particular, compound sentences can potentially contain multiple quadruplets, making the extraction task increasingly difficult as sentence complexity grows. To address this issue, we are focusing on simplifying sentence structures to facilitate the easier recognition of these elements and crafting a model that integrates seamlessly with various ABSA tasks. In this paper, we propose Aspect Term Oriented Sentence Splitter (ATOSS), which simplifies compound sentence into simpler and clearer forms, thereby clarifying their structure and intent. As a plug-and-play module, this approach retains the parameters of the ABSA model while making it easier to identify essential intent within input sentences. Extensive experimental results show that utilizing ATOSS outperforms existing methods in both ASQP and ACOS tasks, which are the primary tasks for extracting sentiment quadruplets.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Designing Logic Pattern Templates for Counter-Argument Logical Structure Analysis

Shoichi Naito, Wenzhi Wang, Paul Reisert, Naoya Inoue, Camelia Guerraoui, Kenshi Yamaguchi, Jungmin Choi, Irfan Robbani, Surawat Potthong, Kentaro Inui

Counterarguments (CAs) are a valuable tool in education for enhancing critical thinking skills. Despite their effectiveness, the logical attack structure of counterarguments in relation to their corresponding opponent argument remains unexplored due to its complexity. Towards tackling this challenge, in this work, we introduce Counter-Argument Logical Structure Analysis (**CALSA**), a new task. We first define 10 new CA logic patterns, each comprised of a unique template and slots. We then conduct an annotation study on top of 778 CAs using our patterns to create a new dataset. Our dataset achieves high annotator agreement (Krippendorff $\alpha=0.50$) and high coverage (86.5%). We perform preliminary experiments employing recent large language models to assess the feasibility of automating CA logical structure analysis. Our findings highlight the task's inherent complexity within a straightforward framework, indicating exciting opportunities for further exploration. We publicly release our dataset and model scripts at <https://github.com/cl-tohoku/CALSA>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Persuasiveness of Generated Free-Text Rationales in Subjective Decisions: A Case Study on Pairwise Argument Ranking

Mohamed Elaraby, Diane Litman, Xiang Lorraine Li, Ahmed Magouda

Generating free-text rationales is among the emergent capabilities of Large Language Models (LLMs). These rationales have been found to enhance LLM performance across various NLP tasks. Recently, there has been growing interest in using these rationales to provide answers for various important downstream tasks. In this paper, we analyze generated free-text rationales in tasks with subjective answers, emphasizing the importance of rationalization in such scenarios. We focus on pairwise argument ranking, a highly subjective task with significant potential for real-world applications, such as debate assistance. We evaluate the persuasiveness of rationales generated by nine LLMs to support their subjective choices. Our findings suggest that open-source LLMs, particularly Llama2-70B-chat, are capable of providing highly persuasive rationalizations, surpassing even GPT models. Additionally, our experiments demonstrate that the persuasiveness of the generated rationales can be enhanced by guiding their persuasive elements through prompting or self-refinement techniques.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Shoes-ACOSI: A Dataset for Aspect-Based Sentiment Analysis with Implicit Opinion Extraction

Joseph J Peper, Wenzhao Qiu, Ryan Bruggeman, Yi Han, Estefania Ciliotta Chehade, Lu Wang

We explore *implicit opinion extraction* as a new component of aspect-based sentiment analysis (ABSA) systems. Prior work in ABSA has investigated opinion extraction as an important subtask, however, these works only label concise, *explicitly*-stated opinion spans. In this work, we present **Shoes-ACOSI***, a new and challenging ABSA dataset in the e-commerce domain with implicit opinion span annotations, the first of its kind. Shoes-ACOSI builds upon the existing Aspect-Category-Opinion-Sentiment (ACOS) quadruple extraction task, extending the task to quintuple extraction—now localizing and differentiating both implicit and explicit opinion. In addition to the new annotation schema, our dataset contains paragraph-length inputs which, importantly, present complex challenges through increased input length, increased number of sentiment expressions, and more mixed-sentiment-polarity examples when compared with existing benchmarks. We quantify the difficulty of our new dataset by evaluating with state-of-the-art fully-supervised and prompted-LLM baselines. We find our dataset presents significant challenges for both supervised models and LLMs, particularly from the new implicit opinion extraction component of the ACOSI task, highlighting the need for continued research into implicit opinion understanding.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Multi-label Sequential Sentence Classification via Large Language Model

Mengfei Lan, Lecheng Zheng, Shufan Ming, Halil Kılıçoglu

Sequential sentence classification (SSC) in scientific publications is crucial for supporting downstream tasks such as fine-grained information retrieval and extractive summarization. However, current SSC methods are constrained by model size, sequence length, and single-label setting. To address these limitations, this paper proposes LLM-SSC, a large language model (LLM)-based framework for both single- and multi-label SSC tasks. Unlike previous approaches that employ small- or medium-sized language models, the proposed framework utilizes LLMs to generate SSC labels through designed prompts, which enhance task understanding by incorporating demonstrations and a query to describe the prediction target. We also present a multi-label contrastive learning loss with auto-weighting scheme, enabling the multi-label classification task. To support our multi-label SSC analysis, we introduce and release a new dataset, biorc800, which mainly contains unstructured abstracts in the biomedical domain with manual annotations. Experiments demonstrate LLM-SSC's strong performance in SSC under both in-context learning and task-specific tuning settings. We release biorc800 and our code at: <https://github.com/ScienceNLP-Lab/LLM-SSC>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

The Overlooked Repetitive Lengthening Form in Sentiment Analysis

Lei Wang, Eduard Dragut

Individuals engaging in online communication frequently express personal opinions with informal styles (e.g., memes and emojis). While Language Models (LMs) with informal communications have been widely discussed, a unique and emphatic style, the Repetitive Lengthening Form (RLF), has been overlooked for years. In this paper, we explore answers to two research questions: 1) Is RLF important for SA? 2) Can LMs understand RLF? Inspired by previous linguistic research, we curate **Lengthening**, the first multi-domain dataset with 850k samples

focused on RLF for sentiment analysis. Moreover, we introduce **Explnstruct**, a two-stage Explainable Instruction Tuning framework aimed at improving both the performance and explainability of LLMs for RLF. We further propose a novel unified approach to quantify LMs' understanding of informal expressions. We show that RLF sentences are expressive expressions and can serve as signatures of document-level sentiment. Additionally, RLF has potential value for online content analysis. Our comprehensive results show that fine-tuned Pre-trained Language Models (PLMs) can surpass zero-shot GPT-4 in performance but not in explanation for RLF. Finally, we show Explnstruct can improve the open-sourced LLMs to match zero-shot GPT-4 in performance and explainability for RLF with limited samples. Code and sample data are available at <https://github.com/Tom-Owl/OverlookedRLF>

Nov 12 (Tue) 16:00-17:30 - Jasmine

ASTE-Transformer: Modelling Dependencies in Aspect-Sentiment Triplet Extraction

Iwo Naglik, Mateusz Lango

Aspect-Sentiment Triplet Extraction (ASTE) is a recently proposed task of aspect-based sentiment analysis that consists in extracting (aspect phrase, opinion phrase, sentiment polarity) triples from a given sentence. Recent state-of-the-art methods approach this task by first extracting all possible text spans from a given text, then filtering the potential aspect and opinion phrases with a classifier, and finally considering all their pairs with another classifier that additionally assigns sentiment polarity to them. Although several variations of the above scheme have been proposed, the common feature is that the final result is constructed by a sequence of independent classifier decisions. This hinders the exploitation of dependencies between extracted phrases and prevents the use of knowledge about the interrelationships between classifier predictions to improve performance. In this paper, we propose a new ASTE approach consisting of three transformer-inspired layers, which enables the modelling of dependencies both between phrases and between the final classifier decisions. Experimental results show that the method achieves higher performance in terms of F1 measure than other methods studied on popular benchmarks. In addition, we show that a simple pre-training technique further improves the performance of the model.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Boosting Large Language Models with Continual Learning for Aspect-based Sentiment Analysis

Xuanwen Ding, Jie Zhou, Liang Dou, Qin Chen, Yuanbin Wu, Arlene Chen, Liang He

Aspect-based sentiment analysis (ABSA) is an important subtask of sentiment analysis, which aims to extract the aspects and predict their sentiments. Most existing studies focus on improving the performance of the target domain by fine-tuning domain-specific models (trained on source domains) based on the target domain dataset. Few works propose continual learning tasks for ABSA, which aim to learn the target domain's ability while maintaining the history domains' abilities. In this paper, we propose a Large Language Model-based Continual Learning (LLM-CL) model for ABSA. First, we design a domain knowledge decoupling module to learn a domain-invariant adapter and separate domain-variant adapters independently with an orthogonal constraint. Then, we introduce a domain knowledge warmup strategy to align the representation between domain-invariant and domain-variant knowledge. In the test phase, we index the corresponding domain-variant knowledge via domain positioning to not require each sample's domain ID. Extensive experiments over 19 datasets indicate that our LLM-CL model obtains new state-of-the-art performance.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Improving Argument Effectiveness Across Ideologies using Instruction-tuned Large Language Models

Roxanne El Baff, Khalid Al Khatib, Midad Alshomary, Kai Konen, Benno Stein, Henning Wachsmuth

Different political ideologies (e.g., liberal and conservative Americans) hold different worldviews, which leads to opposing stances on different issues (e.g., gun control) and, thereby, fostering societal polarization. Arguments are a means of bringing the perspectives of people with different ideologies closer together, depending on how well they reach their audience. In this paper, we study how to computationally turn ineffective arguments into effective arguments for people with certain ideologies by using instruction-tuned large language models (LLMs), looking closely at style features. For development and evaluation, we collect ineffective arguments per ideology from debate.org, and we generate about 30k, which we rewrite using three LLM methods tailored to our task: zero-shot prompting, few-shot prompting, and LLM steering. Our experiments provide evidence that LLMs naturally improve argument effectiveness for liberals. Our LLM-based and human evaluation show a clear preference towards the rewritten arguments. Code and link to the data are available here: <https://github.com/roxaneelbaff/emnlp2024-esta>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Stanceformer: Target-Aware Transformer for Stance Detection

Krishna Garg, Cornelia Caragea

The task of Stance Detection involves discerning the stance expressed in a text towards a specific subject or target. Prior works have relied on existing transformer models that lack the capability to prioritize targets effectively. Consequently, these models yield similar performance regardless of whether we utilize or disregard target information, undermining the task's significance. To address this challenge, we introduce Stanceformer, a target-aware transformer model that incorporates enhanced attention towards the targets during both training and inference. Specifically, we design a *Target Awareness* matrix that increases the self-attention scores assigned to the targets. We demonstrate the efficacy of the Stanceformer with various BERT-based models, including state-of-the-art models and Large Language Models (LLMs), and evaluate its performance across three stance detection datasets, alongside a zero-shot dataset. Our approach Stanceformer not only provides superior performance but also generalizes even to other domains, such as Aspect-based Sentiment Analysis. We make the code publicly available.⁴

Special Theme: Efficiency in Model Algorithms, Training, and Inference 2

Nov 12 (Tue) 16:00-17:30 - Room: Jasmine

Nov 12 (Tue) 16:00-17:30 - Jasmine

Towards Pareto-Efficient RLHF: Paying Attention to a Few High-Reward Samples

Changhun Lee, Chiehyeon Lim

Recently, leveraging reinforcement learning (RL) to fine-tune language models (LMs), known as reinforcement learning from human feedback (RLHF), has become an important research topic. However, there is still a lack of theoretical understanding of how RLHF works, the conditions under which it succeeds or fails, and whether it guarantees optimization of both likelihood $\beta(\cdot)$ and reward $R(\cdot)$ objectives. To address these issues, we consider RLHF as a bi-objective problem that has the nature of a *Pareto* optimization, present a Pareto improvement condition that is necessary to obtain Pareto-efficient policies, and propose a simple yet powerful method named *reward dropout* that guarantees a Pareto improvement. To demonstrate the performance of reward dropout, two benchmark datasets commonly used in text style transfer

⁴<https://github.com/kgarg8/Stanceformer>

tasks were utilized in our study: sentiment and topic datasets sourced from Yelp and AG_News, respectively. Our experiments highlight that paying attention to a few samples with higher rewards leads to greater Pareto improvements regardless of model size. We also demonstrate that the effect of reward dropout is generalizable and most effective with non-pretrained target models, saving the effort of pretraining.

Nov 12 (Tue) 16:00-17:30 - Jasmine

UniGen: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation

Juhwan Choi, Yeonghwa Kim, Seunguk Yu, JungMin Yun, YoungBin Kim

Although pre-trained language models have exhibited great flexibility and versatility with prompt-based few-shot learning, they suffer from the extensive parameter size and limited applicability for inference. Recent studies have suggested that PLMs be used as dataset generators and a tiny task-specific model be trained to achieve efficient inference. However, their applicability to various domains is limited because they tend to generate domain-specific datasets. In this work, we propose a novel approach to universal domain generalization that generates a dataset regardless of the target domain. This allows for generalization of the tiny task model to any domain that shares the label space, thus enhancing the real-world applicability of the dataset generation paradigm. Our experiments indicate that the proposed method accomplishes generalizability across various domains while using a parameter set that is orders of magnitude smaller than PLMs.

Nov 12 (Tue) 16:00-17:30 - Jasmine

AdaZeta: Adaptive Zeroth-Order Tensor-Train Adaption for Memory-Efficient Large Language Models Fine-Tuning

Yifan Yang, Kai Chen, Ershad Banijamali, Athanasios Mouchtaris, Zheng Zhang

Fine-tuning large language models (LLMs) has achieved remarkable performance across various natural language processing tasks, yet it demands more and more memory as model sizes keep growing. To address this issue, the recently proposed Memory-efficient Zeroth-order (MeZO) methods attempt to fine-tune LLMs using only forward passes, thereby avoiding the need for a backpropagation graph. However, significant performance drops and a high risk of divergence have limited their widespread adoption. In this paper, we propose the Adaptive Zeroth-order Tensor-Train Adaption (AdaZeta) framework, specifically designed to improve the performance and convergence of the ZO methods. To enhance dimension-dependent ZO estimation accuracy, we introduce a fast-forward, low-parameter tensorized adapter. To tackle the frequently observed divergence issue in large-scale ZO fine-tuning tasks, we propose an adaptive query number schedule that guarantees convergence. Detailed theoretical analysis and extensive experimental results on Roberta-Large and Llama-2-7B models substantiate the efficacy of our AdaZeta framework in terms of accuracy, memory efficiency, and convergence speed.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Eliminating Biased Length Reliance of Direct Preference Optimization via Down-Sampled KL Divergence

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, di yin, Xing Sun

Direct Preference Optimization (DPO) has emerged as a prominent algorithm for the direct and robust alignment of Large Language Models (LLMs) with human preferences, offering a more straightforward alternative to the complex Reinforcement Learning from Human Feedback (RLHF). Despite its promising efficacy, DPO faces a notable drawback: "verbosity", a common over-optimization phenomenon also observed in RLHF. While previous studies mainly attributed verbosity to biased labels within the data, we propose that the issue also stems from an inherent algorithmic length reliance in DPO. Specifically, we suggest that the discrepancy between sequence-level KullbackLeibler (KL) divergences between chosen and rejected sequences, used in DPO, results in overestimated or underestimated rewards due to varying token lengths. Empirically, we utilize datasets with different label lengths to demonstrate the presence of biased rewards. We then introduce an effective downsampling approach, named SamPO, to eliminate potential length reliance. Our experimental evaluations, conducted across three LLMs of varying scales and a diverse array of conditional and open-ended benchmarks, highlight the efficacy of SamPO in mitigating verbosity, achieving improvements of 5% to 12% over DPO through debiased rewards. Our code can be accessed at: <https://github.com/Lu-Junru/SamPO>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Rethinking Pruning Large Language Models: Benefits and Pitfalls of Reconstruction Error Minimization

Sunghin Shin, Wonpyo Park, Jaeho Lee, Namhoon Lee

This work suggests fundamentally rethinking the current practice of pruning large language models (LLMs). The way it is done is by divide and conquer: split the model into submodels, sequentially prune them, and reconstruct predictions of the dense counterparts on small calibration data one at a time; the final model is obtained simply by putting the resulting sparse submodels together. While this approach enables pruning under memory constraints, it generates high reconstruction errors. In this work, we first present an array of reconstruction techniques that can significantly reduce this error by more than 90%. Unwittingly, however, we discover that minimizing reconstruction error is not always ideal and can overfit the given calibration data, resulting in rather increased language perplexity and poor performance at downstream tasks. We find out that a strategy of self-generating calibration data can mitigate this trade-off between reconstruction and generalization, suggesting new directions in the presence of both benefits and pitfalls of reconstruction for pruning LLMs.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale

Flavio Di Palo, Prateek Singh, Bilal H Fadlallah

Large Language Models (LLMs) face significant challenges at inference time due to their high computational demands. To address this, we present Performance-Guided Knowledge Distillation (PGKD), a cost-effective and high-throughput solution for production text classification applications. PGKD utilizes teacher-student Knowledge Distillation to distill the knowledge of LLMs into smaller, task-specific models. PGKD establishes an active learning routine between the student model and the LLM; the LLM continuously generates new training data leveraging hard-negative mining, student model validation performance, and early-stopping protocols to inform the data generation. By employing a cyclical, performance-aware approach tailored for highly multi-class, sparsely annotated datasets prevalent in industrial text classification, PGKD effectively addresses training challenges and outperforms traditional BERT-base models and other knowledge distillation methods on several multi-class classification datasets. Additionally, cost and latency benchmarking reveals that models fine-tuned with PGKD are up to 130X faster and 25X less expensive than LLMs for inference on the same classification task. While PGKD is showcased for text classification tasks, its versatile framework can be extended to any LLM distillation task, including language generation, making it a powerful tool for optimizing performance across a wide range of AI applications.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Optimized Speculative Sampling for GPU Hardware Accelerators

Dominik Wagner, Seanie Lee, Ilja Baumann, Philipp Seerberger, Korbinian Riedhammer, Tobias Bocklet

In this work, we optimize speculative sampling for parallel hardware accelerators to improve sampling speed. We notice that substantial portions of the intermediate matrices necessary for speculative sampling can be computed concurrently. This allows us to distribute the workload across multiple GPU threads, enabling simultaneous operations on matrix segments within thread blocks. This results in profiling time improvements ranging from 6% to 13% relative to the baseline implementation, without compromising accuracy. To further accelerate speculative sampling, probability distributions parameterized by softmax are approximated by sigmoid. This approximation approach results in significantly greater relative improvements in profiling time, ranging from 37% to 94%, with a minor decline in accuracy. We conduct ex-

tensive experiments on both automatic speech recognition and summarization tasks to validate the effectiveness of our optimization methods.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

MOSEL: Inference Serving Using Dynamic Modality Selection

Bodun Hu, Le Xu, Jeongyo Moon, Neeraja J Yadwadkar, Aditya Akella

Rapid advancements over the years have helped machine learning models reach previously hard-to-achieve goals, sometimes even exceeding human capabilities. However, achieving desired accuracy comes at the cost of larger model sizes and increased computational demands. Thus, serving predictions from these models to meet any latency and cost requirements of applications remains a key challenge, despite recent work in building inference serving systems as well as algorithmic approaches that dynamically adapt models based on inputs. Our paper introduces a new form of dynamism, modality selection, where we adaptively choose modalities from inference inputs while maintaining the model performance. We introduce MOSEL, an automated inference serving system for multi-modal ML models that carefully picks input modalities per request based on user-defined performance and accuracy requirements. MOSEL exploits modality configurations extensively, improving system throughput by $3.6 \times$ with an accuracy guarantee. It also reduces job completion times by $11 \times$ compared to modality-agnostic approaches.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Fast Forwarding Low-Rank Training

Adir Rahamim, Naomi Saphra, Sara Kangaslahti, Yonatan Belinkov

Parameter efficient finetuning methods like low-rank adaptation (LoRA) aim to reduce the computational costs of finetuning pretrained Language Models (LMs). Enabled by these low-rank settings, we propose an even more efficient optimization strategy: Fast Forward, a simple and effective approach to accelerate large segments of SGD training. In a Fast Forward stage, we repeat the most recent optimizer step until the loss stops improving on a tiny validation set. By alternating between regular optimization steps and Fast Forward stages, Fast Forward provides up to an 87% reduction in FLOPs over standard SGD with Adam. We validate Fast Forward by finetuning various models on different tasks and demonstrate that it speeds up training without compromising model performance. Additionally, we analyze when and how to apply Fast Forward.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Breaking ReLU Barrier: Generalized MoEification for Dense Pretrained Models

Jaeseong Lee, seung-won hwang, Wonpyo Park, Mingi Ji

As the scale of language models (LMs) continues to grow, there is a heightened interest in reducing the inference cost associated with these models. Mixture-of-Experts (MoEs) present an efficient alternative to dense models, while the existing methods to convert pretrained dense models, MoEs is limited to ReLU-based models with natural sparsity. This paper introduces G-MoEification, applicable to arbitrary dense models, where ReLU-based activation sparsity assumptions no longer hold. For generalizations, we encounter the dilemma of needing to zero-out deactivated experts, while also avoiding excessive zeroing-out to retain dense activation information. We publicly release our code and report results conducted with mBERT, SantaCoder-1.1B, Phi-2-2.7B, and Falcon-7B demonstrating the efficacy of our approach in general scenarios, from multitask to multilingual, from fine-tuning to zero-shot evaluation.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Towards Fast Multilingual LLM Inference: Speculative Decoding and Specialized Drafters

Eunin Yi, Taeheyon Kim, Hongseok Jeung, Du-Seong Chang, Se-Young Yun

Large language models (LLMs) have revolutionized natural language processing and broadened their applicability across diverse commercial applications. However, the deployment of these models is constrained by high inference time in multilingual settings. To mitigate this challenge, this paper explores a training recipe of an assistant model in speculative decoding, which are leveraged to draft and then its future tokens are verified by the target LLM. We show that language-specific draft models, optimized through a targeted pretrain-and-finetune strategy, substantially brings a speedup of inference time compared to the previous methods. We validate these models across various languages in inference time, out-of-domain speedup, and GPT-4o evaluation.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minshofer, Markus Schödl

Segmenting text into sentences plays an early and crucial role in many NLP systems. This is commonly achieved by using rule-based or statistical methods relying on lexical features such as punctuation. Although some recent works no longer exclusively rely on punctuation, we find that no prior method achieves all of (i) robustness to missing punctuation, (ii) effective adaptability to new domains, and (iii) high efficiency. We introduce a new model Segment Any Text (SaT) to solve this problem. To enhance robustness, we propose a new pretraining scheme that ensures less reliance on punctuation. To address adaptability, we introduce an extra stage of parameter-efficient fine-tuning, establishing state-of-the-art performance in distinct domains such as verses from lyrics and legal documents. Along the way, we introduce architectural modifications that result in a threefold gain in speed over the previous state of the art and solve spurious reliance on context far in the future. Finally, we introduce a variant of our model with fine-tuning on a diverse, multilingual mixture of sentence-segmented data, acting as a drop-in replacement and enhancement for existing segmentation tools. Overall, our contributions provide a universal approach for segmenting any text. Our method outperforms all baselines including strong LLMs across 8 corpora spanning diverse domains and languages, especially in practically relevant situations where text is poorly formatted. Our models and code, including documentation, are readily available at <https://github.com/segment-any-text/wtspsplit> under the MIT license.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Heterogeneous LoRA for Federated Fine-tuning of On-Device Foundation Models

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrizi, Gauri Joshi

Foundation models (FMs) adapt surprisingly well to downstream tasks with fine-tuning. However, their colossal parameter space prohibits their training on resource-constrained edge-devices. For federated fine-tuning, we need to consider the smaller FMs of few billion parameters at most, namely on-device FMs (ODFMs), which can be deployed on-device. Federated fine-tuning of ODFMs has unique challenges non-present in standard fine-tuning: i) ODFMs poorly generalize to downstream tasks due to their limited sizes making proper fine-tuning imperative to their performance, and ii) devices have limited and heterogeneous system capabilities and data that can deter the performance of fine-tuning. Tackling these challenges, we propose HetLoRA, a feasible and effective federated fine-tuning method for ODFMs that leverages the system and data heterogeneity at the edge. HetLoRA allows heterogeneous LoRA ranks across clients for their individual system resources, and efficiently aggregates and distributes these LoRA modules in a data-aware manner by applying rank self-pruning locally and sparsity-weighted aggregation at the server. It combines the advantages of high and low-rank LoRAs, achieving improved convergence speed and final performance compared to homogeneous LoRA. Furthermore, HetLoRA has enhanced computation and communication efficiency compared to full fine-tuning making it more feasible for the edge.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Chain and Causal Attention for Efficient Entity Tracking

Erwan Fagnou, Paul Caillon, Blaise Delattre, Alexandre Allauzen

This paper investigates the limitations of transformers for entity-tracking tasks in large language models. We identify a theoretical constraint, showing that transformers require at least $\log_2(n + 1)$ layers to handle entity tracking with n state changes. To address this issue, we propose an efficient and frugal enhancement to the standard attention mechanism, enabling it to manage long-term dependencies more efficiently. By considering attention as an adjacency matrix, our model can track entity states with a single layer. Empirical results demonstrate significant improvements in entity tracking datasets while keeping competitive performance on standard natural language modeling. Our modified attention allows us to achieve the same performance with drastically fewer layers. Additionally, our enhanced mechanism reveals structured internal representations of attention. Extensive experiments on both toy and complex datasets validate our approach. Our contributions include theoretical insights, an improved attention mechanism, and empirical validation.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Ouroboros: Generating Longer Drafts Phrase by Phrase for Faster Speculative Decoding

Weilin Zhao, Yuxiang Huang, Xu Han, Wang Xu, Chaojun Xiao, Xinrong Zhang, Yewei Fang, Kaihuo Zhang, Zhiyuan Liu, Maosong Sun

Speculative decoding is a widely used method that accelerates the generation process of large language models (LLMs) with no compromise in model performance. It achieves this goal by using an existing smaller model for drafting and then employing the target LLM to verify the draft in a low-cost parallel manner. Under such a drafting-verification framework, drafting efficiency has become a bottleneck in the final speedup of speculative decoding. Therefore, generating longer drafts at less cost can lead to better decoding speedup. To achieve this, we introduce Ouroboros, which can generate draft phrases to parallelize the drafting process and meanwhile lengthen drafts in a training-free manner. The experimental results on various typical text generation tasks show that Ouroboros can achieve speedups of up to $2.4 \times$ over speculative decoding and $3.9 \times$ over vanilla decoding, without fine-tuning draft and target models. Code available at <https://github.com/thunlp/Ouroboros>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Fewer is More: Boosting Math Reasoning with Reinforced Context Pruning

Xijie Huang, Li Yang Zhang, Kwang-Ting Cheng, Fan Yang, Mao Yang

Large Language Models (LLMs) have shown impressive capabilities, yet they still struggle with math reasoning. In this work, we propose CoT-Influx, a novel approach that pushes the boundary of few-shot Chain-of-Thoughts (CoT) learning to improve LLM mathematical reasoning. Motivated by the observation that adding more concise CoT examples in the prompt can improve LLM reasoning performance, CoT-Influx employs a coarse-to-fine pruner to maximize the input of effective and concise CoT examples. The pruner first selects as many crucial CoT examples as possible and then prunes unimportant tokens to fit the context window. As a result, by enabling more CoT examples with double the context window size in tokens, CoT-Influx significantly outperforms various prompting baselines across various LLMs (LLaMA2-7B, 13B, 70B) and 5 math datasets, achieving up to 4.55% absolute improvements. Remarkably, without any fine-tuning, LLaMA2-70B with CoT-Influx surpasses GPT-3.5 and a wide range of larger LLMs (PaLM, Minerva 540B, etc.) on the GSM8K. CoT-Influx is a plug-and-play module for LLMs, adaptable in various scenarios. It's compatible with advanced reasoning prompting techniques, such as self-consistency, and supports different long-context LLMs, including Mistral-7B-v0.3-32K and Yi-6B-200K.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Improve Students Reasoning Generalizability through Cascading Decomposed CoTs Distillation

Chengwei Dai, Kun Li, Wei Zhou, Songlin Hu

Large language models (LLMs) exhibit enhanced reasoning at larger scales, driving efforts to distill these capabilities into smaller models via teacher-student learning. Previous works simply fine-tune student models on teachers' generated Chain-of-Thoughts (CoTs) data. Although these methods enhance in-domain (IND) reasoning performance, they struggle to generalize to out-of-domain (OOD) tasks. We believe that the widespread spurious correlations between questions and answers may lead the model to present a specific answer which restricts the diversity and generalizability of its reasoning process. In this paper, we propose Cascading Decomposed CoTs Distillation (CasCoD) to address these issues by decomposing the traditional single-step learning process into two cascaded learning steps. Specifically, by restructuring the training objectives removing the answer from outputs and concatenating the question with the rationale as input, CasCoD's two-step learning process ensures that students focus on learning rationales without interference from the present answers, thus improving reasoning generalizability. Extensive experiments demonstrate the effectiveness of CasCoD on both IND and OOD benchmark reasoning datasets⁵.

Nov 12 (Tue) 16:00-17:30 - Jasmine

InfiniPot: Infinite Context Processing on Memory-Constrained LLMs

Minsoo Kim, Kyuhong Shim, Jungwook Choi, Simyoung Chang

Handling long input contexts remains a significant challenge for Large Language Models (LLMs), particularly in resource-constrained environments such as mobile devices. Our work aims to address this limitation by introducing InfiniPot, a novel KV cache control framework designed to enable pre-trained LLMs to manage extensive sequences within fixed memory constraints efficiently, without requiring additional training. InfiniPot leverages Continual Context Distillation (CCD), an iterative process that compresses and retains essential information through novel importance metrics, effectively maintaining critical data even without access to future context. Our comprehensive evaluations indicate that InfiniPot significantly outperforms models trained for long contexts in various NLP tasks, establishing its efficacy and versatility. This work represents a substantial advancement toward making LLMs applicable to a broader range of real-world scenarios.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Exploring Step Efficiency in a Tree-based Linear Model for Extreme Multi-label Classification

He-Zhe Lin, Cheng-Hung Liu, Chih-Jen Lin

Extreme multi-label classification (XMC) aims to identify relevant subsets from numerous labels. Among the various approaches for XMC, tree-based linear models are effective due to their superior efficiency and simplicity. However, the space complexity of free-based methods is not well-studied. Many past works assume that storing the model is not affordable and apply techniques such as pruning to save space, which may lead to performance loss. In this work, we conduct both theoretical and empirical analyses on the space to store a tree model under the assumption of sparse data, a condition frequently met in text data. We found that, some features may be unused when training binary classifiers in a tree method, resulting in zero values in the weight vectors. Hence, storing only non-zero elements can greatly save space. Our experimental results indicate that tree models can require less than 10% of the size of the standard one-vs-rest method for multi-label text classification. Our research provides a simple procedure to estimate the size of a tree model before training any classifier in the tree nodes. Then, if the model size is already acceptable, this approach can help avoid modifying the model through weight pruning or other techniques.

Nov 12 (Tue) 16:00-17:30 - Jasmine

FFN-SkipLLM: A Hidden Gem for Autoregressive Decoding with Adaptive Feed Forward Skipping

AJAY KUMAR JAISWAL, Bodan Hu, Lu Yin, Yeonju Ro, Tianlong Chen, Shiwei Liu, Aditya Akella

Autoregressive Large Language Models (e.g., LLaMa, GPTs) are omnipresent achieving remarkable success in language understanding and

⁵Code available at <https://github.com/C-W-D/CasCoD>

generation. However, such impressive capability typically comes with a substantial model size, which presents significant challenges for autoregressive token-by-token generation. To mitigate computation overload incurred during generation, several early-exit and layer-dropping strategies have been proposed. Despite some promising success due to the redundancy across LLMs layers on metrics like Rouge-L/BLUE, our careful knowledge-intensive evaluation unveils issues such as generation collapse, hallucination, and noticeable performance drop even at the trivial exit ratio of 10-15% of layers. We attribute these errors primarily to ineffective handling of the KV cache through state copying during early exit. In this work, we observe the saturation of computationally expensive feed-forward blocks of LLM layers and propose FFN-SkipLLM, which is a novel fine-grained skip strategy for autoregressive LLMs. FFN-SkipLLM leverages an input-adaptive feed-forward skipping approach that can skip 25-30% of FFN blocks of LLMs with marginal change in performance on knowledge-intensive generation tasks without any requirement to handle the KV cache. Our extensive experiments and ablation studies across benchmarks like MT-Bench, Factoid-QA, and variable-length text summarization illustrate how our simple and easy-to-use method can facilitate faster autoregressive decoding.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

LLoCO: Learning Long Contexts Offline

Sijun Tan, Xuyu Li, Shishir G Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E. Gonzalez, Raluca Popa

Processing long contexts remains a challenge for large language models (LLMs) due to the quadratic computational and memory overhead of the self-attention mechanism and the substantial KV cache sizes during generation. We propose LLoCO, a novel approach to address this problem by learning contexts offline through context compression and in-domain parameter-efficient finetuning with LoRA. Our method enables an LLM to create a concise representation of the original context and efficiently retrieve relevant information to answer questions accurately. Our approach extends the effective context window of a 4k token LLaMA2-7B model to handle up to 128k tokens. We evaluate our approach on several long-context question-answering datasets, demonstrating that LLoCO significantly outperforms in-context learning while using 30× fewer tokens during inference. LLoCO achieves up to 7.62× speed-up during inference and 11.52× higher throughput during finetuning, substantially reduces the cost of long document question answering. This makes it a promising solution for efficient long context processing.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Mentor-KD: Making Small Language Models Better Multi-step Reasoners

Hojae Lee, Junho Kim, SangKeun Lee

Large Language Models (LLMs) have displayed remarkable performances across various complex tasks by leveraging Chain-of-Thought (CoT) prompting. Recently, studies have proposed a Knowledge Distillation (KD) approach, reasoning distillation, which transfers such reasoning ability of LLMs through fine-tuning language models of multi-step rationales generated by LLM teachers. However, they have inadequately considered two challenges regarding insufficient distillation sets from the LLM teacher model, in terms of 1) data quality and 2) soft label provision. In this paper, we propose Mentor-KD, which effectively distills the multi-step reasoning capability of LLMs to smaller LMs while addressing the aforementioned challenges. Specifically, we exploit a mentor, intermediate-sized task-specific fine-tuned model, to augment additional CoT annotations and provide soft labels for the student model during reasoning distillation. We conduct extensive experiments and confirm Mentor-KD's effectiveness across various models and complex reasoning tasks.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Is C4 Dataset Enough for Pruning? An Investigation of Calibration Data for LLM Pruning

Abhinav Bandari, Lu Yin, Cheng-Yu Hsieh, AJAY KUMAR JAISWAL, Tianlong Chen, Li Shen, Ranjay Krishna, Shiwei Liu

Network pruning has emerged as a potential solution to make LLMs cheaper to deploy. However, existing LLM pruning approaches universally rely on the C4 dataset as the calibration data for calculating pruning scores, leaving its optimality unexplored. In this study, we evaluate the choice of calibration data on LLM pruning, across a wide range of datasets that are most commonly used in LLM training and evaluation, including four pertaining datasets as well as three categories of downstream tasks encompassing nine datasets. Each downstream dataset is prompted with In-Context Learning (ICL) or Chain-of-Thought (CoT), respectively. Besides the already intriguing observation that the choice of calibration data significantly impacts the performance of pruned LLMs, our results also uncover several subtle and often unexpected findings, summarized as follows: (1) C4 is not the optimal choice for LLM pruning, even among commonly used pre-training datasets; (2) arithmetic datasets when used as calibration data performs on par or even better than pre-training datasets; (3) pruning with downstream datasets does not necessarily help the corresponding downstream task, compared to pre-training data; (4) ICL is widely beneficial to all data categories, whereas CoT is only useful on certain tasks. Our findings shed light on the importance of carefully selecting calibration data for LLM pruning and pave the way for more efficient deployment of these powerful models in real-world applications. We release our code at: <https://github.com/abx393/lm-pruning-calibration-data>.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

PALM: Few-Shot Prompt Learning for Audio Language Models

Asif Hanif, Maha Tufail Agro, Mohammad Areeb Qazi, Hanan Aldarmaki

Audio-Language Models (ALMs) have recently achieved remarkable success in zero-shot audio recognition tasks, which match features of audio waveforms with class-specific text prompt features, inspired by advancements in Vision-Language Models (VLMs). Given the sensitivity of zero-shot performance to the choice of hand-crafted text prompts, many prompt learning techniques have been developed for VLMs. We explore the efficacy of these approaches in ALMs and propose a novel method, Prompt Learning in Audio Language Models (PALM), which optimizes the feature space of the text encoder branch. Unlike existing methods that work in the input space, our approach results in greater training efficiency. We demonstrate the effectiveness of our approach on 11 audio recognition datasets, encompassing a variety of speech-processing tasks, and compare the results with three baselines in a few-shot learning setup. Our method is either on par with or outperforms other approaches while being computationally less demanding. Our code is publicly available at <https://asif-hanif.github.io/palm/>.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

SpecHub: Provable Acceleration to Multi-Draft Speculative Decoding

Hanchi Sun, Tianyi Zhou, Xun Chen, Lichao Sun

Large Language Models (LLMs) have become essential in advancing natural language processing (NLP) tasks, but their sequential token generation limits inference speed. Multi-Draft Speculative Decoding (MDSD) offers a promising solution by using a smaller draft model to generate multiple token sequences, which the target LLM verifies in parallel. However, current heuristic approaches, such as Recursive Rejection Sampling (RRS), suffer from low acceptance rates in subsequent drafts, limiting the advantages of using multiple drafts. Meanwhile, Optimal Transport with Membership Cost (OTM) can theoretically improve acceptance rates, but its computational cost is too high for real-time use. We present SpecHub, a novel, efficient sampling-verification method for MDSD that improves acceptance rates with only linear computational overhead. By simplifying the OTM problem into a compact Linear Programming model, SpecHub significantly reduces computational complexity. It further accelerates sampling by leveraging a sparse joint distribution, focusing computation on high-probability token sequences. It integrates seamlessly into existing MDSD frameworks. In extensive experiments, SpecHub consistently generates 0.05-0.27 and 0.02-0.16 more tokens per step than RRS and RRS without replacement. We attach our code at <https://github.com/MasterGodzilla/Specula>.

tive_decoding_OT.

Nov 12 (Tue) 16:00-17:30 - Jasmine

ApiQ: Finetuning of 2-Bit Quantized Large Language Model

Baohao Liao, Christian Herold, Shahram Khadivi, Christof Monz

Memory-efficient finetuning of large language models (LLMs) has recently attracted huge attention with the increasing size of LLMs, primarily due to the constraints posed by GPU memory limitations and the effectiveness of these methods compared to full finetuning. Despite the advancements, current strategies for memory-efficient finetuning, such as QLoRA, exhibit inconsistent performance across diverse bit-width quantizations and multifaceted tasks. This inconsistency largely stems from the detrimental impact of the quantization process on preserved knowledge, leading to catastrophic forgetting and undermining the utilization of pretrained models for finetuning purposes. In this work, we introduce a novel quantization framework named ApiQ, designed to restore the lost information from quantization by concurrently initializing the LoRA components and quantizing the weights of LLMs. This approach ensures the maintenance of the original LLM's activation precision while mitigating the error propagation from shallower into deeper layers. Through comprehensive evaluations conducted on a spectrum of language tasks with various LLMs, ApiQ demonstrably minimizes activation error during quantization. Consequently, it consistently achieves superior finetuning results across various bit-widths. Notably, one can even finetune a 2-bit Llama-2-70b with ApiQ on a single NVIDIA A100-80GB GPU without any memory-saving techniques, and achieve promising results.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Memory-Efficient Fine-Tuning of Transformers via Token Selection

Antoine Simoulin, Namyoung Park, Xiaoyi Liu, Grey Yang

Fine-tuning provides an effective means to specialize pre-trained models for various downstream tasks. However, fine-tuning often incurs high memory overhead, especially for large transformer-based models, such as LLMs. While existing methods may reduce certain parts of the memory required for fine-tuning, they still require caching all intermediate activations computed in the forward pass to update weights during the backward pass. In this work, we develop TokenTune, a method to reduce memory usage, specifically the memory to store intermediate activations, in the fine-tuning of transformer-based models. During the backward pass, TokenTune approximates the gradient computation by backpropagating through just a subset of input tokens. Thus, with TokenTune, only a subset of intermediate activations are cached during the forward pass. Also, TokenTune can be easily combined with existing methods like LoRA, further reducing the memory cost. We evaluate our approach on pre-trained transformer models with up to billions of parameters, considering the performance on multiple downstream tasks such as text classification and question answering in a few-shot learning setup. Overall, TokenTune achieves performance on par with full fine-tuning or representative memory-efficient fine-tuning methods, while greatly reducing the memory footprint, especially when combined with other methods with complementary memory reduction mechanisms. We hope that our approach will facilitate the fine-tuning of large transformers, in specializing them for specific domains or co-training them with other neural components from a larger system. Our code is available at <https://github.com/facebookresearch/tokentune>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Sprout: Green Generative AI with Carbon-Efficient LLM Inference

Baolin Li, Yankai Jiang, Vijay Gadepally, Devesh Tiwari

The rapid advancement of generative AI has heightened environmental concerns, particularly regarding carbon emissions. Our framework, Sprout, addresses these challenges by reducing the carbon footprint of inference in large language models (LLMs). Sprout introduces "generation directives" to guide the autoregressive generation process, achieving a balance between ecological sustainability and high-quality outputs. By employing a strategic optimizer for directive assignment and a novel offline quality evaluator, Sprout reduces the carbon footprint of generative LLM inference by over 40% in real-world evaluations, using the Llama model and global electricity grid data. This work is crucial as the rising interest in inference time compute scaling laws amplifies environmental concerns, emphasizing the need for eco-friendly AI solutions.

Nov 12 (Tue) 16:00-17:30 - Jasmine

DisGeM: Distractor Generation for Multiple Choice Questions with Span Masking

Devrim Çavuolu, Seçil en, Ula Sert

Recent advancements in Natural Language Processing (NLP) have impacted numerous sub-fields such as natural language generation, natural language inference, question answering, and more. However, in the field of question generation, the creation of distractors for multiple-choice questions (MCQ) remains a challenging task. In this work, we present a simple, generic framework for distractor generation using readily available Pre-trained Language Models (PLMs). Unlike previous methods, our framework relies solely on pre-trained language models and does not require additional training on specific datasets. Building upon previous research, we introduce a two-stage framework consisting of candidate generation and candidate selection. Our proposed distractor generation framework outperforms previous methods without the need for training or fine-tuning. Human evaluations confirm that our approach produces more effective and engaging distractors. The related codebase is publicly available at <https://github.com/obss/disgem>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

QEFT: Quantization for Efficient Fine-Tuning of LLMs

Changhun Lee, Jun-gyu Jin, YoungHyun Cho, Eunhyeok Park

With the rapid growth in the use of fine-tuning for large language models (LLMs), optimizing fine-tuning while keeping inference efficient has become highly important. However, this is a challenging task as it requires improvements in all aspects, including inference speed, fine-tuning speed, memory consumption, and, most importantly, model quality. Previous studies have attempted to achieve this by combining quantization with fine-tuning, but they have failed to enhance all four aspects simultaneously. In this study, we propose a new lightweight technique called Quantization for Efficient Fine-Tuning (QEFT). QEFT accelerates both inference and fine-tuning, is supported by robust theoretical foundations, offers high flexibility, and maintains good hardware compatibility. Our extensive experiments demonstrate that QEFT matches the quality and versatility of full-precision parameter-efficient fine-tuning, while using fewer resources. Our code is available at <https://github.com/xvyyaward/qeft>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Long Sequence Modeling with Attention Tensorization: From Sequence to Tensor Learning

Aosong Feng, Rex Ying, Leandros Tassios

As the demand for processing extended textual data grows, the ability to handle long-range dependencies and maintain computational efficiency is more critical than ever. One of the key issues for long-sequence modeling using attention-based model is the mismatch between the limited-range modeling power of full attention and the long-range token dependency in the input sequence. In this work, we propose to scale up the attention receptive field by tensorizing long input sequences into compact tensor representations followed by attention on each transformed dimension. The resulting Tensorized Attention can be adopted as efficient transformer backbones to extend input context length with improved memory and time efficiency. We show that the proposed attention tensorization encodes token dependencies as a multi-hop attention process, and is equivalent to Kronecker decomposition of full attention. Extensive experiments show that tensorized attention can be

used to adapt pretrained LLMs with improved efficiency. Notably, using customized Triton kernels, tensorization enables Llama-8B training with 32,768 context length and can steadily extrapolate to 128k length during inference with 11 times speedup (compared to full attention with FlashAttention-2).

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

Merge to Learn: Efficiently Adding Skills to Language Models with Model Merging

Jacob Morrison, Noah A. Smith, Hanneh Hajishirzi, Pang Wei Koh, Jesse Dodge, Pradeep Dasigi

Adapting general-purpose language models to new skills is currently an expensive process that must be repeated as new instruction datasets targeting new skills are created, or can cause the models to forget older skills. In this work, we investigate the effectiveness of adding new skills to preexisting models by training on the new skills in isolation and later merging with the general model (e.g. using task vectors). In experiments focusing on scientific literature understanding, safety, and coding, we find that the parallel-train-then-merge procedure, which is significantly cheaper than retraining the models on updated data mixtures, is often comparably effective. Our experiments also show that parallel training is especially well-suited for enabling safety features in LMs relative to continued finetuning and retraining, as it dramatically improves model compliance with safe prompts while preserving its ability to refuse dangerous or harmful prompts.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

QUEST: Efficient Extreme Multi-Label Text Classification with Large Language Models on Commodity Hardware

Chuang Zhou, Junnan Dong, Xiao Huang, Zirui Liu, Kaixiong Zhou, Zhaozhuo Xu

Extreme multi-label text classification (EMTC) involves predicting multiple labels from a vast pool of candidates based on a user's textual query. While traditional BERT-based methods have shown limited success, large language models (LLMs) have brought new possibilities. It is promising to leverage their remarkable comprehension ability to understand textual queries. However, implementing LLMs is non-trivial for two main reasons. Firstly, real-world EMTC datasets can be extremely large, with candidate product pairs reaching up to ten million in real-world scenarios, which poses significant challenges in data ingestion. Secondly, the large size of LLMs makes computation and memory demands prohibitive for EMTC applications. To this end, we propose QUEST, a Quantized and Efficient Learning with Sampling Technique. QUEST includes a tailored hash sampling module that reduces the data volume to one-fourth of its original size. Additionally, we perform compressive fine-tuning LLMs with only twenty thousand trainable parameters, largely reducing computational requirements. Extensive experiments demonstrate that QUEST outperforms existing methods while requiring fewer computational resources, unlocking efficient EMTC on commodity hardware such as a single Nvidia RTX 3090 GPU with 24 GB of memory.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

LOOK-M: Look-Once Optimization in KV Cache for Efficient Multimodal Long-Context Inference

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, Li Yuan

Long-context Multimodal Large Language Models (MLLMs) demand substantial computational resources for inference as the growth of their multimodal Key-Value (KV) cache, in response to increasing input lengths, challenges memory and time efficiency. Unlike single-modality LLMs that manage only textual contexts, the KV cache of long-context MLLMs includes representations from multiple images with temporal and spatial relationships and related textual contexts. The predominance of image tokens means traditional optimizations for LLMs' KV caches are unsuitable for multimodal long-context settings, and no prior works have addressed this challenge. In this work, we introduce **LOOK-M***, a pioneering, fine-tuning-free approach that efficiently reduces the multimodal KV cache size while maintaining performance comparable to a full cache. We observe that during prompt prefill, the model prioritizes more textual attention over image features, and based on the multimodal interaction observation, a new proposed text-prior method is explored to compress the KV cache. Furthermore, to mitigate the degradation of image contextual information, we propose several compensatory strategies using KV pairs merging. **LOOK-M*** demonstrates that with a significant reduction in KV Cache memory usage, such as reducing it by **80%** in some cases, it not only achieves approximately **1.3x** faster decoding but also maintains or even **enhances** performance across a variety of long context multimodal tasks.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

OneGen: Efficient One-Pass Unified Generation and Retrieval for LLMs

Jintian Zhang, Cheng Peng, Mengshu Sun, Xiang Chen, Lei Liang, Zhiqiang Zhang, JUN ZHOU, Huajun Chen, Ningyu Zhang

Despite the recent advancements in Large Language Models (LLMs), which have significantly enhanced the generative capabilities for various NLP tasks, LLMs still face limitations in directly handling retrieval tasks. However, many practical applications demand the seamless integration of both retrieval and generation. This paper introduces a novel and efficient One-pass Generation and retrieval framework (OneGen), designed to improve LLMs' performance on tasks that require both generation and retrieval. The proposed framework bridges the traditionally separate training approaches for generation and retrieval by incorporating retrieval tokens generated autoregressively. This enables a single LLM to handle both tasks simultaneously in a unified forward pass. We conduct experiments on two distinct types of composite tasks, RAG and Entity Linking, to validate the pluggability, effectiveness, and efficiency of OneGen in training and inference. Furthermore, our results show that integrating generation and retrieval within the same context preserves the generative capabilities of LLMs while improving retrieval performance. To the best of our knowledge, OneGen is the first to enable LLMs to conduct vector retrieval during the generation.

Summarization

Nov 12 (Tue) 16:00-17:30 - Room: *Jasmine*

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, Hwanhee Lee

Through the advent of pre-trained language models, there have been notable advancements in abstractive summarization systems. Simultaneously, a considerable number of novel methods for evaluating factual consistency in abstractive summarization systems has been developed. But these evaluation approaches incorporate substantial limitations, especially on refinement and interpretability. In this work, we propose highly effective and interpretable factual inconsistency detection method FIZZ (Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document) for abstractive summarization systems that is based on fine-grained atomic facts decomposition. Moreover, we align atomic facts decomposed from the summary with the source document through adaptive granularity expansion. These atomic facts represent a more fine-grained unit of information, facilitating detailed understanding and interpretability of the summary's factual inconsistency. Experimental results demonstrate that our proposed factual consistency checking system significantly outperforms existing systems. We release the code at <https://github.com/plm332/FIZZ>.

Nov 12 (Tue) 16:00-17:30 - *Jasmine*

When Reasoning Meets Information Aggregation: A Case Study with Sports Narratives

Yeben Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Wenlin Yao, Hassan Foroosh, Dong Yu, Fei Liu

Reasoning is most powerful when an LLM accurately aggregates relevant information. We examine the critical role of information aggregation in reasoning by requiring the LLM to analyze sports narratives. To succeed at this task, an LLM must infer points from actions, identify related entities, attribute points accurately to players and teams, and compile key statistics to draw conclusions. We conduct comprehensive experiments with real NBA basketball data and present SportsGen, a new method to synthesize game narratives. By synthesizing data, we can rigorously evaluate LLMs' reasoning capabilities under complex scenarios with varying narrative lengths and density of information. Our findings show that most models, including GPT-4o, often fail to accurately aggregate basketball scores due to frequent scoring patterns. Open-source models like Llama-3 further suffer from significant score hallucinations. Finally, the effectiveness of reasoning is influenced by narrative complexity, information density, and domain-specific terms, highlighting the challenges in analytical reasoning tasks.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems

Philippe Laban, Alexander Fabri, Caiming Xiong, Chien-Sheng Wu

LLMs and RAG systems are now capable of handling millions of input tokens or more. However, evaluating the output quality of such systems on long-context tasks remains challenging, as tasks like Needle-in-a-Haystack lack complexity. In this work, we argue that summarization can play a central role in such evaluation. We design a procedure to synthesize Haystacks of documents, ensuring that specific insights repeat across documents. The "Summary of a Haystack" (SummHay) task requires a system to process the Haystack and generate, given a query, a summary that identifies the relevant insights and precisely cites the source documents. Since we have precise knowledge of what insights should appear in a haystack summary and what documents should be cited, we implement a highly reproducible automatic evaluation that can score summaries on two aspects – Coverage and Citation. We generate Haystacks in two domains (conversation, news), and perform a large-scale evaluation of 10 LLMs and corresponding 50 RAG systems. Our findings indicate that SummHay is an open challenge for current systems, as even systems provided with an Oracle signal of document relevance lag our estimate of human performance (56%) by 10+ points on a Joint Score. Without a retriever, long-context LLMs like GPT-4o and Claude 3 Opus score below 20% on SummHay. We show SummHay can also be used to study enterprise RAG systems and position bias in long-context models. We hope future systems can equal and surpass human performance on SummHay.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Related Work and Citation Text Generation: A Survey

Xiangci Li, Jessica Ouyang

To convince readers of the novelty of their research paper, authors must perform a literature review and compose a coherent story that connects and relates prior works to the current work. This challenging nature of literature review writing makes automatic related work generation (RWG) academically and computationally interesting, and also makes it an excellent test bed for examining the capability of SOTA natural language processing (NLP) models. Since the initial proposal of the RWG task, its popularity has waxed and waned, following the capabilities of mainstream NLP approaches. In this work, we survey the zoo of RWG historical works, summarizing the key approaches and task definitions and discussing the ongoing challenges of RWG.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Learning to Rank Salient Content for Query-focused Summarization

Sajad Sotudeh, Nazli Goharian

This study examines the potential of integrating Learning-to-Rank (LTR) with Query-focused Summarization (QFS) to enhance the summary relevance via content prioritization. Using a shared secondary decoder with the summarization decoder, we carry out the LTR task at the segment-level. Compared to the state-of-the-art, our model outperforms on QMSum benchmark (all metrics) and matches on SQuALITY benchmark (2 metrics) as measured by Rouge and BertScore while offering a lower training overhead. Specifically, on the QMSum benchmark, our proposed system achieves improvements, particularly in Rouge-L (+0.42) and BertScore (+0.34), indicating enhanced understanding and relevance. While facing minor challenges in Rouge-1 and Rouge-2 scores on the SQuALITY benchmark, the model significantly excels in Rouge-L (+1.47), underscoring its capability to generate coherent summaries. Human evaluations emphasize the efficacy of our method in terms of relevance and faithfulness of the generated summaries, without sacrificing fluency. A deeper analysis reveals our model's superiority over the state-of-the-art for broad queries, as opposed to specific ones, from a qualitative standpoint. We further present an error analysis of our model, pinpointing challenges faced and suggesting potential directions for future research in this field.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Are Large Language Models In-Context Personalized Summarizers? Get an iCOPERNICUS Test Done!

Divya Patel, Pathik Patel, Ankush Chander, Sourish Dasgupta, Tammo Chakraborty

Large Language Models (LLMs) have succeeded considerably in In-Context-Learning (ICL) based summarization. However, saliency is subject to the users' specific preference histories. Hence, we need reliable *In-Context Personalization Learning* (ICPL) capabilities within such LLMs. For any arbitrary LLM to exhibit ICPL, it needs to have the **ability to discern contrast in user profiles**. A recent study proposed a measure for *degree-of-personalization* called EGISES for the first time. EGISES measures a model's responsiveness to user profile differences. However, it cannot test if a model utilizes all three types of cues provided in ICPL prompts: (i) example summaries, (ii) user's reading histories, and (iii) contrast in user profiles. To address this, we propose the iCOPERNICUS framework, a novel *In-Context Personalization Learning Scrutiny* of Summarization capability in LLMs that uses EGISES as a comparative measure. As a case-study, we evaluate 17 state-of-the-art LLMs based on their reported ICL performances and observe that 15 models' ICPL degrades (min: 1.6%↓; max: 3.6%↓) when probed with richer prompts, thereby showing lack of *true* ICPL.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Model-based Preference Optimization in Abstractive Summarization without Human Feedback

Jaepill Choi, Kyubyoung Chae, Jiwoo Song, Yohan Jo, Taesup Kim

In abstractive summarization, the challenge of producing concise and accurate summaries arises from the vast amount of information contained in the source document. Consequently, although Large Language Models (LLMs) can generate fluent text, they often introduce inaccuracies by hallucinating content not found in the original source. While supervised fine-tuning methods that maximize likelihood contribute to this issue, they do not consistently enhance the faithfulness of the summaries. Preference-based optimization methods, such as Direct Preference Optimization (DPO), can further refine the model to align with human preferences. However, these methods still heavily depend on costly human feedback. In this work, we introduce a novel and straightforward approach called Model-based Preference Optimization (MPO) to fine-tune LLMs for improved summarization abilities without any human feedback. By leveraging the model's inherent summarization capabilities, we create a preference dataset that is fully generated by the model using different decoding strategies. Our experiments on standard summarization datasets and various metrics demonstrate that our proposed MPO significantly enhances the quality of generated summaries without relying on human feedback. The code is publicly available at <https://github.com/cjaep/MPO>.

Nov 12 (Tue) 16:00-17:30 - Jasmine

SYNFACT-EDIT: Synthetic Imitation Edit Feedback for Factual Alignment in Clinical Summarization

Prakanya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, hong yu
Large Language Models (LLMs) such as GPT & Llama have demonstrated significant achievements in summarization tasks but struggle with factual inaccuracies, a critical issue in clinical NLP applications where errors could lead to serious consequences. To counter the high costs and limited availability of expert-annotated data for factual alignment, this study introduces an innovative pipeline that utilizes >100B parameter GPT variants like GPT-3.5 & GPT-4 to act as synthetic experts to generate high-quality synthetics feedback aimed at enhancing factual consistency in clinical note summarization. Our research primarily focuses on edit feedback generated by these synthetic feedback experts without additional human annotations, mirroring and optimizing the practical scenario in which medical professionals refine AI system outputs. Although such 100B+ parameter GPT variants have proven to demonstrate expertise in various clinical NLP tasks, such as the Medical Licensing Examination, there is scant research on their capacity to act as synthetic feedback experts and deliver expert-level edit feedback for improving the generation quality of weaker (<10B parameter) LLMs like GPT-2 (1.5B) & Llama 2 (7B) in clinical domain. So in this work, we leverage 100B+ GPT variants to act as synthetic feedback experts offering expert-level edit feedback, that is used to reduce hallucinations and align weaker (<10B parameter) LLMs with medical facts using two distinct alignment algorithms (DPO & SALT), endeavoring to narrow the divide between AI-generated content and factual accuracy. This highlights the substantial potential of LLM-based synthetic edit in enhancing the alignment of clinical factuality.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Whats under the hood: Investigating Automatic Metrics on Meeting Summarization

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, Bela Gipp

Meeting summarization has become a critical task considering the increase in online interactions. Despite new techniques being proposed regularly, the evaluation of meeting summarization techniques relies on metrics not tailored to capture meeting-specific errors, leading to ineffective assessment. This paper explores what established automatic metrics capture and the errors they mask by correlating metric scores with human evaluations across a comprehensive error taxonomy. We start by reviewing the literature on English meeting summarization to identify key challenges, such as speaker dynamics and contextual turn-taking, and error types, including missing information and linguistic inaccuracy, concepts previously loosely defined in the field. We then examine the relationship between these challenges and errors using human annotated transcripts and summaries from encoder-decoder-based and autoregressive Transformer models on the QMSum dataset. Experiments reveal that different model architectures respond variably to the challenges, resulting in distinct links between challenges and errors. Current established metrics struggle to capture the observable errors, showing weak to moderate correlations, with a third of the correlations indicating error masking. Only a subset of metrics accurately reacts to specific errors, while most correlations show either unresponsiveness or failure to reflect the error's impact on summary quality.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Event-Keyed Summarization

William Gant, Alexander Martin, Pavlo Kuchmichechuk, Aaron Steven White

We introduce “event-keyed summarization” (EKS), a novel task that marries traditional summarization and document-level event extraction, with the goal of generating a contextualized summary for a specific event, given a document and an extracted event structure. We introduce a dataset for this task, MUCSUM, consisting of summaries of all events in the classic MUC-4 dataset, along with a set of baselines that comprises both pretrained LM standards in the summarization literature, as well as larger frontier models. We show that ablations that reduce EKS to traditional summarization or structure-to-text yield inferior summaries of target events and that MUCSUM is a robust benchmark for this task. Lastly, we conduct a human evaluation of both reference and model summaries, and provide some detailed analysis of the results.

Nov 12 (Tue) 16:00-17:30 - Jasmine

HealthAlignSumm : Utilizing Alignment for Multimodal Summarization of Code-Mixed Healthcare Dialogues

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, Setu Sinha

As generative AI progresses, collaboration be-tween doctors and AI scientists is leading to the development of personalized models to streamline healthcare tasks and improve productivity. Summarizing doctor-patient dialogues has be-come important, helping doctors understand conversations faster and improving patient care. While previous research has mostly focused on context data, incorporating visual cues from pa-tient interactions allows doctors to gain deeper insights into medical conditions. Most of this research has centered on English datasets, but real-world conversations often mix languages for better communication. To address the lack of resources for multimodal summarization of code-mixed dialogues in healthcare, we de veloped the MCDH dataset. Additionally, we created HealthAlignSumm, a new model that integrates visual components with the BART architecture. This represents a key advancement in multimodal fusion, applied within both the encoder and decoder of the BART model. Our work is the first to use alignment techniques, including state-of-the-art algorithms like DirectPreference Optimization, on encoder-decoder models with synthetic datasets for multimodal summarization. Through extensive experiments, we demonstrated the superior performance of HealthAlignSumm across several metrics validated by both automated assessments and human evaluations. The dataset MCDH and our proposed model HealthAlign-Summ will be available in this GitHub account: <https://github.com/AkashGhosh/HealthAlignSumm-Utilizing-Alignment-for-Multimodal-Summarization-of-Code-Mixed-Healthcare-Dialogues>. Disclaimer: This work involves medical im-agery based on the subject matter of the topic.

Nov 12 (Tue) 16:00-17:30 - Jasmine

A Decoding Algorithm Based on Directed Acyclic Transformers for Length-Control Summarization

Chenyang Huang, Hao Zhou, Cameron Jen, Kangjie Zheng, Osmar Zaiane, Lili Mou

Length-control summarization aims to condense long texts into a short one within a certain length limit. Previous approaches often use autoregressive (AR) models and treat the length requirement as a soft constraint, which may not always be satisfied. In this study, we propose a novel length-control decoding algorithm based on the directed acyclic Transformer (DAT). Our approach allows for multiple plausible sequence fragments and predicts a path to connect them. In addition, we propose a Sequence Maximum a Posteriori (Seq-MAP) decoding algorithm that marginalizes different possible paths and finds the most probable summary satisfying the length budget. Our algorithm is based on beam search, which further facilitates a reranker for performance improvement. Experimental results on the Gigaword dataset demonstrate our state-of-the-art performance for length-control summarization.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Learning to Refine with Fine-Grained Natural Language Feedback

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, Greg Durrett

Recent work has explored the capability of large language models (LLMs) to identify and correct errors in LLM-generated responses. These refinement approaches frequently evaluate what sizes of models are able to do refinement for what problems, but less attention is paid to what effective feedback for refinement looks like. In this work, we propose looking at refinement with feedback as a composition of three distinct LLM competencies: (1) detection of bad generations; (2) fine-grained natural language critique generation; (3) refining with fine-grained feedback. The first step can be implemented with a high-performing discriminative model and steps 2 and 3 can be implemented either via prompted or fine-tuned LLMs. A key property of the proposed Detect, Critique, Refine (“DCR”) method is that the step 2 critique model can give fine-grained feedback about errors, made possible by offloading the discrimination to a separate model in step 1. We show that models of

different capabilities benefit from refining with DCR on the task of improving factual consistency of document grounded summaries. Overall, DCR consistently outperforms existing end-to-end refinement approaches and current trained models not fine-tuned for factuality critiquing.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Unveiling Implicit Table Knowledge with Question-Then-Pinpoint Reasoner for Insightful Table Summarization

Kwangwook Seo, Jinyoung Yeo, Dongha Lee

Implicit knowledge hidden within the explicit table cells, such as data insights, is the key to generating a high-quality table summary. However, unveiling such implicit knowledge is a non-trivial task. Due to the complex nature of structured tables, it is challenging even for large language models (LLMs) to mine the implicit knowledge in an insightful and faithful manner. To address this challenge, we propose a novel table reasoning framework Question-then-Pinpoint. Our work focuses on building a plug-and-play table reasoner that can self-question the insightful knowledge and answer it by faithfully pinpointing evidence on the table to provide explainable guidance for the summarizer. To train a reliable reasoner, we collect table knowledge by guiding a teacher LLM to follow the coarse-to-fine reasoning paths and refine it through two quality enhancement strategies to selectively distill the high-quality knowledge to the reasoner. Extensive experiments on two table summarization datasets, including our newly proposed InsTaSumm, validate the general effectiveness of our framework.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Multi-Target Cross-Lingual Summarization: a novel task and a language-neutral approach

Diogo Pernes, Gonçalo M. Correia, Afonso Mendes

Cross-lingual summarization aims to bridge language barriers by summarizing documents in different languages. However, ensuring semantic coherence across languages is an overlooked challenge and can be critical in several contexts. To fill this gap, we introduce multi-target cross-lingual summarization as the task of summarizing a document into multiple target languages while ensuring that the produced summaries are semantically similar. We propose a principled re-ranking approach to this problem and a multi-criteria evaluation protocol to assess semantic coherence across target languages, marking a first step that will hopefully stimulate further research on this problem.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Attacks against Abstractive Text Summarization Models through Lead Bias and Influence Functions

Poojitha Thota, Shirin Nilizadeh

Large Language Models (LLMs) have introduced novel opportunities for text comprehension and generation. Yet, they are vulnerable to adversarial perturbations and data poisoning attacks, particularly in tasks like text classification and translation. However, the adversarial robustness of abstractive text summarization models remains less explored. In this work, we unveil a novel approach by exploiting the inherent lead bias in summarization models, to perform adversarial perturbations. Furthermore, we introduce an innovative application of influence functions, to execute data poisoning, which compromises the model's integrity. This approach not only shows a skew in the models' behavior to produce desired outcomes but also shows a new behavioral change, where models under attack tend to generate extractive summaries rather than abstractive summaries.

Nov 12 (Tue) 16:00-17:30 - Jasmine

SummaCoz: A Dataset for Improving the Interpretability of Factual Consistency Detection for Summarization

Ge Luo, Weisi Fan, Miaoan Li, Yuoruizhe Sun, Runlong Zhang, Chenyu Xu, Forrest Sheng Bao

Summarization is an important application of Large Language Models (LLMs). When judging the quality of a summary, factual consistency holds a significant weight. Despite numerous efforts dedicated to building factual inconsistency detectors, the exploration of explainability remains limited among existing effort. In this study, we incorporate both human-annotated and model-generated natural language explanations elucidating how a summary deviates and thus becomes inconsistent with its source article. We build our explanation-augmented dataset on top of the widely used SummaCz summarization consistency benchmark. Additionally, we develop an inconsistency detector that is jointly trained with the collected explanations. Our findings demonstrate that integrating explanations during training not only enables the model to provide rationales for its judgments but also enhances its accuracy significantly.

Nov 12 (Tue) 16:00-17:30 - Jasmine

Enhancing Incremental Summarization with Structured Representations

Eunjeong Hwang, Yichao Zhou, James Bradley Wendt, Beliz Günel, Nguyen Vo, Jing Xie, Sandeep Tata

Large language models (LLMs) often struggle with processing extensive input contexts, which can lead to redundant, inaccurate, or incoherent summaries. Recent methods have used unstructured memory to incrementally process these contexts, but they still suffer from information overload due to the volume of unstructured data handled. In our study, we introduce structured knowledge representations (GU.json), which significantly improve summarization performance by 40% and 14% across two public datasets. Most notably, we propose the Chain-of-Key strategy (CoK.json) that dynamically updates or augments these representations with new information, rather than recreating the structured memory for each new source. This method further enhances performance by 7% and 4% on the datasets.

Nov 12 (Tue) 16:00-17:30 - Jasmine

UniSumEval: Towards Unified, Fine-grained, Multi-dimensional Summarization Evaluation for LLMs

Yuhu Lee, Taewon Yun, Jason Cai, Hang Su, Hwanjun Song

Existing benchmarks for summarization quality evaluation often lack diverse input scenarios, focus on narrowly defined dimensions (e.g., faithfulness), and struggle with subjective and coarse-grained annotation schemes. To address these shortcomings, we create UniSumEval benchmark, which extends the range of input context (e.g., domain, length) and provides fine-grained, multi-dimensional annotations. We use AI assistance in data creation, identifying potentially hallucinogenic input texts, and also helping human annotators reduce the difficulty of fine-grained annotation tasks. With UniSumEval, we benchmark nine latest language models as summarizers, offering insights into their performance across varying input contexts and evaluation dimensions. Furthermore, we conduct a thorough comparison of SOTA automated summary evaluators. Our benchmark data will be available at <https://github.com/DISL-Lab/UniSumEval-v1.0>.

Session 06 - Nov 13 (Wed) 10:30-12:00

Demo

Nov 13 (Wed) 10:30-12:00 - Room: Riverfront Hall

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

PyMarian: Fast Neural Machine Translation and Evaluation in Python

Elijah Rippeth, Marcin Junczys-Dowmunt, Matt Post, Roman Grundkiewicz, Thamme Gowda

The deep learning language of choice these days is Python; measured by factors such as available libraries and technical support, it is hard to beat. At the same time, software written in lower-level programming languages like C++ retain advantages in speed. We describe a Python interface to Marian NMT, a C++-based training and inference toolkit for sequence-to-sequence models, focusing on machine translation. This interface enables models trained with Marian to be connected to the rich, wide range of tools available in Python. A highlight of the interface is the ability to compute state-of-the-art COMET metrics from Python but using Marian's inference engine, with a speedup factor of up to $7.8 \times$ the existing implementations. We also briefly spotlight a number of other integrations, including Jupyter notebooks, connection with prebuilt models, and a web app interface provided with the package. PyMarian is available in PyPI via `pip install pymarian`.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Translation Canvas: An Explainable Interface to Pinpoint and Analyze Translation Systems

Chinmay Dandekar, Lei Li, Sijia Ouyang, Wenda Xu, Xi Xu

With the rapid advancement of machine translation research, evaluation toolkits have become essential for benchmarking system progress. Tools like COMET and SacreBLEU offer single quality score assessments that are effective for pairwise system comparisons. However, these tools provide limited insights for fine-grained system-level comparisons and the analysis of instance-level defects. To address these limitations, we introduce **Translation Canvas**, an explainable interface designed to pinpoint and analyze translation systems' performance: 1) Translation Canvas assists machine translation researchers in comprehending system-level model performance by identifying common errors (their frequency and severity) and analyzing relationships between different systems based on various evaluation metrics. 2) It supports fine-grained analysis by highlighting error spans with explanations and selectively displaying systems' predictions. According to human evaluation, Translation Canvas demonstrates superior performance over COMET and SacreBLEU packages under enjoyability and understandability criteria.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

sign.mt: Real-Time Multilingual Sign Language Translation Application

Amit Moryossef

This paper presents *sign.mt*, an open-source application for real-time multilingual bi-directional translation between spoken and signed languages. Harnessing state-of-the-art open-source models, this tool aims to address the communication divide between the hearing and the deaf, facilitating seamless translation in both spoken-to-signed and signed-to-spoken translation directions. To provide reliable and unrestricted communication, *sign.mt* offers offline functionality, crucial in areas with limited internet connectivity. It enhances user engagement by providing customizable photorealistic sign language avatars, encouraging a more personalized and authentic user experience. Licensed under CC BY-NC-SA 4.0, *sign.mt* signifies an important stride towards open, inclusive communication. The app can be used and modified for personal and academic purposes and even supports a translation API, fostering integration into a wider range of applications. However, it is by no means a finished product. We invite the NLP community to contribute towards the evolution of *sign.mt*. Whether it be the integration of more refined models, the development of innovative pipelines, or user experience improvements, your contributions can propel this project to new heights. Available at <https://sign.mt>, it stands as a testament to what we can achieve together, as we strive to make communication accessible to all.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

WebOlympus: An Open Platform for Web Agents on Live Websites

Boyu Gou, Boyuan Zheng, Huan Sun, Scott Salisbury, Yu Su, Zheng Du

Web agents are emerging as powerful tools capable of performing complex tasks across diverse web environments. The rapid development of large multimodal models is further enhancing this advancement. However, there is a lack of standard, easy-to-use tools for research and development, as well as experimental platforms on live websites. To address this challenge, we present WebOlympus, an open platform for web agents operating on live websites. WebOlympus facilitates the deployment of web agents with various designs, providing an accessible toolkit for both researchers and engineers. To ensure the trustworthiness of web agents, WebOlympus incorporates a safety monitor module that prevents harmful actions through human supervision or model. Additionally, WebUI enables users without programming experience to utilize the platform through a chrome extension based UI. WebOlympus also supports diverse applications, including annotation interfaces for web agent trajectories and data crawling.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

ChatHF: Collecting Rich Human Feedback from Real-time Conversations

Alan Ritter, Andrew Li, Duong Minh Le, Wei Xu, Zhenduo Wang, Ethan Adrian Mendes

We introduce ChatHF, an interactive annotation framework for chatbot evaluation, which integrates configurable annotation within a chat interface. ChatHF can be flexibly configured to accommodate various chatbot evaluation tasks, for example detecting offensive content, identifying incorrect or misleading information in chatbot responses, and chatbot responses that might compromise privacy. It supports post-editing of chatbot outputs and supports visual inputs, in addition to an optional voice interface. ChatHF is suitable for collection and annotation of NLP datasets, and Human-Computer Interaction studies, as demonstrated in case studies on image geolocation and assisting older adults with daily activities.

Human-centered NLP 2

Nov 13 (Wed) 10:30-12:00 - Room: Riverfront Hall

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, Nancy F. Chen

Intelligent Tutoring Systems (ITSs) can provide personalized and self-paced learning experience. The emergence of large language models (LLMs) further enables better human-machine interaction, and facilitates the development of conversational ITSs in various disciplines such as math and language learning. In dialogic teaching, recognizing and adapting to individual characteristics can significantly enhance student engagement and learning efficiency. However, characterizing and simulating student's persona remain challenging in training and evaluating conversational ITSs. In this work, we propose a framework to construct profiles of different student groups by refining and integrating both cognitive and noncognitive aspects, and leverage LLMs for personality-aware student simulation in a language learning scenario. We further enhance the framework with multi-aspect validation, and conduct extensive analysis from both teacher and student perspectives. Our experimental results show that state-of-the-art LLMs can produce diverse student responses according to the given language ability and personality traits, and trigger teacher's adaptive scaffolding strategies.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration

Shangbin Feng, Taylor Sorensen, Yuhua Liu, Jillian Fisher, Chan Young Park, Yejin Choi, Yulia Tsvetkov

While existing alignment paradigms have been integral in developing large language models (LLMs), LLMs often learn an averaged human preference and struggle to model diverse preferences across cultures, demographics, and communities. We propose Modular Pluralism, a modular framework based on multi-LLM collaboration for pluralistic alignment: it "plugs into" a base LLM a pool of smaller but specialized community LMs, where models collaborate in distinct modes to flexibility support three modes of pluralism: Overton, steerable, and distributional. Modular Pluralism is uniquely compatible with black-box LLMs and offers the modular control of adding new community LMs for previously underrepresented communities. We evaluate Modular Pluralism with six tasks and four datasets featuring questions/instructions with value-laden and perspective-informed responses. Extensive experiments demonstrate that Modular Pluralism advances the three pluralism objectives across six black-box and open-source LLMs. Further analysis reveals that LLMs are generally faithful to the inputs from smaller community LLMs, allowing seamless patching by adding a new community LM to better cover previously underrepresented communities.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Visual Prompting in LLMs for Enhancing Emotion Recognition

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Yang Liu, Zhenyu Qin, Wenjia Niu, Sabrina Caldwell, Tom Gedeon

Vision Large Language Models (VLLMs) are transforming the intersection of computer vision and natural language processing; however, the potential of using visual prompts for emotion recognition in these models remains largely unexplored and untapped. Traditional methods in VLLMs struggle with spatial localization and often discard valuable global context. We propose a novel Set-of-Vision prompting (SoV) approach that enhances zero-shot emotion recognition by using spatial information, such as bounding boxes and facial landmarks, to mark targets precisely. SoV improves accuracy in face count and emotion categorization while preserving the enriched image context. Through comprehensive experimentation and analysis of recent commercial or open-source VLLMs, we evaluate the SoV model's ability to comprehend facial expressions in natural environments. Our findings demonstrate the effectiveness of integrating spatial visual prompts into VLLMs for improving emotion recognition performance.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

KidLM: Advancing Language Models for Children Early Insights and Future Directions

Mir Taseer Nayeen, Davood Rafiei

Recent studies highlight the potential of large language models in creating educational tools for children, yet significant challenges remain in maintaining key child-specific properties such as linguistic nuances, cognitive needs, and safety standards. In this paper, we explore foundational steps toward the development of child-specific language models, emphasizing the necessity of high-quality pre-training data. We introduce a novel user-centric data collection pipeline that involves gathering and validating a corpus specifically written for and sometimes by children. Additionally, we propose a new training objective, Stratified Masking, which dynamically adjusts masking probabilities based on our domain-specific child language data, enabling models to prioritize vocabulary and concepts more suitable for children. Experimental evaluations demonstrate that our model excels in understanding lower grade-level text, maintains safety by avoiding stereotypes, and captures children's unique preferences. Furthermore, we provide actionable insights for future research and development in child-specific language modeling.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

How Does the Disclosure of AI Assistance Affect the Perceptions of Writing?

Zhuoyan Li, Chen Liang, Jing Peng, Ming Yu

Recent advances in generative AI technologies like large language models have boosted the incorporation of AI assistance in writing workflows, leading to the rise of a new paradigm of human-AI co-creation in writing. To understand how people perceive writings that are produced under this paradigm, in this paper, we conduct an experimental study to understand whether and how the disclosure of the level and type of AI assistance in the writing process would affect people's perceptions of the writing on various aspects, including their evaluation on the quality of the writing, and their ranking of different writings. Our results suggest that disclosing the AI assistance in the writing process, especially if AI has provided assistance in generating new content, decreases the average quality ratings for both argumentative essays and creative stories. This decrease in the average quality ratings often comes with an increased level of variations in different individuals' quality evaluations of the same writing. Indeed, factors such as an individual's writing confidence and familiarity with AI writing assistants are shown to moderate the impact of AI assistance disclosure on their writing quality evaluations. We also find that disclosing the use of AI assistance may significantly reduce the proportion of writings produced with AI's content generation assistance among the top-ranked writings.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Annotation alignment: Comparing LLM and human annotations of conversational safety

Rajiv Movva, Pang Wei Koh, Emma Pierson

Do LLMs align with human perceptions of safety? We study this question via *annotation alignment*, the extent to which LLMs and humans agree when annotating the safety of user-chatbot conversations. We leverage the recent DICES dataset (Aroyo et al. 2023), in which 350 conversations are each rated for safety by 112 annotators spanning 10 race-gender groups. GPT-4 achieves a Pearson correlation of $r = 0.59$ with the average annotator rating, higher than the median annotator's correlation with the average ($r = 0.51$). We show that larger datasets are needed to resolve whether GPT-4 exhibits disparities in how well it correlates with different demographic groups. Also, there is substantial idiosyncratic variation in correlation within groups, suggesting that race & gender do not fully capture differences in alignment. Finally, we find that GPT-4 cannot predict when one demographic group finds a conversation more unsafe than another.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, Monica Lam

While language model (LM)-powered chatbots and generative search engines excel at answering concrete queries, discovering information in the terrain of unknown unknowns remains challenging for users. To emulate the common educational scenario where children/students learn by listening to and participating in conversations of their parents/teachers, we create Collaborative STORM (Co-STORM). Unlike QA systems that require users to ask all the questions, Co-STORM lets users observe and occasionally steer the discourse among several LM agents. The agents ask questions on the user's behalf, allowing the user to discover unknown unknowns serendipitously. To facilitate user interaction, Co-STORM assists users in tracking the discourse by organizing the uncovered information into a dynamic mind map, ultimately generating a comprehensive report as takeaways. For automatic evaluation, we construct the WildSeek dataset by collecting real information-seeking records with user goals. Co-STORM outperforms baseline methods on both discourse trace and report quality. In a further human evaluation, 70% of participants prefer Co-STORM over a search engine, and 78% favor it over a RAG chatbot.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CommVQA: Situating Visual Question Answering in Communicative Contexts

Nandita Shankar Naik, Christopher Potts, Elisa Kreiss

Current visual question answering (VQA) models tend to be trained and evaluated on image-question pairs in isolation. However, the questions people ask are dependent on their informational needs and prior knowledge about the image content. To evaluate how situating images within naturalistic contexts shapes visual questions, we introduce CommVQA, a VQA dataset consisting of images, image descriptions, real-world communicative scenarios where the image might appear (e.g., a travel website), and follow-up questions and answers conditioned on the scenario and description. CommVQA, which contains 1000 images and 8,949 question-answer pairs, poses a challenge for current models. Error analyses and a human-subjects study suggest that generated answers still contain high rates of hallucinations, fail to fittingly address unanswerable questions, and don't suitably reflect contextual information.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Annotator-Centric Active Learning for Subjective NLP Tasks

Michiel van der Meer, Nele Falk, Pradeep K. Murukannaiyah, Enrico Liscio

Active Learning (AL) addresses the high costs of collecting human annotations by strategically annotating the most informative samples. However, for subjective NLP tasks, incorporating a wide range of perspectives in the annotation process is crucial to capture the variability in human judgments. We introduce Annotator-Centric Active Learning (ACAL), which incorporates an annotator selection strategy following data sampling. Our objective is two-fold: (1) to efficiently approximate the full diversity of human judgments, and (2) to assess model performance using annotator-centric metrics, which value minority and majority perspectives equally. We experiment with multiple annotator selection strategies across seven subjective NLP tasks, employing both traditional and novel, human-centered evaluation metrics. Our findings indicate that ACAL improves data efficiency and excels in annotator-centric performance evaluations. However, its success depends on the availability of a sufficiently large and diverse pool of annotators to sample from.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Updating CLIP to Prefer Descriptions Over Captions

Amir Zur, Elisa Kreiss, Karel D' Oostervliet, Christopher Potts, Atticus Geiger

Although CLIPScore is a powerful generic metric that captures the similarity between a text and an image, it fails to distinguish between a caption that is meant to complement the information in an image and a description that is meant to replace an image entirely, e.g., for accessibility. We address this shortcoming by updating the CLIP model with the Concordia dataset to assign higher scores to descriptions than captions using parameter efficient fine-tuning and a loss objective derived from work on causal interpretability. This model correlates with the judgements of blind and low-vision people while preserving transfer capabilities and has interpretable structure that sheds light on the caption–description distinction.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

The Death and Life of Great Prompts: Analyzing the Evolution of LLM Prompts from the Structural Perspective

Yihua Ma, Xinyue Shen, Yixin Wu, Boyang Zhang, Michael Backes, Yang Zhang

Effective utilization of large language models (LLMs), such as ChatGPT, relies on the quality of input prompts. This paper explores prompt engineering, specifically focusing on the disparity between experimentally designed prompts and real-world "in-the-wild" prompts. We analyze 10,538 in-the-wild prompts collected from various platforms and develop a framework that decomposes the prompts into eight key components. Our analysis shows that Role and Requirement are the most prevalent two components. Roles specified in the prompts, along with their capabilities, have become increasingly varied over time, signifying a broader range of application scenarios for LLMs. However, from the response of GPT-4, there is a marginal improvement with a specified role, whereas leveraging less prevalent components such as Capability and Demonstration can result in a more satisfying response. Overall, our work sheds light on the essential components of in-the-wild prompts and the effectiveness of these components on the broader landscape of LLM prompt engineering, providing valuable guidelines for the LLM community to optimize high-quality prompts.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?

Alexander Choi, Syeda Sabrina Akter, J.P. Singh, Antonios Anastasopoulos

Large Language Models (LLMs) have shown capabilities close to human performance in various analytical tasks, leading researchers to use them for time and labor-intensive analyses. However, their capability to handle highly specialized and open-ended tasks in domains like policy studies remains in question. This paper investigates the efficiency and accuracy of LLMs in specialized tasks through a structured user study focusing on Human-LLM partnership. The study, conducted in two stages Topic Discovery and Topic Assignment, integrates LLMs with expert annotators to observe the impact of LLM suggestions on what is usually human-only analysis. Results indicate that LLM-generated topic lists have significant overlap with human generated topic lists, with minor hiccups in missing document-specific topics. However, LLM suggestions may significantly improve task completion speed, but at the same time introduce anchoring bias, potentially affecting the depth and nuance of the analysis, raising a critical question about the trade-off between increased efficiency and the risk of biased analysis.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MUSCLE: A Model Update Strategy for Compatible LLM Evolution

Jessica Maria Echterhoff, Farash Faghri, Raviteja Venulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, Hadi Pouransari

Large Language Models (LLMs) are regularly updated to enhance performance, typically through changes in data or architecture. Within the update process, developers often prioritize improving overall performance metrics, paying less attention to maintaining compatibility with earlier model versions. Instance-level degradation (instance regression) of performance from one model version to the next can interfere with a user's mental model of the capabilities of a particular language model. Users having to adapt their mental model with every update can lead to dissatisfaction, especially when the new model has degraded compared to a prior version for a known use case (model update regression). We find that when pretrained LLM base models are updated, fine-tuned user-facing downstream task adapters experience negative flips – previously correct instances are now predicted incorrectly. We observe model update regression between different model versions on a diverse set of tasks and models, even when the downstream task training procedures remain identical. We argue for the importance of maintaining model update compatibility during updates, and present evaluation metrics designed specifically for generative tasks, while also being applicable to discriminative tasks. We propose a training strategy to minimize the extent of instance regression in model updates, involving training of a compatibility adapter that can enhance task fine-tuned language models. We show negative flips reduce by up to 40% e.g. when updating Llama 1 to Llama 2 with our proposed method.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, Frauke Kreuter

Recent advances in Large Language Models (LLMs) have sparked wide interest in validating and comprehending the human-like cognitive-behavioral traits LLMs may capture and convey. These cognitive-behavioral traits include typically Attitudes, Opinions, Values (AOVs). However, measuring AOVs embedded within LLMs remains opaque, and different evaluation methods may yield different results. This has

led to a lack of clarity on how different studies are related to each other and how they can be interpreted. This paper aims to bridge this gap by providing a comprehensive overview of recent works on the evaluation of AOVs in LLMs. Moreover, we survey related approaches in different stages of the evaluation pipeline in these works. By doing so, we address the potential and challenges with respect to understanding the model, human-AI alignment, and downstream application in social sciences. Finally, we provide practical insights into evaluation methods, model enhancement, and interdisciplinary collaboration, thereby contributing to the evolving landscape of evaluating AOVs in LLMs.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Can LLM be a Personalized Judge?

Yijiang River Dong, Tiancheng Hu, Nigel Collier

As large language models (LLMs) gain widespread adoption, ensuring they cater to diverse user needs has become increasingly important. While many researchers have studied LLM personalization and role-playing, they primarily use LLM-as-a-Judge for evaluation without thoroughly examining its validity. This paper investigates the reliability of LLM-as-a-Personalized-Judge asking LLMs to judge user preferences based on persona. Our results suggest that LLM-as-a-Personalized-Judge is less reliable for personalization than previously believed, showing low agreement with human ground truth. We observed that the personas provided to the LLM often have limited predictive power for the tasks, leading us to introduce verbal uncertainty estimation. We find that powerful LLMs are aware of the certainty of their prediction and can achieve high agreement with ground truth on high-certainty samples, indicating a promising approach for building reliable and scalable proxies for evaluating LLM personalization. Our human annotation reveals that third-person crowd worker evaluations of personalized preferences are even worse than LLM predictions, highlighting the challenges of evaluating LLM personalization.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Will LLMs Sink or Swim? Exploring Decision-Making Under Pressure

Kyusik Kim, Hyeonseok Jeon, Jeongwoo Ryu, Bongwon Suh

Recent advancements in Large Language Models (LLMs) have demonstrated their ability to simulate human-like decision-making, yet the impact of psychological pressures on their decision-making processes remains underexplored. To understand how psychological pressures influence decision-making in LLMs, we tested LLMs on various high-level tasks, using both explicit and implicit pressure prompts. Moreover, we examined LLM responses under different personas to compare with human behavior under pressure. Our findings show that pressures significantly affect LLMs' decision-making, varying across tasks and models. Persona-based analysis suggests some models exhibit human-like sensitivity to pressure, though with some variability. Furthermore, by analyzing both the responses and reasoning patterns, we identified the values LLMs prioritize under specific social pressures. These insights deepen our understanding of LLM behavior and demonstrate the potential for more realistic social simulation experiments.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Implicit Personalization in Language Models: A Systematic Study

Zhijing Jin, Nils Heil, Jiarui Liu, Sheliaad Dhulaiwala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, Mrimaya Sachan

Implicit Personalization (IP) is a phenomenon of language models inferring a user's background from the implicit cues in the input prompts and tailoring the response based on this inference. While previous work has touched upon various instances of this problem, there lacks a unified framework to study this behavior. This work systematically studies IP through a rigorous mathematical formulation, a multi-perspective moral reasoning framework, and a set of case studies. Our theoretical foundation for IP relies on a structural causal model and introduces a novel method, indirect intervention, to estimate the causal effect of a mediator variable that cannot be directly intervened upon. Beyond the technical approach, we also introduce a set of moral reasoning principles based on three schools of moral philosophy to study when IP may or may not be ethically appropriate. Equipped with both mathematical and ethical insights, we present three diverse case studies illustrating the varied nature of the IP problem and offer recommendations for future research.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Cognitive Bias in Decision-Making with LLMs

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, Zexue He

Large language models (LLMs) offer significant potential as tools to support an expanding range of decision-making tasks. Given their training on human (created) data, LLMs have been shown to inherit societal biases against protected groups, as well as be subject to bias functionally resembling cognitive bias. Human-like bias can impede fair and explainable decisions made with LLM assistance. Our work introduces BiasBuster, a framework designed to uncover, evaluate, and mitigate cognitive bias in LLMs, particularly in high-stakes decision-making tasks. Inspired by prior research in psychology and cognitive science, we develop a dataset containing 13,463 prompts to evaluate LLM decisions on different cognitive biases (e.g., prompt-induced, sequential, inherent). We test various bias mitigation strategies, while proposing a novel method utilizing LLMs to debias their own human-like cognitive bias within prompts. Our analysis provides a comprehensive picture of the presence and effects of cognitive bias across commercial and open-source models. We demonstrate that our selfhelp debiasing effectively mitigates model answers that display patterns akin to human cognitive bias without having to manually craft examples for each bias.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

"Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations?

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, Barbara Plank

Human label variation (HLV) is a valuable source of information that arises when multiple human annotators provide different labels for valid reasons. In Natural Language Inference (NLI) earlier approaches to capturing HLV involve either collecting annotations from many crowd workers to represent human judgment distribution (HJD) or use expert linguists to provide detailed explanations for their chosen labels. While the former method provides denser HJD information, obtaining it is resource-intensive. In contrast, the latter offers richer textual information but it is challenging to scale up to many human judges. Besides, large language models (LLMs) are increasingly used as evaluators ("LLM judges") but with mixed results, and few works aim to study HJDs. This study proposes to exploit LLMs to approximate HJDs using a small number of expert labels and explanations. Our experiments show that a few explanations significantly improve LLMs' ability to approximate HJDs with and without explicit labels, thereby providing a solution to scale up annotations for HJD. However, fine-tuning smaller soft-label aware models with the LLM-generated model judgment distributions (MJDs) presents partially inconsistent results: while similar in distance, their resulting fine-tuned models and visualized distributions differ substantially. We show the importance of complementing instance-level distance measures with a global-level shape metric and visualization to more effectively evaluate MJDs against human judgment distributions.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Minimal Yet Big Impact: How AI Agent Back-channeling Enhances Conversational Engagement through Conversation Persistence and Context Richness

Jin Yea Jang, Saim Shin, gahgane gweon

The increasing use of AI agents in conversational services, such as counseling, highlights the importance of back-channeling (BC) as an active listening strategy to enhance conversational engagement. BC improves conversational engagement by providing timely acknowledgments and encouraging the speaker to talk. This study investigates the effect of BC provided by an AI agent on conversational engagement, offering insights for future AI conversational service design. We conducted an experiment with 55 participants, divided into Todak_BC and

Todak_NoBC groups based on the presence or absence of the BC feature in Todak, a conversational agent. Each participant engaged in nine sessions with predetermined subjects and questions. We collected and analyzed approximately 6 hours and 30 minutes of conversation logs to evaluate conversational engagement using both quantitative (conversation persistence, including conversation duration and number of utterances) and qualitative metrics (context richness, including self-disclosure and topic diversity). The findings reveal significantly higher conversational engagement in the Todak_BC group compared to the Todak_NoBC group across all metrics ($p < 0.05$). Additionally, the impact of BC varies across sessions, suggesting that conversation characteristics such as question type and topic sensitivity can influence BC effectiveness.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, Yun-Nung Chen

The concept of "persona", originally adopted in dialogue literature, has re-surfaced as a promising framework for tailoring large language models (LLMs) to specific context (*e.g.* personalized search, LLM-as-a-judge). However, the growing research on leveraging persona in LLMs is relatively disorganized and lacks a systematic taxonomy. To close the gap, we present a comprehensive survey to categorize the current state of the field. We identify two lines of research, namely (1) *LLM Role-Playing*, where personas are assigned to LLMs, and (2) *LLM Personalization*, where LLMs take care of user personas. Additionally, we introduce existing methods for LLM personality evaluation. To the best of our knowledge, we present the first survey for role-playing and personalization in LLMs under the unified view of persona. We continuously maintain a paper collection to foster future endeavors.

Interpretability and Analysis of Models for NLP 3

Nov 13 (Wed) 10:30-12:00 - Room: Jasmine

Nov 13 (Wed) 10:30-12:00 - Jasmine

Measuring the Robustness of NLP Models to Domain Shifts

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chaparin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, Roi Reichart

Existing research on Domain Robustness (DR) suffers from disparate setups, limited task variety, and scarce research on recent capabilities such as in-context learning. Furthermore, the common practice of measuring DR might not be fully accurate. Current research focuses on challenge set and relies solely on the Source Drop (SD): Using the source in-domain performance as a reference point for degradation. However, we argue that the Target Drop (TD), which measures degradation from the target in-domain performance, should be used as a complementary point of view. To address these issues, we first curated a DR benchmark comprised of 7 diverse NLP tasks, which enabled us to measure both the SD and the TD. We then conducted a comprehensive large-scale DR study involving over 14,000 domain shifts across 21 fine-tuned models and few-shot LLMs. We found that both model types suffer from drops upon domain shifts. While fine-tuned models excel in-domain, few-shot LLMs often surpass them cross-domain, showing better robustness. In addition, we found that a large SD can often be explained by shifting to a harder domain rather than by a genuine DR challenge, and this highlights the importance of TD as a complementary metric. We hope our study will shed light on the current DR state of NLP models and promote improved evaluation practices toward more robust models.

Nov 13 (Wed) 10:30-12:00 - Jasmine

In Search of the Long-Tail: Systematic Generation of Long-Tail Inferential Knowledge via Logical Rule Guided Search

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenling Zhao, Faeze Brahman, Yejin Choi, Xiang Ren

To effectively use large language models (LLMs) for real-world queries, it is imperative that they generalize to the long-tail distribution, i.e. rare examples where models exhibit low confidence. In this work, we take the first step towards evaluating LLMs in the long-tail distribution of inferential knowledge. We exemplify long-tail evaluation on the Natural Language Inference task. First, we introduce Logic-Induced-Knowledge-Search (LINK), a systematic long-tail data generation framework, to obtain factually-correct yet long-tail inferential statements. LINK uses variable-wise prompting grounded on symbolic rules to seek low-confidence statements while ensuring factual correctness. We then use LINK to curate Logic-Induced-Long-Tail (LINT), a large-scale long-tail inferential knowledge dataset that contains 108K statements spanning four domains. We evaluate popular LLMs on LINT; we find that state-of-the-art LLMs show significant performance drop (21% relative drop for GPT4) on long-tail data as compared to on head distribution data, and smaller models show even more generalization weakness. These results further underscore the necessity of long-tail evaluation in developing generalizable LLMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Neuron-Level Knowledge Attribution in Large Language Models

ZEPING YU, Sophia Ananiadou

Identifying important neurons for final predictions is essential for understanding the mechanisms of large language models. Due to computational constraints, current attribution techniques struggle to operate at neuron level. In this paper, we propose a static method for pinpointing significant neurons. Compared to seven other methods, our approach demonstrates superior performance across three metrics. Additionally, since most static methods typically only identify "value neurons" directly contributing to the final prediction, we propose a method for identifying "query neurons" which activate these "value neurons". Finally, we apply our methods to analyze six types of knowledge across both attention and feed-forward network (FFN) layers. Our method and analysis are helpful for understanding the mechanisms of knowledge storage and set the stage for future research in knowledge editing. The code is available on <https://github.com/zepingyu0512/neuron-attribution>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

How do Large Language Models Learn In-Context? Query and Key Matrices of In-Context Heads are Two Towers for Metric Learning

ZEPING YU, Sophia Ananiadou

We investigate the mechanism of in-context learning (ICL) on sentence classification tasks with semantically-unrelated labels ("foo"/"bar"). We find intervening in only 1% heads (named "in-context heads") significantly affects ICL accuracy from 87.6% to 24.4%. To understand this phenomenon, we analyze the value-output vectors in these heads and discover that the vectors at each label position contain substantial information about the corresponding labels. Furthermore, we observe that the prediction shift from "foo" to "bar" is due to the respective reduction and increase in these heads' attention scores at "foo" and "bar" positions. Therefore, we propose a hypothesis for ICL: in-in-context heads, the value-output matrices extract label features, while the query-key matrices compute the similarity between the features at the last position and those at each label position. The query and key matrices can be considered as two towers that learn the similarity metric between the last position's features and each demonstration at label positions. Using this hypothesis, we explain the majority label bias and recency bias in ICL and propose two methods to reduce these biases by 22% and 17%, respectively.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Contextual and Parametric Knowledge: More Context, More Focus

Yufei Tao, Adam Hiatt, Erik Haake, Antonie J. Jetter, Ameeta Agrawal

Large language models (LLMs) have demonstrated remarkable progress in leveraging diverse knowledge sources. This study investigates how nine widely used LLMs allocate knowledge between local context and global parameters when answering open-ended questions in knowledge-consistent scenarios. We introduce a novel dataset, WikiAtomic, and systematically vary context sizes to analyze how LLMs prioritize and utilize the provided information and their parametric knowledge in knowledge-consistent scenarios. Additionally, we also study their tendency to hallucinate under varying context sizes. Our findings reveal consistent patterns across models, including a consistent reliance on both contextual (around 70%) and parametric (around 30%) knowledge, and a decrease in hallucinations with increasing context. These insights highlight the importance of more effective context organization and developing models that use input more deterministically for robust performance.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Investigating How Large Language Models Leverage Internal Knowledge to Perform Complex Reasoning

Miyoung Ko, Sue Hyun Park, Joonsuk Park, Minjoon Seo

Despite the advances in large language models (LLMs), how they use their knowledge for reasoning is not yet well understood. In this study, we propose a method that deconstructs complex real-world questions into a graph, representing each question as a node with predecessors of background knowledge needed to solve the question. We develop the DepthQA dataset, deconstructing questions into three depths: (i) recalling conceptual knowledge, (ii) applying procedural knowledge, and (iii) analyzing strategic knowledge. Based on a hierarchical graph, we quantify forward discrepancy, a discrepancy in LLM performance on simpler sub-problems versus complex questions. We also measure backward discrepancy where LLMs answer complex questions but struggle with simpler ones. Our analysis shows that smaller models exhibit more discrepancies than larger models. Distinct patterns of discrepancies are observed across model capacity and possibility of training data memorization. Additionally, guiding models from simpler to complex questions through multi-turn interactions improves performance across model sizes, highlighting the importance of structured intermediate steps in knowledge reasoning. This work enhances our understanding of LLM reasoning and suggests ways to improve their problem-solving abilities.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Secured Weight Release for Large Language Models via Taylor Expansion

Guanchu Wang, Yu-Neng Chuang, Ruixiang Tang, Shaochen Zhong, Jiayi Yuan, Hongye Jin, Zirui Liu, Vipin Chaudhary, Shuai Xu, James Caverlee, Xia Hu

Ensuring the security of released large language models (LLMs) poses a significant dilemma, as existing mechanisms either compromise ownership rights or raise data privacy concerns. To address this dilemma, we introduce TaylorMLP to protect the ownership of released LLMs and prevent their abuse. Specifically, TaylorMLP preserves the ownership of LLMs by transforming the weights of LLMs into parameters of Taylor-series. Instead of releasing the original weights, developers can release the Taylor-series parameters with users, thereby ensuring the security of LLMs. Moreover, TaylorMLP can prevent abuse of LLMs by adjusting the generation speed. It can induce low-speed token generation for the protected LLMs by increasing the terms in the Taylor-series. This intentional delay helps LLM developers prevent potential large-scale unauthorized uses of their models. Empirical experiments across five datasets and three LLM architectures demonstrate that TaylorMLP induces over increase in latency, producing the tokens precisely matched with original LLMs. Subsequent defensive experiments further confirm that TaylorMLP effectively prevents users from reconstructing the weight values based on downstream datasets.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models

Tianyi Men, Pengfei Cao, Zhiuruan Jin, Yubo Chen, Kang Liu, Jun Zhao

Planning, as the core module of agents, is crucial in various fields such as embodied agents, web navigation, and tool using. With the development of large language models (LLMs), some researchers treat large language models as intelligent agents to stimulate and evaluate their planning capabilities. However, the planning mechanism is still unclear. In this work, we focus on exploring the look-ahead planning mechanism in large language models from the perspectives of information flow and internal representations. First, we study how planning is done internally by analyzing the multi-layer perception (MLP) and multi-head self-attention (MHSA) components at the last token. We find that the output of MHSA in the middle layers at the last token can directly decode the decision to some extent. Based on this discovery, we further trace the source of MHSA by information flow, and we reveal that MHSA extracts information from spans of the goal states and recent steps. According to information flow, we continue to study what information is encoded within it. Specifically, we explore whether future decisions have been considered in advance in the representation of flow. We demonstrate that the middle and upper layers encode a few short-term future decisions. Overall, our research analyzes the look-ahead planning mechanisms of LLMs, facilitating future research on LLMs performing planning tasks.

Nov 13 (Wed) 10:30-12:00 - Jasmine

The Best Defense is Attack: Repairing Semantics in Textual Adversarial Examples

Heng Yang

Recent studies have revealed the vulnerability of pre-trained language models to adversarial attacks. Adversarial defense techniques have been proposed to reconstruct adversarial examples within feature or text spaces. However, these methods struggle to effectively repair the semantics in adversarial examples, resulting in unsatisfactory defense performance. To repair the semantics in adversarial examples, we introduce a novel approach named Reactive Perturbation Defocusing (Rapid), which employs an adversarial detector to identify the fake labels of adversarial examples and leverages adversarial attackers to repair the semantics in adversarial examples. Our extensive experimental results, conducted on four public datasets, demonstrate the consistent effectiveness of Rapid in various adversarial attack scenarios. For easy evaluation, we provide a click-to-run demo of Rapid at <https://tinyurl.com/22ercuf8>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Perceptions of Linguistic Uncertainty by Language Models and Humans

Catarina G Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, Padhraic Smyth

Uncertainty expressions^a such as ‘probably’ or ‘highly unlikely’ are pervasive in human language. While prior work has established that there is population-level agreement in terms of how humans quantitatively interpret these expressions, there has been little inquiry into the abilities of language models in the same context. In this paper, we investigate how language models map linguistic expressions of uncertainty to numerical responses. Our approach assesses whether language models can employ theory of mind in this setting: understanding the uncertainty of another agent about a particular statement, independently of the model’s own certainty about that statement. We find that 7 out of 10 models are able to map uncertainty expressions to probabilistic responses in a human-like manner. However, we observe systematically different behavior depending on whether a statement is actually true or false. This sensitivity indicates that language models are substantially more susceptible to bias based on their prior knowledge (as compared to humans). These findings raise important questions and have broad implications for human-AI and AI-AI communication.

Nov 13 (Wed) 10:30-12:00 - Jasmine

The effects of distance on NPI illusive effects in BERT

So Young Lee, Mai Ha Vu

Previous studies have examined the syntactic capabilities of large pre-trained language models, such as BERT, by using stimuli from psycholinguistic studies. Studying well-known processing errors, such as NPI illusive effects can reveal whether a model prioritizes linear or hierarchical information when processing language. Recent experiments have found that BERT is mildly susceptible to Negative Polarity Item (NPI) illusion effects (Shin et al., 2023; Vu and Lee, 2022). We expand on these results by examining the effect of distance on the illusive effect, using and modifying stimuli from Parker and Phillips (2016). We also further tease apart whether the model is more affected by hierarchical distance or linear distance. We find that BERT is highly sensitive to syntactic hierarchical information: added hierarchical layers affected its processing capabilities compared to added linear distance.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Unlocking Memorization in Large Language Models with Dynamic Soft Prompting

Zhepeng Wang, Runxue Bao, Yawen Wu, Jackson Taylor, Cao Xiao, Feng Zheng, Weiwen Jiang, Shangqian Gao, Yanfu Zhang

Pretrained large language models (LLMs) have excelled in a variety of natural language processing (NLP) tasks, including summarization, question answering, and translation. However, LLMs pose significant security risks due to their tendency to memorize training data, leading to potential privacy breaches and copyright infringement. Therefore, accurate measurement of the memorization is essential to evaluate and mitigate these potential risks. However, previous attempts to characterize memorization are constrained by either using prefixes only or by prepending a constant soft prompt to the prefixes, which cannot react to changes in input. To address this challenge, we propose a novel method for estimating LLM memorization using dynamic, prefix-dependent soft prompts. Our approach involves training a transformer-based generator to produce soft prompts that adapt to changes in input, thereby enabling more accurate extraction of memorized data. Our method not only addresses the limitations of previous methods but also demonstrates superior performance in diverse experimental settings compared to state-of-the-art techniques. In particular, our method can achieve the maximum relative improvement of 135.3% and 39.8% over the vanilla baseline on average in terms of *discoverable memorization rate* for the text generation task and code generation task, respectively. Our code is available at <https://github.com/wangger/lm-memorization-dsp>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

LLMs Are Prone to Fallacies in Causal Inference

Nitish Joshi, Abulhair Saparov, Yixin Wang, He He

Recent work shows that causal facts can be effectively extracted from LLMs through prompting, facilitating the creation of causal graphs for causal inference tasks. However, it is unclear if this success is limited to explicitly-mentioned causal facts in the pretraining data which the model can memorize. Thus, this work investigates: Can LLMs infer causal relations from other relational data in text? To disentangle the role of memorized causal facts vs inferred causal relations, we finetune LLMs on synthetic data containing temporal, spatial and counterfactual relations, and measure whether the LLM can then infer causal relations. We find that: (a) LLMs are susceptible to inferring causal relations from the order of two entity mentions in text (e.g. X mentioned before Y implies X causes Y); (b) if the order is randomized, LLMs still suffer from the post hoc fallacy, i.e. X occurs before Y (temporal relation) implies X causes Y. We also find that while LLMs can correctly deduce the absence of causal relations from temporal and spatial relations, they have difficulty inferring causal relations from counterfactuals, questioning their understanding of causality.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Paraphrase Types Elicit Prompt Engineering Capabilities

Jan Philipp Wahle, Terry Ruas, Yang Xu, Bela Gipp

Much of the success of modern language models depends on finding a suitable prompt to instruct the model. Until now, it has been largely unknown how variations in the linguistic expression of prompts affect these models. This study systematically and empirically evaluates which linguistic features influence models through paraphrase types, i.e., different linguistic changes at particular positions. We measure behavioral changes for five models across 120 tasks and six families of paraphrases (i.e., morphology, syntax, lexicor, lexico-syntax, discourse, and others). We also control for other prompt engineering factors (e.g., prompt length, lexical diversity, and proximity to training data). Our results show a potential for language models to improve tasks when their prompts are adapted in specific paraphrase types (e.g., 6.7% median gain in Mixtral 8x7B; 5.5% in LLaMA 3.8B). In particular, changes in morphology and lexicor, i.e., the vocabulary used, showed promise in improving prompts. These findings contribute to developing more robust language models capable of handling variability in linguistic expression.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Towards Interpretable Sequence Continuation: Analyzing Shared Circuits in Large Language Models

Michael Lan, Philip Torr, Fazl Barez

While transformer models exhibit strong capabilities on linguistic tasks, their complex architectures make them difficult to interpret. Recent work has aimed to reverse engineer transformer models into human-readable representations called circuits that implement algorithmic functions. We extend this research by analyzing and comparing circuits for similar sequence continuation tasks, which include increasing sequences of Arabic numerals, number words, and months. By applying circuit interpretability analysis, we identify a key sub-circuit in both GPT-2 Small and Llama-2-7B responsible for detecting sequence members and for predicting the next member in a sequence. Our analysis reveals that semantically related sequences rely on shared circuit subgraphs with analogous roles. Additionally, we show that this sub-circuit has effects on various math-related prompts, such as on intervalized circuits, Spanish number word and months continuation, and natural language word problems. Overall, documenting shared computational structures enables better model behavior predictions, identification of errors, and safer editing procedures. This mechanistic understanding of transformers is a critical step towards building more robust, aligned, and interpretable language models.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Why Does New Knowledge Create Messy Ripple Effects in LLMs?

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, Heng Ji

Extensive previous research has focused on post-training knowledge editing (KE) for language models (LMs) to ensure that knowledge remains accurate and up-to-date. One desired property and open question in KE is to let edited LMs correctly handle ripple effects, where LM is expected to answer its logically related knowledge accurately. In this paper, we answer the question of why most KE methods still create messy ripple effects. We conduct extensive analysis and identify a salient indicator, GradSim, that effectively reveals when and why updated knowledge ripples in LMs. GradSim is computed by the cosine similarity between gradients of the original fact and its related knowledge. We observe a strong positive correlation between ripple effect performance and GradSim across different LMs, KE methods, and evaluation metrics. Further investigations into three counter-intuitive failure cases (Negation, Over-Ripple, Multi-Lingual) of ripple effects demonstrate that these failures are often associated with very low GradSim. This finding validates that GradSim is an effective indicator of when knowledge ripples in LMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, Daniel Khashabi
 Humans regularly engage in analogical thinking, relating personal experiences to current situations (X is analogous to Y because of Z). Analogical thinking allows humans to solve problems in creative ways, grasp difficult concepts, and articulate ideas more effectively. Can language models (LMs) do the same? To answer this question, we propose AnaloBench, a benchmark to determine analogical reasoning ability in LMs. Our benchmarking approach focuses on aspects of this ability that are common among humans: (i) recalling related experiences from a large amount of information, and (ii) applying analogical reasoning to complex and lengthy scenarios. We collect a set of 340 high quality, human written analogies for use in our benchmark, which constitutes the largest such collection to date. We then test a broad collection of models consisting of 12 open source and 3 proprietary in various sizes and architectures. As in prior results, scaling up LMs results in some performance boosts. Surprisingly, scale offers minimal gains when, (i) analogies involve lengthy scenarios, or (ii) recalling relevant scenarios from a large pool of information, a process analogous to finding a needle in a haystack. We hope these observations encourage further research in this field.

Nov 13 (Wed) 10:30-12:00 - Jasmine

An Analysis and Mitigation of the Reversal Curse

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhua Chen, Ji-Rong Wen, Rui Yan

Recent research observed a noteworthy phenomenon in large language models (LLMs), referred to as the "reversal curse." The reversal curse is that when dealing with two entities, denoted as a and b , connected by their relation R and its inverse R^{-1} , LLMs excel in handling sequences in the form of " aRb ," but encounter challenges when processing " $bR^{-1}a$," whether in generation or comprehension. For instance, GPT-4 can accurately respond to the query "Tom Cruise's mother is?" with "Mary Lee Pfeiffer," but it struggles to provide a satisfactory answer when asked "Mary Lee Pfeiffer's son is?" In this paper, we undertake the first-ever study of how the reversal curse happens in LLMs. Our investigations reveal that the reversal curse can stem from the specific training objectives, which become particularly evident in the widespread use of next-token prediction within most causal language models. We hope this initial investigation can draw more attention to the reversal curse, as well as other underlying limitations in current LLMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, Amir Globerson

Large language models (LLMs) can solve complex multi-step problems, but little is known about how these computations are implemented internally. Motivated by this, we study how LLMs answer multi-hop queries such as "The spouse of the performer of Imagine is?". These queries require two information extraction steps: a latent one for resolving the first hop ("the performer of Imagine") into the bridge entity (John Lennon), and another for resolving the second hop ("the spouse of John Lennon") into the target entity (Yoko Ono). Understanding how the latent step is computed internally is key to understanding the overall computation. By carefully analyzing the internal computations of transformer-based LLMs, we discover that the bridge entity is resolved in the early layers of the model. Then, only after this resolution, the two-hop query is solved in the later layers. Because the second hop commences in later layers, there could be cases where these layers no longer encode the necessary knowledge for correctly predicting the answer. Motivated by this, we propose a novel "back-patching" analysis method whereby a hidden representation from a later layer is patched back to an earlier layer. We find that in up to 66% of previously incorrect cases there exists a back-patch that results in the correct generation of the answer, showing that the later layers indeed sometimes lack the needed functionality. Overall our methods and findings open further opportunities for understanding and improving latent reasoning in transformer-based LLMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

HalluMeasure: Fine-grained Hallucination Measurement Using Chain-of-Thought Reasoning

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Victor Alvarez, Erica M Salinas, Erwin Cornejo

Automating the measurement of hallucinations in LLM generated responses is a challenging task as it requires careful investigation of each factual claim in a response. In this paper, we introduce HalluMeasure, a new LLM-based hallucination detection mechanism that decomposes an LLM response into atomic claims, and evaluates each atomic claim against the provided reference context. The model uses a step-by-step reasoning process called Chain-of-Thought and can identify 3 major categories of hallucinations (e.g., contradiction) as well as 10 more specific subtypes (e.g., overgeneralization) which help to identify reasons behind the hallucination errors. Specifically, we explore four different configurations for HalluMeasure's classifier: with and without CoT prompting, and using a single classifier call to classify all claims versus separate calls for each claim. The best-performing configuration (with CoT and separate calls for each claim) demonstrates significant improvements in detecting hallucinations, achieving a 10-point increase in F1 score on our TechNewsSumm dataset, and a 3-point increase in AUC ROC on the SummEval dataset, compared to three baseline models (RefChecker, AlignScore, and Vectara HHEM). We further show reasonable accuracy on detecting 10 novel error subtypes of hallucinations (where even humans struggle in classification) derived from linguistic analysis of the errors made by the LLMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Exploring the Learning Capabilities of Language Models using LEVERWORLDS

Eitan Wagner, Amir Feder, Omri Abend

Learning a model of a stochastic setting often involves learning both general structure rules and specific properties of the instance. This paper investigates the interplay between learning the general and the specific in various learning methods, with emphasis on sample efficiency. We design a framework called LEVERWORLDS, which allows the generation of simple physics-inspired worlds that follow a similar generative process with different distributions, and their instances can be expressed in natural language. These worlds allow for controlled experiments to assess the sample complexity of different learning methods. We experiment with classic learning algorithms as well as Transformer language models, both with fine-tuning and In-Context Learning (ICL). Our general finding is that (1) Transformers generally succeed in the task; but (2) they are considerably less sample efficient than classic methods that make stronger assumptions about the structure, such as Maximum Likelihood Estimation and Logistic Regression. This finding is in tension with the recent tendency to use Transformers as general-purpose estimators. We propose an approach that leverages the ICL capabilities of contemporary language models to apply simple algorithms for this type of data. Our experiments show that models currently struggle with the task but show promising potential.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Rationale-Aware Answer Verification by Pairwise Self-Evaluation

Akira Kawabata, Saku Sugawara

Answer verification identifies correct solutions among candidates generated by large language models (LLMs). Current approaches typically train verifier models by labeling solutions as correct or incorrect based solely on whether the final answer matches the gold answer. However, this approach neglects any flawed rationale in the solution yielding the correct answer, undermining the verifier's ability to distinguish between sound and flawed rationales. We empirically show that in StrategyQA, only 19% of LLM-generated solutions with correct answers have valid rationales, thus leading to an unreliable verifier. Furthermore, we demonstrate that training a verifier on valid rationales signifi-

cantly improves its ability to distinguish valid and flawed rationale. To make a better verifier without extra human supervision, we introduce REPS (Rationale Enhancement through Pairwise Selection), a method for selecting valid rationales from candidates by iteratively applying pairwise self-evaluation using the same LLM that generates the solutions. Verifiers trained on solutions selected by REPS outperform those trained using conventional training methods on three reasoning benchmarks (ARC-Challenge, DROP, and StrategyQA). Our results suggest that training reliable verifiers requires ensuring the validity of rationales in addition to the correctness of the final answers, which would be critical for models assisting humans in solving complex reasoning tasks.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Interpreting Context Look-ups in Transformers: Investigating Attention-MLP Interactions

Clement Neo, Shay B Cohen, Faiz Barez

Understanding the inner workings of large language models (LLMs) is crucial for advancing their theoretical foundations and real-world applications. While the attention mechanism and multi-layer perceptrons (MLPs) have been studied independently, their interactions remain largely unexplored. This study investigates how attention heads and next-token neurons interact in LLMs to predict new words. We propose a methodology to identify next-token neurons, find prompts that highly activate them, and determine the upstream attention heads responsible. We then generate and evaluate explanations for the activity of these attention heads in an automated manner. Our findings reveal that some attention heads recognize specific contexts relevant to predicting a token and activate a downstream token-predicting neuron accordingly. This mechanism provides a deeper understanding of how attention heads work with MLP neurons to perform next-token prediction. Our approach offers a foundation for further research into the intricate workings of LLMs and their impact on text generation and understanding.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Fill In The Gaps: Model Calibration and Generalization with Synthetic Data

Yang Ba, Michelle V Mancenido, Rong Pan

As machine learning models continue to swiftly advance, calibrating their performance has become a major concern prior to practical and widespread implementation. Most existing calibration methods often negatively impact model accuracy due to the lack of diversity of validation data, resulting in reduced generalizability. To address this, we propose a calibration method that incorporates synthetic data without compromising accuracy. We derive the expected calibration error (ECE) bound using the Probably Approximately Correct (PAC) learning framework. Large language models (LLMs), known for their ability to mimic real data and generate text with mixed class labels, are utilized as a synthetic data generation strategy to lower the ECE bound and improve model accuracy on real test data. Additionally, we propose data generation mechanisms for efficient calibration. Testing our method on four different natural language processing tasks, we observed an average up to 34% increase in accuracy and 33% decrease in ECE.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Low-rank Subspace for Binding in Large Language Models

Qin Dai, Benjamin Heinzerling, Kenaro Inui

Entity tracking is essential for complex reasoning. To perform in-context entity tracking, language models (LMs) must bind an entity to its attribute (e.g., bind a container to its content) to recall attribute for a given entity. For example, given a context mentioning "The coffee is in Box Z, the stone is in Box M, the map is in Box H", to infer "Box Z contains the coffee" later, LMs must bind "Box Z" to "coffee". To explain the binding behaviour of LMs, existing research introduces a Binding ID mechanism and states that LMs use a abstract concept called Binding ID (BI) to internally mark entity-attribute pairs. However, they have not directly captured the BI information from entity activations. In this work, we provide a novel view of the Binding ID mechanism by localizing the BI information. Specifically, we discover that there exists a low-rank subspace in the hidden state (or activation) of LMs, that primarily encodes BIs. To identify this subspace, we take principle component analysis as our first attempt and it is empirically proven to be effective. Moreover, we also discover that when editing representations along directions in the subspace, LMs tend to bind a given entity to other attributes accordingly. For example, by patching activations along the BI encoding direction we can make the LM to infer "Box Z contains the stone" and "Box Z contains the map".

Nov 13 (Wed) 10:30-12:00 - Jasmine

Enhancing Post-Hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition

Pritika Ramu, Koustava Goswami, Apoorv Saxena, Balaji Vasavada Srinivasan

Accurately attributing answer text to its source document is crucial for developing a reliable question-answering system. However, attribution for long documents remains largely unexplored. Post-hoc attribution systems are designed to map answer text back to the source document, yet the granularity of this mapping has not been addressed. Furthermore, a critical question arises: What exactly should be attributed? This involves identifying the specific information units within an answer that require grounding. In this paper, we propose and investigate a novel approach to the factual decomposition of generated answers for attribution, employing template-based in-context learning. To accomplish this, we utilize the question and integrate negative sampling during few-shot in-context learning for decomposition. This approach enhances the semantic understanding of both abstractive and extractive answers. We examine the impact of answer decomposition by providing a thorough examination of various attribution approaches, ranging from retrieval-based techniques to LLM-based attributors.

Nov 13 (Wed) 10:30-12:00 - Jasmine

On the Universal Truthfulness Hyperplane Inside LLMs

Junteng Liu, Shiqi Chen, Yu Cheng, Junxian He

While large language models (LLMs) have demonstrated remarkable abilities across various fields, hallucination remains a significant challenge. Recent studies have explored hallucinations through the lens of internal representations, proposing mechanisms to decipher LLMs' adherence to facts. However, these approaches often fail to generalize to out-of-distribution data, leading to concerns about whether internal representation patterns reflect fundamental factual awareness, or only overfit spurious correlations on the specific datasets. In this work, we investigate whether a universal truthfulness hyperplane that distinguishes the model's factually correct and incorrect outputs exists within the model. To this end, we scale up the number of training datasets and conduct an extensive evaluation – we train the truthfulness hyperplane on a diverse collection of over 40 datasets and examine its cross-task, cross-domain, and in-domain generalization. Our results indicate that increasing the diversity of the training datasets significantly enhances the performance in all scenarios, while the volume of data samples plays a less critical role. This finding supports the optimistic hypothesis that a universal truthfulness hyperplane may indeed exist within the model, offering promising directions for future research.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Please note that I'm just an AI: Analysis of Behavior Patterns of LLMs in (Non-)offensive Speech Identification

Ezra Dönmez, Thang Yu, Agnieszka Fälenska

Offensive speech is highly prevalent on online platforms. Being trained on online data, Large Language Models (LLMs) display undesirable behaviors, such as generating harmful text or failing to recognize it. Despite these shortcomings, the models are becoming a part of our everyday lives by being used as tools for information search, content creation, writing assistance, and many more. Furthermore, the research explores using LLMs in applications with immense social risk, such as late-life companions and online content moderators. Despite the potential harms from LLMs in such applications, whether LLMs can reliably identify offensive speech and how they behave when they

fail are open questions. This work addresses these questions by probing sixteen widely used LLMs and showing that most fail to identify (non-)offensive online language. Our experiments reveal undesirable behavior patterns in the context of offensive speech detection, such as erroneous response generation, over-reliance on profanity, and failure to recognize stereotypes. Our work highlights the need for extensive documentation of model reliability, particularly in terms of the ability to detect offensive language.

Nov 13 (Wed) 10:30-12:00 - Jasmine

A Morphology-Based Investigation of Positional Encodings

Poulami Ghosh, Shikhar Vashishth, Raj Dabre, Pushpak Bhattacharyya

Contemporary deep learning models effectively handle languages with diverse morphology despite not being directly integrated into them. Morphology and word order are closely linked, with the latter incorporated into transformer-based models through positional encodings. This prompts a fundamental inquiry: Is there a correlation between the morphological complexity of a language and the utilization of positional encoding in pre-trained language models? In pursuit of an answer, we present the first study addressing this question, encompassing 22 languages and 5 downstream tasks. Our findings reveal that the importance of positional encoding diminishes with increasing morphological complexity in languages. Our study motivates the need for a deeper understanding of positional encoding, augmenting them to better reflect the different languages under consideration.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Evaluating Concurrent Robustness of Language Models Across Diverse Challenge Sets

Vatsal Gupia, Pranshu Pandya, Tushar Kataria, Vivek Gupta, Dan Roth

Language models, characterized by their black-box nature, often hallucinate and display sensitivity to input perturbations, causing concerns about trust. To enhance trust, it is imperative to gain a comprehensive understanding of the model's failure modes and develop effective strategies to improve their performance. In this study, we introduce a methodology designed to examine how input perturbations affect language models across various scales, including pre-trained models and large language models (LLMs). Utilizing fine-tuning, we enhance the model's robustness to input perturbations. Additionally, we investigate whether exposure to one perturbation enhances or diminishes the model's performance with respect to other perturbations. To address robustness against multiple perturbations, we present three distinct fine-tuning strategies. Furthermore, we broaden the scope of our methodology to encompass large language models (LLMs) by leveraging a chain of thought (CoT) prompting approach augmented with exemplars. We employ the Tabular-NLI task to showcase how our proposed strategies adeptly train a robust model, enabling it to address diverse perturbations while maintaining accuracy on the original dataset.

Nov 13 (Wed) 10:30-12:00 - Jasmine

On Evaluating Explanation Utility for Human-AI Decision Making in NLP

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbind bambroo, Ana Marasovic

Is explainability a false promise? This debate has emerged from the insufficient evidence that explanations help people in situations they are introduced for. More human-centered, application-grounded evaluations of explanations are needed to settle this. Yet, with no established guidelines for such studies in NLP, researchers accustomed to standardized proxy evaluations must discover appropriate measurements, tasks, datasets, and sensible models for human-AI teams in their studies. To aid with this, we first review existing metrics suitable for application-grounded evaluation. We then establish criteria to select appropriate datasets, and using them, we find that only 4 out of over 50 datasets available for explainability research in NLP meet them. We then demonstrate the importance of reassessing the state of the art to form and study human-AI teams: teaming people with models for certain tasks might only now start to make sense, and for others, it remains unsound. Finally, we present the exemplar studies of human-AI decision-making for one of the identified tasks, verifying the correctness of a legal claim given a contract. Our results show that providing AI predictions, with or without explanations, does not cause decision makers to speed up their work without compromising performance. We argue for revisiting the setup of human-AI teams and improving automatic deferral of instances to AI, where explanations could play a useful role.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Variational Language Concepts for Interpreting Pretrained Language Models

Hengyi Wang, Zhiqiang Hong, Shiwei Tan, Desheng Zhang, Hao Wang

Foundation Language Models (FLMs) such as BERT and its variants have achieved remarkable success in natural language processing. To date, the interpretability of FLMs has primarily relied on the attention weights in their self-attention layers. However, these attention weights only provide word-level interpretations, failing to capture higher-level structures, and are therefore lacking in readability and intuitiveness. To address this challenge, we first provide a formal definition of *conceptual interpretation* and then propose a variational Bayesian framework, dubbed VAriational Language Concept (VALC), to go beyond word-level interpretations and provide concept-level interpretations. Our theoretical analysis shows that our VALC finds the optimal language concepts to interpret FLM predictions. Empirical results on several real-world datasets show that our method can successfully provide conceptual interpretation for FLMs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, Elena Tutubalina

The recent integration of chemistry with natural language processing (NLP) has advanced drug discovery. Molecule representation in language models (LMs) is crucial in enhancing chemical understanding. We propose Augmented Molecular Retrieval (AMORE), a flexible zero-shot framework for assessment of Chemistry LMs of different natures: trained solely on molecules for chemical tasks and on a combined corpus of natural language texts and string-based structures. The framework relies on molecule augmentations that preserve an underlying chemical, such as kekulization and cycle replacements. We evaluate encoder-only and generative LMs by calculating a metric based on the similarity score between distributed representations of molecules and their augmentations. Our experiments on ChEBI-20 and QM9 benchmarks show that these models exhibit significantly lower scores than graph-based molecular models trained without language modeling objectives. Additionally, our results on the molecule captioning task for cross-domain models, MolT5 and Text+Chem T5, demonstrate that the lower the representation-based evaluation metrics, the lower the classical text generation metrics like ROUGE and METEOR.

Nov 13 (Wed) 10:30-12:00 - Jasmine

LLMs for Generating and Evaluating Counterfactuals: A Comprehensive Study

Van Bach Nguyen, Paul Youssef, Jörg Schröterer, Christin Seifert

As NLP models become more complex, understanding their decisions becomes more crucial. Counterfactuals (CFs), where minimal changes to inputs flip a model's prediction, offer a way to explain these models. While Large Language Models (LLMs) have shown remarkable performance in NLP tasks, their efficacy in generating high-quality CFs remains uncertain. This work fills this gap by investigating how well LLMs generate CFs for three tasks. We conduct a comprehensive comparison of several common LLMs, and evaluate their CFs, assessing both intrinsic metrics, and the impact of these CFs on data augmentation. Moreover, we analyze differences between human and LLM-generated CFs, providing insights for future research directions. Our results show that LLMs generate fluent CFs, but struggle to keep the induced changes minimal. Generating CFs for Sentiment Analysis (SA) is less challenging than NLI and Hate Speech (HS) where LLMs show weaknesses in generating CFs that flip the original label. This also reflects on the data augmentation performance, where we observe a

large gap between augmenting with human and LLM CFs. Furthermore, we evaluate LLMs' ability to assess CFs in a mislabelled data setting, and show that they have a strong bias towards agreeing with the provided labels. GPT4 is more robust against this bias, but it shows strong preference to its own generations. Our analysis suggests that safety training is causing GPT4 to prefer its generations, since these generations do not contain harmful content. Our findings reveal several limitations and point to potential future work directions.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

On the Empirical Complexity of Reasoning and Planning in LLMs

Lwei Kang, Zirui Zhao, David Hsu, Wee Sun Lee

Chain-of-thought (CoT), tree-of-thought (ToT), and related techniques work surprisingly well in practice for some complex reasoning tasks with Large Language Models (LLMs), but why? This work seeks the underlying reasons by conducting experimental case studies and linking the performance benefit to well-established sample and computational complexity principles in machine learning. We experimented with six reasoning tasks, ranging from grade school math, air travel planning, ... to Blocksworld. The results suggest that (i) both CoT and ToT benefit significantly from task decomposition, which breaks a complex reasoning task into a sequence of steps with low sample complexity and explicitly outlines the reasoning structure; (ii) for computationally hard reasoning tasks, the more sophisticated tree structure of ToT outperforms the linear structure of CoT; (iii) explicitly annotating important variables is important for good performance. These findings provide useful guidelines for using LLM in solving reasoning tasks in practice.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

Tending Towards Stability: Convergence Challenges in Small Language Models

Richard Diehl Martinez, Pietro Lesci, Paula Buttery

Increasing the number of parameters in language models is a common strategy to enhance their performance. However, smaller language models remain valuable due to their lower operational costs. Despite their advantages, smaller models frequently underperform compared to their larger counterparts, even when provided with equivalent data and computational resources. Specifically, their performance tends to degrade in the late pretraining phase. This is anecdotally attributed to their reduced representational capacity. Yet, the exact causes of this performance degradation remain unclear. We use the Pythia model suite to analyse the training dynamics that underlie this phenomenon. Across different model sizes, we investigate the convergence of the Attention and MLP activations to their final state and examine how the effective rank of their parameters influences this process. We find that nearly all layers in larger models stabilise early in training - within the first 20% - whereas layers in smaller models exhibit slower and less stable convergence, especially when their parameters have lower effective rank. By linking the convergence of layers' activations to their parameters' effective rank, our analyses can guide future work to address inefficiencies in the learning dynamics of small models.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

Self-contradictory reasoning evaluation and detection

Ziyi Liu, Soumya Sanyal, Isabelle Lee, Yongkang Du, Rahul Gupta, Yang Liu, Jieyu Zhao

In a plethora of recent work, large language models (LLMs) demonstrated impressive reasoning ability, but many proposed downstream reasoning tasks only focus on performance-wise evaluation. Two fundamental questions persist: 1) how consistent is the reasoning, and 2) can models detect unreliable reasoning? In this paper, we investigate self-contradictory (Self-Contra) reasoning, where the model reasoning does not support answers. To answer 1), we define and assess the Self-Contra rate across three datasets and delve into finer-grained categories of Self-Contra reasoning. We find that LLMs often contradict themselves in reasoning tasks involving contextual information understanding or commonsense. The model may generate correct answers by taking shortcuts in reasoning or overlooking contextual evidence, leading to compromised reasoning. For 2), we task the state-of-the-art model GPT-4 with identifying Self-Contra reasoning and finer-grained fallacies. We find that finer-grained adetion detection can improve GPT-4's ability to detect Self-Contra. However, it is only able to detect Self-Contra with a 52.2% F1 score, much lower compared to 66.7% for humans. Our results indicate that current LLMs lack the robustness necessary for reliable reasoning and we emphasize the urgent need for establishing best practices in comprehensive reasoning evaluations beyond pure performance-based metrics.

Low-resource Methods for NLP 2

Nov 13 (Wed) 10:30-12:00 - Room: *Jasmine*

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

On Sensitivity of Learning with Limited Labelled Data to the Effects of Randomness: Impact of Interactions and Systematic Choices

Branislav Pečer, Ivan Šrba, Maria Bielikova

While learning with limited labelled data can effectively deal with a lack of labels, it is also sensitive to the effects of uncontrolled randomness introduced by so-called randomness factors (i.e., non-deterministic decisions such as choice or order of samples). We propose and formalise a method to systematically investigate the effects of individual randomness factors while taking the interactions (dependence) between them into consideration. To this end, our method mitigates the effects of other factors while observing how the performance varies across multiple runs. Applying our method to multiple randomness factors across in-context learning and fine-tuning approaches on 7 representative text classification tasks and meta-learning on 3 tasks, we show that: 1) disregarding interactions between randomness factors in existing works led to inconsistent findings due to incorrect attribution of the effects of randomness factors, such as disproving the consistent sensitivity of in-context learning to sample order even with random sample selection; and 2) besides mutual interactions, the effects of randomness factors, especially sample order, are also dependent on more systematic choices unexplored in existing works, such as number of classes, samples per class or choice of prompt format.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

Rethinking Token Reduction for State Space Models

Zheng Zhan, Yushu Wu, Zhenglin Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, Yanzhi Wang

Recent advancements in State Space Models (SSMs) have attracted significant interest, particularly in models optimized for parallel training and handling long-range dependencies. Architectures like Mamba have scaled to billions of parameters with selective SSM. To facilitate broader applications using Mamba, exploring its efficiency is crucial. While token reduction techniques offer a straightforward post-training strategy, we find that applying existing methods directly to SSMs leads to substantial performance drops. Through insightful analysis, we identify the reasons for this failure and the limitations of current techniques. In response, we propose a tailored, unified post-training token reduction method for SSMs. Our approach integrates token importance and similarity, thus taking advantage of both pruning and merging, to devise a fine-grained intra-layer token reduction strategy. Extensive experiments show that our method improves the average accuracy by 5.7% to 13.1% on six benchmarks with Mamba-2 compared to existing methods, while significantly reducing computational demands and memory requirements.

Nov 13 (Wed) 10:30-12:00 - Jasmine

MetaGPT: Merging Large Language Models Using Model Exclusive Task Arithmetic

Yuyan Zhou, Liang Song, Bingning Wang, weipeng chen

The advent of large language models (LLMs) like GPT-4 has catalyzed the exploration of multi-task learning (MTL), in which a single model demonstrates proficiency across diverse tasks. Task arithmetic has emerged as a cost-effective approach for MTL. It enables performance enhancement across multiple tasks by adding their corresponding task vectors to a pre-trained model. However, the current lack of a method that can simultaneously achieve optimal performance, computational efficiency, and data privacy limits their application to LLMs. In this paper, we propose **Model Exclusive Task Arithmetic** for merging **GPT-scale** models (MetaGPT) which formalizes the objective of model merging into a multi-task learning framework, aiming to minimize the average loss difference between the merged model and each individual task model. Since data privacy limits the use of multi-task training data, we leverage LLMs' local linearity and task vectors' orthogonality to separate the data term and scaling coefficients term and derive a model-exclusive task arithmetic method. Our proposed MetaGPT is data-agnostic and bypasses the heavy search process, making it cost-effective and easy to implement for LLMs. Extensive experiments demonstrate that MetaGPT leads to improvement of task arithmetic and achieves state-of-the-art performance on multiple tasks.

Nov 13 (Wed) 10:30-12:00 - Jasmine

An Effective Deployment of Diffusion LM for Data Augmentation in Low-Resource Sentiment Classification

Zhuowei Chen, Lianxi Wang, Xinfeng Liao, Yufia Tian, Junyang Zhong

Sentiment classification (SC) often suffers from low-resource challenges such as domain-specific contexts, imbalanced label distributions, and few-shot scenarios. The potential of the diffusion language model (LM) for textual data augmentation (DA) remains unexplored, moreover, textual DA methods struggle to balance the diversity and consistency of new samples. Most DA methods either perform logical modifications or rephrase less important tokens in the original sequence with the language model. In the context of SC, strong emotional tokens could act critically on the sentiment of the whole sequence. Therefore, contrary to rephrasing less important context, we propose Diffusion-CLS to leverage a diffusion LM to capture in-domain knowledge and generate pseudo samples by reconstructing strong label-related tokens. This approach ensures a balance between consistency and diversity, avoiding the introduction of noise and augmenting crucial features of datasets. DiffusionCLS also comprises a Noise-Resistant Training objective to help the model generalize. Experiments demonstrate the effectiveness of our method in various low-resource scenarios including domain-specific and domain-general problems. Ablation studies confirm the effectiveness of our framework's modules, and visualization studies highlight optimal deployment conditions, reinforcing our conclusions.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Self-Refine Instruction-Tuning for Aligning Reasoning in Language Models

Leonardo Ranaldi, Andre Freitas

The alignment of reasoning abilities between smaller and larger Language Models are largely conducted via supervised fine-tuning using demonstrations generated from robust Large Language Models (LLMs). Although these approaches deliver more performant models, they do not show sufficiently strong generalization ability as the training only relies on the provided demonstrations. In this paper, we propose the Self-refine Instruction-tuning method that elicits Smaller Language Models to self-improve their abilities. Our approach is based on a two-stage process, where reasoning abilities are first transferred between LLMs and Small Language Models (SLMs) via Instruction-tuning on synthetic demonstrations provided by LLMs, and then the instructed models self-improve their abilities through preference optimization strategies. In particular, the second phase operates refinement heuristics based on Direct Preference Optimization, where the SLMs are elicited to deliver a series of reasoning paths by automatically sampling the generated responses and providing rewards using ground truths from the LLMs. Results obtained on commonsense and math reasoning tasks show that this approach consistently outperforms Instruction-tuning in both in-domain and out-domain scenarios, aligning the reasoning abilities of Smaller and Larger language models.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Collaborative Performance Prediction for Large Language Models

Qiyuan Zhang, Fuyuan Lyu, Xue Liu, Chen Ma

Comprehensively understanding and accurately predicting the performance of large language models across diverse downstream tasks has emerged as a pivotal challenge in NLP research. The pioneering scaling law on downstream works demonstrated intrinsic similarities within model families and utilized such similarities for performance prediction. However, they tend to overlook the similarities between model families and only consider design factors listed in the original scaling law. To overcome these limitations, we introduce a novel framework, Collaborative Performance Prediction (CPP), which significantly enhances prediction accuracy by leveraging the historical performance of various models on downstream tasks and other design factors for both model and task. We also collect a collaborative data sourced from online platforms containing both historical performance and additional design factors. With the support of the collaborative data, CPP not only surpasses traditional scaling laws in predicting the performance of scaled LLMs but also facilitates a detailed analysis of factor importance, an area previously overlooked.

Nov 13 (Wed) 10:30-12:00 - Jasmine

A Generic Method for Fine-grained Category Discovery in Natural Language Texts

Chang Tian, Matthew B. Blaschko, Wengpeng Yin, Mingzhe Xing, Yintiang Yue, Marie-Francine Moens

Fine-grained category discovery using only coarse-grained supervision is a cost-effective yet challenging task. Previous training methods focus on aligning query samples with positive samples and distancing them from negatives. They often neglect intra-category and inter-category semantic similarities of fine-grained categories when navigating sample distributions in the embedding space. Furthermore, some evaluation techniques that rely on pre-collected test samples are inadequate for real-time applications. To address these shortcomings, we introduce a method that successfully detects fine-grained clusters of semantically similar texts guided by a novel objective function. The method uses semantic similarities in a logarithmic space to guide sample distributions in the Euclidean space and to form distinct clusters that represent fine-grained categories. We also propose a centroid inference mechanism to support real-time applications. The efficacy of the method is both theoretically justified and empirically confirmed on three benchmark tasks. The proposed objective function is integrated in multiple contrastive learning based neural models. Its results surpass existing state-of-the-art approaches in terms of Accuracy, Adjusted Rand Index and Normalized Mutual Information of the detected fine-grained categories. Code and data are publicly available at <https://github.com/changtianluckyforever/F-grained-STAR>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Text Grafting: Near-Distribution Weak Supervision for Minority Classes in Text Classification

Letian Peng, Yi Gu, Chengyu Dong, Zihan Wang, Jingbo Shang

For extremely weak-supervised text classification, pioneer research generates pseudo labels by mining texts similar to the class names from the raw corpus, which may end up with very limited or even no samples for the minority classes. Recent works have started to generate the relevant texts by prompting LLMs using the class names or definitions; however, there is a high risk that LLMs cannot generate in-distribution (i.e., similar to the corpus where the text classifier will be applied) data, leading to ungeneralizable classifiers. In this paper, we combine the advantages of these two approaches and propose to bridge the gap via a novel framework, *text grafting*, which aims to obtain clean and

near-distribution weak supervision for minority classes. Specifically, we first use LLM-based logits to mine masked templates from the raw corpus, which have a high potential for data synthesis into the target minority class. Then, the templates are filled by state-of-the-art LLMs to synthesize near-distribution texts falling into minority classes. Text grafting shows significant improvement over direct mining or synthesis on minority classes. We also use analysis and case studies to comprehend the property of text grafting.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Incubating Text Classifiers Following User Instruction with Nothing but LLM

Letian Peng, Zilong Wang, Jingbo Shang

In this paper, we aim to generate text classification data given arbitrary class definitions (i.e., user instruction), so one can train a text classifier without any human annotation or raw corpus. Recent advances in large language models (LLMs) lead to pioneer attempts to individually generate texts for each class via prompting. In this paper, we propose Incubator, the first framework that can handle complicated and even mutually dependent classes (e.g., "TED Talk given by Educator" and "Other"). Specifically, our Incubator is a fine-tuned LLM that takes the instruction of all class definitions as input, and in each inference, it can jointly generate one sample for every class. First, we tune Incubator on the instruction-to-data mappings that we obtained from classification datasets and descriptions on Hugging Face together with in-context augmentation by GPT-4. To emphasize the uniformity and diversity in generations, we refine Incubator by fine-tuning with the cluster centers of semantic textual embeddings of the generated samples. We compare Incubator on various classification tasks with strong baselines such as direct LLM-based inference and training data generation by prompt engineering. Experiments show Incubator is able to (1) outperform previous methods on traditional benchmarks, (2) take label interdependency and user preference into consideration, and (3) enable logical text mining by incubating multiple classifiers.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Interpretability-based Tailored Knowledge Editing in Transformers

Yihuai Hong, Aldo Lipani

Language models recognized as a new form of knowledge bases, face challenges of outdated, erroneous, and privacy-sensitive information, necessitating knowledge editing to rectify errors without costly retraining. Existing methods, spanning model's parameters modification, external knowledge integration, and in-context learning, lack in-depth analysis from a model interpretability perspective. Our work explores the instability in in-context learning outcomes, providing insights into its reasons and distinctions from other methods. Leveraging findings on the critical role of feed-forward MLPs in decoder-only models, we propose a tailored knowledge editing method, TailoredKE, that considers the unique information flow of each sample. Model interpretability reveals diverse attribute recall across transformer layers, guiding edits to specific features at different depths and mitigating over-editing issues.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Order of Magnitude Speedups for LLM Membership Inference

Rongting Zhang, Martin Andres Bertran, Aaron Roth

Large Language Models (LLMs) have the promise to revolutionize computing broadly, but their complexity and extensive training data also expose significant privacy vulnerabilities. One of the simplest privacy risks associated with LLMs is their susceptibility to membership inference attacks (MIA), wherein an adversary aims to determine whether a specific data point was part of the models training set. Although this is a known risk, state of the art methodologies for MIA rely on training multiple computationally costly 'shadow models', making risk evaluation prohibitive for large models. Here we adapt a recent line of work which uses quantile regression to mount membership inference attacks; we extend this work by proposing a low-cost MIA that leverages an ensemble of small quantile regression models to determine if a document belongs to the model's training set or not. We demonstrate the effectiveness of this approach on fine-tuned LLMs of varying families (OPT, Pythia, Llama) and across multiple datasets. Across all scenarios we obtain comparable or improved accuracy compared to state of the art 'shadow model' approaches, with as little as 6% of their computation budget. We demonstrate increased effectiveness across multi-epoch trained target models, and architecture miss-specification robustness, that is, we can mount an effective attack against a model using a different tokenizer and architecture, without requiring knowledge on the target model.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Do Not Worry if You Do Not Have Data: Building Pretrained Language Models Using Translationese

Meet Doshi, Raj Dabre, Pushpak Bhattacharya

In this paper, we explore the utility of Translationese as synthetic data created using machine translation for pre-training language models (LMs) for low-resource languages (LRLs). Our simple methodology consists of translating large amounts of web-crawled monolingual documents (clean) into the LRLs, followed by filtering the translated documents using tiny LMs trained on small but clean LRL data. Taking the case of Indian languages, we pre-train LMs from scratch with 28M and 85M parameters, and then fine-tune them for 5 downstream natural language understanding (NLU) and 4 generative (NLG) tasks. We observe that pre-training on filtered synthetic data leads to relative performance drops of only 0.87% for NLU and 2.35% for NLG, compared to pre-training on clean data, and this gap further diminishes upon the inclusion of a small amount of clean data. We also study the impact of synthetic data filtering and the choice of source language for synthetic data generation. Furthermore, evaluating continually pre-trained larger models like Gemma-2B and Llama-3-8B in few-shot settings, we observe that using synthetic data is competitive with using clean data. Our findings suggest that synthetic data shows promise for bridging the pre-training gap between English and LRLs.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Less is More: Parameter-Efficient Selection of Intermediate Tasks for Transfer Learning

David Schulte, Felix Hamborg, Alan Akbik

Intermediate task transfer learning can greatly improve model performance. If, for example, one has little training data for emotion detection, first fine-tuning a language model on a sentiment classification dataset may improve performance strongly. But which task to choose for transfer learning? Prior methods producing useful task rankings are infeasible for large source pools, as they require forward passes through all source language models. We overcome this by introducing Embedding Space Maps (ESMs), light-weight neural networks that approximate the effect of fine-tuning a language model. We conduct the largest study on NLP task transferability and task selection with 12k source-target pairs. We find that applying ESMs on a prior method reduces execution time and disk space usage by factors of 10 and 278, respectively, while retaining high selection performance (avg. regret@5 score of 2.95).

Nov 13 (Wed) 10:30-12:00 - Jasmine

Fisher Information-based Efficient Curriculum Federated Learning with Large Language Models

Ji Liu, Jiaxiang Ren, Ruoming Jin, Zijie Zhang, Yang Zhou, Patrick Valdoriez, Dejing Dou

As a promising paradigm to collaboratively train models with decentralized data, Federated Learning (FL) can be exploited to fine-tune Large Language Models (LLMs). While LLMs correspond to huge size, the scale of the training data significantly increases, which leads to tremendous amounts of computation and communication costs. The training data is generally non-Independent and Identically Distributed (non-IID), which requires adaptive data processing within each device. Although Low-Rank Adaptation (LoRA) can significantly reduce the scale of parameters to update in the fine-tuning process, it still takes unaffordable time to transfer the low-rank parameters of all the layers in LLMs. In

this paper, we propose a Fisher Information-based Efficient Curriculum Federated Learning framework (FibecFed) with two novel methods, i.e., adaptive federated curriculum learning and efficient sparse parameter update. First, we propose a fisher information-based method to adaptively sample data within each device to improve the effectiveness of the FL fine-tuning process. Second, we dynamically select the proper layers for global aggregation and sparse parameters for local update with LoRA so as to improve the efficiency of the FL fine-tuning process. Extensive experimental results based on 10 datasets demonstrate that FibecFed yields excellent performance (up to 45.35% in terms of accuracy) and superb fine-tuning speed (up to 98.61% faster) compared with 17 baseline approaches.

Nov 13 (Wed) 10:30-12:00 - Jasmine

TEMA: Token Embeddings Mapping for Enriching Low-Resource Language Models

Rodolfo Zevallos, Núria Bel, Mireia Farrús

The objective of the research we present is to remedy the problem of the low quality of language models for low-resource languages. We introduce an algorithm, the Token Embedding Mapping Algorithm (TEMA), that maps the token embeddings of a richly pre-trained model L1 to a poorly trained model L2, thus creating a richer L2 model. Our experiments show that the L2 model reduces perplexity with respect to the original monolingual model L2, and that for downstream tasks, including SuperGLUE, the results are state-of-the-art or better for the most semantic tasks. The models obtained with TEMA are also competitive or better than multilingual or extended models proposed as solutions for mitigating the low-resource language problems.

Nov 13 (Wed) 10:30-12:00 - Jasmine

DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging

Tzu-Han Lin, Chen-An Li, Hung-yi Lee, Yun-Nung Chen

Reinforcement learning from human feedback (RLHF) is a popular strategy for aligning large language models (LLMs) with desired behaviors. Reward modeling is a crucial step in RLHF. However, collecting paired preference data for training reward models is often costly and time-consuming, especially for domain-specific preferences requiring expert annotation. To address this challenge, we propose the **Do**-*main knowledge*-ge** merged ***R***-reward ***M***-odel (**DogeRM**), a novel framework that integrates domain-specific knowledge into a general reward model by model merging. The experiments demonstrate that DogeRM enhances performance across different benchmarks and provide a detailed analysis showcasing the effects of model merging, showing the great potential of facilitating model alignment.

Nov 13 (Wed) 10:30-12:00 - Jasmine

HiFT: A Hierarchical Full Parameter Fine-Tuning Strategy

YongKang Liu, Yiqun Zhang, Qian Li, Tong Liu, Shi Feng, Daling Wang, Yifei Zhang, Hinrich Schuetze

Full-parameter fine-tuning (FPFT) has become the go-to choice for adapting language models (LMs) to downstream tasks due to its excellent performance. As LMs grow in size, fine-tuning the full parameters of LMs requires a prohibitively large amount of GPU memory. Existing approaches utilize zeroth-order optimizer to conserve GPU memory, which potentially compromises the performance of LMs as non-zero order optimizers tend to converge more readily on most downstream tasks. We propose a novel, memory-efficient, optimizer-independent, end-to-end hierarchical fine-tuning strategy, HiFT, which only updates a subset of parameters at each training step. HiFT significantly reduces the amount of gradients and optimizer state parameters residing in GPU memory at the same time, thereby reducing GPU memory usage. Our results demonstrate that: (1) HiFT achieves comparable performance with parameter-efficient fine-tuning and standard FPFT. (2) Results on six models show that HiFT reduces the number of trainable parameters by about 89.18% on average compared to FPFT. (3) HiFT supports FPFT of 7B models for 24G GPU memory devices under mixed precision without using any memory saving techniques. (4) HiFT supports various optimizers including AdamW, AdaGrad, SGD, etc. The source code link is <https://github.com/misnksy/HiFT>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Link, Synthesize, Retrieve: Universal Document Linking for Zero-Shot Information Retrieval

Dae Yon Hwang, Bilal Tahar, Harshit Pande, Yaroslav Nechaev

Despite the recent advancements in information retrieval (IR), zero-shot IR remains a significant challenge, especially when dealing with new domains, languages, and newly-released use cases that lack historical query traffic from existing users. For such cases, it is common to use query augmentations followed by fine-tuning pre-trained models on the document data paired with synthetic queries. In this work, we propose a novel Universal Document Linking (UDL) algorithm, which links similar documents to enhance synthetic query generation across multiple datasets with different characteristics. UDL leverages entropy for the choice of similarity models and named entity recognition (NER) for the link decision of documents using similarity scores. Our empirical studies demonstrate the effectiveness and universality of the UDL across diverse datasets and IR models, surpassing state-of-the-art methods in zero-shot cases. The developed code for reproducibility is included in <https://github.com/edoduself/UDL>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

A Two-Step Approach for Data-Efficient French Pronunciation Learning

Hoyeon Lee, Hyeun Jang, JONGHWAN KIM, Jaemin Kim

Recent studies have addressed intricate phonological phenomena in French, relying on either extensive linguistic knowledge or a significant amount of sentence-level pronunciation data. However, creating such resources is expensive and non-trivial. To this end, we propose a novel two-step approach that encompasses two pronunciation tasks: grapheme-to-phone and post-lexical processing. We then investigate the efficacy of the proposed approach with a notably limited amount of sentence-level pronunciation data. Our findings demonstrate that the proposed two-step approach effectively mitigates the lack of extensive labeled data, and serves as a feasible solution for addressing French phonological phenomena even under resource-constrained environments.

Nov 13 (Wed) 10:30-12:00 - Jasmine

DEFT-UCS: Data Efficient Fine-Tuning for Pre-Trained Language Models via Unsupervised Core-Set Selection

Develeena Das, Vivek Khetan

Recent advances have led to the availability of many pre-trained language models (PLMs); however, a question that remains is how much data is truly needed to fine-tune PLMs for downstream tasks? In this work, we introduce DEFT-UCS, a data-efficient fine-tuning framework that leverages unsupervised core-set selection to identify a smaller, representative dataset to fine-tune PLMs for text-generation needed for text editing tasks such as simplification, grammar correction, clarity, etc. We examine the efficacy of DEFT-UCS across multiple text-editing tasks, and compare to the state-of-the-art text-editing model, CoEDIT. Our results demonstrate that DEFT-UCS models are just as accurate as CoEDIT, across eight different datasets consisting of six different editing tasks, while finetuned on 70% less data.

Nov 13 (Wed) 10:30-12:00 - Jasmine

CoverICL: Selective Annotation for In-Context Learning via Active Graph Coverage

Costas Mavromatis, Balasubramanian Srinivasan, Zhengyuan Shen, Jian Zhang, Huzeifa Rangwala, Christos Faloutsos, George Karypis

In-context learning (ICL) adapts Large Language Models (LLMs) to new tasks, without requiring any parameter updates, but few annotated examples as input. In this work, we investigate selective annotation for ICL, where there is a limited budget for annotating examples, similar to low-budget active learning (AL). Although uncertainty-based selection is unreliable with few annotated data, we present CoverICL, an

adaptive graph-based selection algorithm, that effectively incorporates uncertainty sampling into selective annotation for ICL. First, CoverICL builds a nearest-neighbor graph based on the semantic similarity between candidate ICL examples. Then, CoverICL employs uncertainty estimation by the LLM to identify hard examples for the task. Selective annotation is performed over the active graph of the hard examples, adapting the process to the particular LLM used and the task tackled. CoverICL selects the most representative examples by solving a Maximum Coverage problem, approximating diversity-based sampling. Extensive experiments on ten datasets and seven LLMs show that, by incorporating uncertainty via coverage on the active graph, CoverICL (1) outperforms existing AL methods for ICL by 2–4.6% accuracy points, (2) is up to 2x more budget-efficient than SOTA methods for low-budget AL, and (3) generalizes better across tasks compared to non-graph alternatives.

Nov 13 (Wed) 10:30-12:00 - Jasmine

DADEE: Unsupervised Domain Adaptation in Early Exit PLMs

Divya Jyoti Bajpai, Manjesh Kumar Hanawal

Pre-trained Language Models (PLMs) exhibit good accuracy and generalization ability across various tasks using self-supervision, but their large size results in high inference latency. Early Exit (EE) strategies handle the issue by allowing the samples to exit from classifiers attached to the intermediary layers, but they do not generalize well, as exit classifiers can be sensitive to domain changes. To address this, we propose Unsupervised Domain Adaptation in EE framework (DADEE) that employs multi-level adaptation using knowledge distillation. DADEE utilizes GAN-based adversarial adaptation at each layer to achieve domain-invariant representations, reducing the domain gap between the source and target domain across all layers. The attached exits not only speed up inference but also enhance domain adaptation by reducing catastrophic forgetting and mode collapse, making it more suitable for real-world scenarios. Experiments on tasks such as sentiment analysis, entailment classification, and natural language inference demonstrate that DADEE consistently outperforms not only early exit methods but also various domain adaptation methods under domain shift scenarios. The anonymized source code is available at <https://github.com/Div290/DADEE>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

CapEEN: Image Captioning with Early Exits and Knowledge Distillation

Divya Jyoti Bajpai, Manjesh Kumar Hanawal

Deep neural networks (DNNs) have made significant progress in recognizing visual elements and generating descriptive text in image-captioning tasks. However, their improved performance comes from increased computational burden and inference latency. Early Exit (EE) strategies can be used to enhance their efficiency, but their adaptation presents challenges in image captioning as it requires varying levels of semantic information for accurate predictions. To overcome this, we introduce CapEEN to improve the performance of EE strategies using knowledge distillation. Inference in CapEEN is completed at intermediary layers if prediction confidence exceeds a predefined value learned from the training data. To account for real-world deployments, where target distributions could drift from that of training samples, we introduce a variant A-CapEEN to adapt the thresholds on the fly using Multi-armed bandits framework. Experiments on the MS COCO and Flickr30k datasets show that CapEEN gains speedup of $1.77 \times$ while maintaining competitive performance compared to the final layer, and A-CapEEN additionally offers robustness against distortions. The source code is available at <https://github.com/Div290/CapEEN>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Intermediate Layer Distillation with the Reused Teacher Classifier: A Study on the Importance of the Classifier of Attention-based Models

Hang Zhang, Seyyed Hasan Mozafari, James J. Clark, Brett H. Meyer, Warren J. Gross

Intermediate Layer Distillation (ILD) effectively compresses large-scale pre-trained language models (PLMs). Existing ILD methods underestimate the importance of utilizing the teacher's discriminative classifier and face challenges in establishing proper layer mappings. Therefore, we propose ILD-RTC, to show that a straightforward implementation of reusing the pre-trained teacher classifier improves student performance even with simple uniform layer mapping. Through extensive experiments, our method outperforms other ILD techniques, maintaining 97.7% performance of the original teacher BERT._{base} without additional trainable parameters. Projectors are developed to help the student match the hidden size of the teacher model, making our ILD-RTC applicable to students with different sizes. In addition, our technique achieves the same average GLUE score as students initialized by pre-trained LMs, saving over $80 \times$ cost resulting from the pre-training step. Our method emphasizes the reuse of pre-trained teacher classifiers as an alternative to pre-training the student for initialization.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Active Learning for Abstractive Text Summarization via LLM-Determined Curriculum and Certainty Gain Maximization

Dongyuan Li, Ying Zhang, Zhen Wang, Shiyin Tan, Satoshi Kosugi, Manabu Okumura

For abstractive text summarization, laborious data annotation and time-consuming model training become two high walls, hindering its further progress. Active Learning, selecting a few informative instances for annotation and model training, sheds light on solving these issues. However, only few active learning-based studies focus on abstractive text summarization and suffer from low stability, effectiveness, and efficiency. To solve the problems, we propose a novel LLM-determined curriculum active learning framework. Firstly, we design a prompt to ask large language models to rate the difficulty of instances, which guides the model to train on from easier to harder instances. Secondly, we design a novel active learning strategy, i.e., Certainty Gain Maximization, enabling to select instances whose distribution aligns well with the overall distribution. Experiments show our method can improve stability, effectiveness, and efficiency of abstractive text summarization backbones.

Nov 13 (Wed) 10:30-12:00 - Jasmine

LaMDA: Large Model Fine-Tuning via Spectrally Decomposed Low-Dimensional Adaptation

Seyedarmin Azizi, Souvik Kundu, Massoud Pedram

Low-rank adaptation (LoRA) has become the default approach to fine-tune large language models (LLMs) due to its significant reduction in trainable parameters. However, trainable parameter demand for LoRA increases with increasing model embedding dimensions, leading to high compute costs. Additionally, its backward updates require storing high-dimensional intermediate activations and optimizer states, demanding high peak GPU memory. In this paper, we introduce LaMDA₋, a novel approach to fine-tuning large language models, which leverages low-dimensional adaptation to achieve significant reductions in trainable parameters and peak GPU memory footprint. LaMDA freezes a first projection matrix (PMA) in the adaptation path while introducing a low-dimensional trainable square matrix, resulting in substantial reductions in trainable parameters and peak GPU memory usage. LaMDA gradually freezes a second projection matrix (PMB) during the early fine-tuning stages, reducing the compute cost associated with weight updates to enhance parameter efficiency further. We also present an enhancement, LaMDA++, incorporating a "lite-weight" adaptive rank allocation for the LoRA path via normalized spectrum analysis of pre-trained model weights. We evaluate LaMDA/LaMDA++ across various tasks, including natural language understanding with the GLUE benchmark, text summarization, natural language generation, and complex reasoning on different LLMs. Results show that LaMDA matches or surpasses the performance of existing alternatives while requiring up to $**17.7 \times **$ fewer parameter updates and up to $**1.32 \times **$ lower peak GPU memory usage during fine-tuning. Code will be publicly available at <https://github.com/ArminAzizi98/LaMDA>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

MORL-Prompt: An Empirical Analysis of Multi-Objective Reinforcement Learning for Discrete Prompt Optimization

Yasaman Jafari, Dheeraj Mekala, Rose Yu, Taylor Berg-Kirkpatrick

RL-based techniques can be employed to search for prompts that, when fed into a target language model, maximize a set of user-specified reward functions. However, in many target applications, the natural reward functions are in tension with one another – for example, content preservation vs. style matching in style transfer tasks. Current techniques focus on maximizing the average of reward functions, which does not necessarily lead to prompts that achieve balance across rewards – an issue that has been well-studied in the multi-objective and robust optimization literature. In this paper, we conduct an empirical comparison of several existing multi-objective optimization techniques adapted to this new setting: RL-based discrete prompt optimization. We compare two methods optimizing the volume of the Pareto reward surface and one method that chooses an update direction that benefits all rewards simultaneously. We evaluate performance on two NLP tasks: style transfer and machine translation, each using three competing reward functions. Our experiments demonstrate that multi-objective methods that directly optimize the volume of the Pareto reward surface perform better and achieve a better balance of all rewards than those that attempt to find monotonic update directions.

Nov 13 (Wed) 10:30-12:00 - Jasmine

TAP4LLM: Table Provider on Sampling, Augmenting, and Packing Semi-structured Data for Large Language Model Reasoning

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, Dongmei Zhang

Table reasoning tasks have shown remarkable progress with the development of large language models (LLMs), which involve interpreting and drawing conclusions from tabular data based on natural language (NL) questions. Existing solutions mainly tested on smaller tables face scalability issues and struggle with complex queries due to incomplete or dispersed data across different table sections. To alleviate these challenges, we propose TAP4LLM as a versatile pre-processor suite for leveraging LLMs in table-based tasks effectively. It covers several distinct components: (1) table sampling to decompose large tables into manageable sub-tables based on query semantics, (2) table augmentation to enhance tables with additional knowledge from external sources or models, and (3) table packing & serialization to convert tables into various formats suitable for LLMs' understanding. In each module, we design and compare several common methods for usage in various scenarios, aiming to shed light on the best practices for leveraging LLMs for table-reasoning tasks. Our experiments show that our method improves LLMs' reasoning capabilities in various tabular tasks and enhances the interaction between LLMs and tabular data by employing effective pre-processing.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Self-training Language Models in Arithmetic Reasoning

Marek Kadlik, Michal Šefánik

Recent language models achieve impressive results in tasks involving complex multistep reasoning, but scaling these capabilities further traditionally requires expensive collection of more annotated data. In this work, we explore the potential of improving models' reasoning capabilities without new data, merely using automated feedback to the validity of their predictions in arithmetic reasoning (self-training). In systematic experimentation across six different arithmetic reasoning datasets, we find that models can substantially improve in both single-round (offline) and online self-training, reaching a correct result in +13.9% and +25.9% more cases, respectively, underlining the importance of actuality of self-training feedback. We further find that in the single-round, offline self-training, traditional supervised training can deliver gains comparable to preference optimization, but in online self-training, preference optimization methods largely outperform supervised training thanks to their superior stability and robustness on unseen types of problems.

Nov 13 (Wed) 10:30-12:00 - Jasmine

All You Need is Attention: Lightweight Attention-based Data Augmentation for Text Classification

Junehyung Kim, Sungjae Hwang

This paper introduces LADAM, a novel method for enhancing the performance of text classification tasks. LADAM employs attention mechanisms to exchange semantically similar words between sentences. This approach generates a greater diversity of synthetic sentences compared to simpler operations like random insertions, while maintaining the context of the original sentences. Additionally, LADAM is an easy-to-use, lightweight technique that does not require external datasets or large language models. Our experimental results across five datasets demonstrate that LADAM consistently outperforms baseline methods across diverse text classification conditions.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Generate then Refine: Data Augmentation for Zero-shot Intent Detection

I-Fan Lin, Faegheh Hasibi, Suzan Verberne

In this short paper we propose a data augmentation method for intent detection in zero-resource domains. Existing data augmentation methods rely on few labelled examples for each intent category, which can be expensive in settings with many possible intents. We use a two-stage approach: First, we generate utterances for intent labels using an open-source large language model in a zero-shot setting. Second, we develop a smaller sequence-to-sequence model (the Refiner), to improve the generated utterances. The Refiner is fine-tuned on seen domains and then applied to unseen domains. We evaluate our method by training an intent classifier on the generated data, and evaluating it on real (human) data. We find that the Refiner significantly improves the data utility and diversity over the zero-shot LLM baseline for unseen domains and over common baseline approaches. Our results indicate that a two-step approach of a generative LLM in zero-shot setting and a smaller sequence-to-sequence model can provide high-quality data for intent detection.

Nov 13 (Wed) 10:30-12:00 - Jasmine

BanglaTLit: A Benchmark Dataset for Back-Transliteration of Romanized Bangla

Md Fahim, Fariba Tanjim Shifat, Md Farhan Ishamm, Deeparghya Dutta Barua, Fabiha Haider, MD SAKIB UL RAHMAN SOUROVE, Md Farhad Alam Bhuiyan

Low-resource languages like Bangla are severely limited by the lack of datasets. Romanized Bangla texts are ubiquitous on the internet, offering a rich source of data for Bangla NLP tasks and extending the available data sources. However, due to the informal nature of romanized text, they often lack the structure and consistency needed to provide insights. We address these challenges by proposing: (1) BanglaTLit, the large-scale Bangla transliteration dataset consisting of 42.7k samples, (2) BanglaTLit-PT, a pre-training corpus on romanized Bangla with 245.7k samples, (3) encoders further-pretrained on BanglaTLit-PT achieving state-of-the-art performance in several romanized Bangla classification tasks, and (4) multiple back-transliteration baseline methods, including a novel encoder-decoder architecture using further pre-trained encoders. Our results show the potential of automated Bangla back-transliteration in utilizing the untapped sources of romanized Bangla to enrich this language. The code and datasets are publicly available: <https://github.com/farhanishmam/BanglaTLit>.

Nov 13 (Wed) 10:30-12:00 - Jasmine

Self-training Large Language Models through Knowledge Detection

Yeo Wei Jie, Teddy Ferdinand, Przemyslaw Kazienko, Ranjan Satapathy, Erik Cambria

Large language models (LLMs) often necessitate extensive labeled datasets and training compute to achieve impressive performance across downstream tasks. This paper explores a self-training paradigm, where the LLM autonomously curates its own labels and selectively trains on unknown data samples identified through a reference-free consistency method. Empirical evaluations demonstrate significant improvements

in reducing hallucination in generation across multiple subjects. Furthermore, the selective training framework mitigates catastrophic forgetting in out-of-distribution benchmarks, addressing a critical limitation in training LLMs. Our findings suggest that such an approach can substantially reduce the dependency on large labeled datasets, paving the way for more scalable and cost-effective language model training.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

XC-Cache: Cross-Attending to Cached Context for Efficient LLM Inference

Joao Monteiro, Etienne Marcotte, Pierre-Andre Noel, Valentina Zantedeschi, David Vazquez, Nicolas Chapados, Christopher Pal, Perouz Taslakian

Prompts are often employed to condition decoder-only language model generation on reference information. Just-in-time processing of a context is inefficient due to the quadratic cost of self-attention operations, and caching is desirable. However, caching transformer states can easily require almost as much space as the model parameters. When the right context is not known in advance, caching the prompt can be challenging. This work addresses these limitations by introducing models that, inspired by the encoder-decoder architecture, use cross-attention to condition generation on reference text without the prompt. More precisely, we leverage pre-trained decoder-only models and only train a small number of added layers. We use Question-Answering (QA) as a testbed to evaluate the ability of our models to perform conditional generation and observe that they outperform prompt-based inference methods, are comparable to fine-tuned prompted LLMs, and drastically reduce the space footprint relative to standard KV caching by two orders of magnitude. Specifically, we introduced XC-Llama which converts a pre-trained Llama 2 into an encoder-decoder architecture by integrating cross-attention layers interleaved in between existing self-attention layers.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

Gloss2Text: Sign Language Gloss translation using LLMs and Semantically Aware Label Smoothing

Pooya Fayazsanavi, Antonios Anastasopoulos, Jana Kosecka

Sign language translation from video to spoken text presents unique challenges owing to the distinct grammar, expression nuances, and high variation of visual appearance across different speakers and contexts. Gloss annotations serve as an intermediary to guide the translation process. In our work, we focus on *Gloss2Text* translation stage and propose several advances by leveraging pre-trained large language models (LLMs), data augmentation, and novel label-smoothing loss function exploiting gloss translation ambiguities improving significantly the performance of state-of-the-art approaches. Through extensive experiments and ablation studies on the PHOENIX Weather 2014T dataset, our approach surpasses state-of-the-art performance in *Gloss2Text* translation, indicating its efficacy in addressing sign language translation and suggesting promising avenues for future research and development.

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

Laywiser Importance Matters: Less Memory for Better Performance in Parameter-efficient Fine-tuning of Large Language Models

Kai Yao, Penglei Gao, Lichun Li, Yuan Zhao, Xiaofeng Wang, Wei Wang, Jianke Zhu

Parameter-Efficient Fine-Tuning (PEFT) methods have gained significant popularity for adapting pre-trained Large Language Models (LLMs) to downstream tasks, primarily due to their potential to significantly reduce memory and computational overheads. However, a common limitation in most PEFT approaches is their application of a uniform architectural design across all layers. This uniformity involves identical trainable modules and ignores the varying importance of each layer, leading to sub-optimal fine-tuning results. To overcome the above limitation and obtain better performance, we develop a novel approach, Importance-aware Sparse Tuning (IST), to fully utilize the inherent sparsity and select the most important subset of full layers with effective layer-wise importance scoring. The proposed IST is a versatile and plug-and-play technique compatible with various PEFT methods that operate on a per-layer basis. By leveraging the estimated importance scores, IST dynamically updates these selected layers in PEFT modules, leading to reduced memory demands. We further provide theoretical proof of convergence and empirical evidence of superior performance to demonstrate the advantages of IST over uniform updating strategies. Extensive experiments on a range of LLMs, PEFTs, and downstream tasks substantiate the effectiveness of our proposed method, showcasing IST's capacity to enhance existing layer-based PEFT methods. Our code is available at <https://github.com/Kaiseem/IST>

Nov 13 (Wed) 10:30-12:00 - *Jasmine*

InfuserKI: Enhancing Large Language Models with Knowledge Graphs via Infuser-Guided Knowledge Integration

Fali Wang, Runxue Bao, Suhang Wang, Wenchao Yu, Yanchi Liu, Wei Cheng, Haifeng Chen

Large Language Models (LLMs) have achieved exceptional capabilities in open generation across various domains, yet they encounter difficulties with tasks that require intensive knowledge. To address these challenges, methods for integrating knowledge have been developed, which augment LLMs with domain-specific knowledge graphs through external modules. These approaches, however, face data inefficiency issues as they necessitate the processing of both known and unknown knowledge for fine-tuning. Thus, our research focuses on a novel problem: efficiently integrating unknown knowledge into LLMs without unnecessary overlap of known knowledge. A risk of introducing new knowledge is the potential forgetting of existing knowledge. To mitigate this risk, we propose the innovative InfuserKI framework. This framework employs transformer internal states to determine when to enrich LLM outputs with additional information, effectively preventing knowledge forgetting. Performance evaluations using the UMLS-2.5k and MetaQA domain knowledge graphs reveal that InfuserKI not only successfully integrates new knowledge but also outperforms state-of-the-art baselines, reducing knowledge forgetting by 9% and 6%, respectively.

NLP Applications 2

Nov 13 (Wed) 10:30-12:00 - Room: Riverfront Hall

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

"Glue pizza and eat rocks" - Exploiting Vulnerabilities in Retrieval-Augmented Generative Models

Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, huan liu

Retrieval-Augmented Generative (RAG) models enhance Large Language Models (LLMs) by integrating external knowledge bases, improving their performance in applications like fact-checking and information searching. In this paper, we demonstrate a security threat where adversaries can exploit the openness of these knowledge bases by injecting deceptive content into the retrieval database, intentionally changing the models behavior. This threat is critical as it mirrors real-world usage scenarios where RAG systems interact with publicly accessible knowledge bases, such as web scrapings and user-contributed data pools. To be more realistic, we target a realistic setting where the adversary has no knowledge of users' queries, knowledge base data, and the LLM parameters. We demonstrate that it is possible to exploit the model successfully through crafted content uploads with access to the retriever. Our findings emphasize an urgent need for security measures in the design and deployment of RAG systems to prevent potential manipulation and ensure the integrity of machine-generated content.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

A B B A: Evaluating and Improving Logical Reasoning Ability of Large Language Models

Yuxuan WAN, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, Michael Lyu

We introduce LogicAsker, a novel approach for evaluating and enhancing the logical reasoning capabilities of large language models (LLMs) such as ChatGPT and GPT-4. Despite LLMs' prowess in tasks like writing assistance, code generation, and machine translation, assessing their ability to reason has been challenging. Traditional evaluations often prioritize accuracy on downstream tasks over direct assessments of reasoning processes. LogicAsker addresses this gap by employing a set of atomic reasoning skills grounded in propositional and predicate logic to systematically examine and improve the reasoning prowess of LLMs. Our methodology reveals significant gaps in LLMs' learning of logical rules, with identified reasoning failures ranging from 29% to 90% across different models. Moreover, we leverage these findings to construct targeted demonstration examples and fine-tune data, notably enhancing logical reasoning in models like GPT-4o by up to 5%. To our knowledge, this is the first effort to utilize test case outcomes to effectively refine LLMs' formal reasoning capabilities. We make our code, data, and results publicly available(<https://github.com/yxwan123/LogicAsker>) to facilitate further research and replication of our findings.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

PTD-SQL: Partitioning and Targeted Drilling with LLMs in Text-to-SQL

Ruilin Luo, Liuyan Wang, Binghuai Lin, Zicheng Lin, Yujiu Yang

Large Language Models (LLMs) have emerged as powerful tools for Text-to-SQL tasks, exhibiting remarkable reasoning capabilities. Different from tasks such as math word problem and commonsense reasoning, SQL solutions have a relatively fixed pattern. This facilitates the investigation of whether LLMs can benefit from categorical thinking, mirroring how humans acquire knowledge through inductive reasoning based on comparable examples. In this study, we propose that employing query group partitioning allows LLMs to focus on learning the thought processes specific to a single problem type, consequently enhancing their reasoning abilities across diverse difficulty levels and problem categories. Our experiments reveal that multiple advanced LLMs, when equipped with PTD-SQL, can either surpass or match previous state-of-the-art (SOTA) methods on the Spider and BIRD datasets. Intriguingly, models with varying initial performances have exhibited significant improvements mainly at the boundary of their capabilities after targeted drilling, suggesting a parallel with human progress. Code is available at <https://github.com/rflrbzl/PTD-SQL>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Seeing the Forest through the Trees: Data Leakage from Partial Transformer Gradients

Weijun Li, Qiongkai Xu, Mark Dras

Recent studies have shown that distributed machine learning is vulnerable to gradient inversion attacks, where private training data can be reconstructed by analyzing the gradients of the models shared in training. Previous attacks established that such reconstructions are possible using gradients from all parameters in the entire models. However, we hypothesize that most of the involved modules, or even their sub-modules, are at risk of training data leakage, and we validate such vulnerabilities in various intermediate layers of language models. Our extensive experiments reveal that gradients from a single Transformer layer, or even a single linear component with 0.54% parameters, are susceptible to training data leakage. Additionally, we show that applying differential privacy on gradients during training offers limited protection against the novel vulnerability of data disclosure.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Learning from Natural Language Explanations for Generalizable Entity Matching

Somita Wadhwa, ADIT KRISHNAN, Runhui Wang, Byron C Wallace, Luyang Kong

Entity matching is the task of linking records from different sources that refer to the same real-world entity. Past work has primarily treated entity linking as a standard supervised learning problem. However, supervised entity matching models often do not generalize well to new data, and collecting exhaustively labeled training data is often cost prohibitive. Further, recent efforts have adopted LLMs for this task in few/zero-shot settings, exploiting their general knowledge. But LLMs are prohibitively expensive for performing inference at scale for real-world entity matching tasks. As an efficient alternative, we re-cast entity matching as a conditional generation task as opposed to binary classification. This enables us to "distill" LLM reasoning into smaller entity matching models via natural language explanations. This approach achieves strong performance, especially on out-of-domain generalization tests (10.85% F-1) where stand-alone generative methods struggle. We perform ablations that highlight the importance of explanations, both for performance and model robustness.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

On the Reliability of Psychological Scales on Large Language Models

Jen-tse Huang, Wenxuan Wang, Man HO LAM, Eric John Li, Wenxiang Jiao, Michael Lyu

Recent research has focused on examining Large Language Models (LLMs) characteristics from a psychological standpoint, acknowledging the necessity of understanding their behavioral characteristics. The administration of personality tests to LLMs has emerged as a noteworthy area in this context. However, the suitability of employing psychological scales, initially devised for humans, on LLMs is a matter of ongoing debate. Our study aims to determine the reliability of applying personality assessments to LLMs, explicitly investigating whether LLMs demonstrate consistent personality traits. Analysis of 2,500 settings per model, including GPT-3.5, GPT-4, Gemini-Pro, and LLaMA-3.1, reveals that various LLMs show consistency in responses to the Big Five Inventory, indicating a satisfactory level of reliability. Furthermore, our research explores the potential of GPT-3.5 to emulate diverse personalities and represent various groups capability increasingly sought after in social sciences for substituting human participants with LLMs to reduce costs. Our findings reveal that LLMs have the potential to represent different personalities with specific prompt instructions.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, Meng Jiang

Personalization in large language models (LLMs) is increasingly important, aiming to align the LLMs' interactions, content, and recommendations with individual user preferences. Recent advances have highlighted effective prompt design by enriching user queries with non-parametric knowledge through behavior history retrieval and textual profiles. However, these methods faced limitations due to a lack of model ownership, resulting in constrained customization and privacy issues, and often failed to capture complex, dynamic user behavior patterns. To address these shortcomings, we introduce One PEFT Per User (OPPU), employing personalized parameter-efficient fine-tuning (PEFT) modules to store user-specific behavior patterns and preferences. By plugging in personal PEFT parameters, users can own and use their LLMs individually. OPPU integrates parametric user knowledge in the personal PEFT parameters with non-parametric knowledge from retrieval and profiles, adapting LLMs to user behavior shifts. Experimental results demonstrate that OPPU significantly outperforms existing prompt-based methods across seven diverse tasks in the LaMP benchmark. Further studies reveal OPPU's enhanced capabilities in handling user behavior shifts, modeling users at different activity levels, maintaining robustness across various user history formats, and displaying versatility with different PEFT methods.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MetaReflection: Learning Instructions for Language Agents using Past Reflections

Priyanshu Gupta, Shashank Kirtania, Ananya Singha, Sumit Gulwani, Arjun Radhakrishna, Gustavo Soares, Sherry Shi

The popularity of Large Language Models (LLMs) have unleashed a new age of Language Agents for solving a diverse range of tasks. While contemporary frontier LLMs are capable enough to perform reasonably good Language agents, the closed-API model makes it hard to improve in cases they perform sub-optimally. To address this, recent works have explored techniques to improve their performance using self reflection and prompt optimization techniques. While techniques like self reflection work well in an online setup, contemporary prompt optimization techniques are designed to work on simpler tasks. To address this, we introduce METAREFLECTION, a novel offline reinforcement learning technique that enhances the performance of Language Agents by augmenting a semantic memory based on experiential learnings from past trials. We demonstrate the efficacy of METAREFLECTION by evaluating across multiple domains, including complex logical reasoning, biomedical semantic similarity, open world question answering, and vulnerability threat detection, in Infrastructure-as-Code, with different agent design. METAREFLECTION boosts Language agents performance by 4 % to 16.82 % over the raw GPT-4 baseline and performs on par with existing state-of-the-art prompt optimization techniques while requiring fewer LLM calls.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

More Insightful Feedback for Tutoring: Enhancing Generation Mechanisms and Automatic Evaluation

Wencke Liermann, Jin-Xia Huang, Yohan Lee, Kong Joo Lee

Incorrect student answers can become valuable learning opportunities, provided that the student understands where they went wrong and why. To this end, rather than being given the correct answer, students should receive elaborated feedback on how to correct a mistake on their own. Highlighting the complex demands that the generation of such feedback places on a model's input utilization abilities, we propose two extensions to the training pipeline. Firstly, we employ a KL regularization term between a standard and enriched input format to achieve more targeted input representations. Secondly, we add a preference optimization step to encourage student answer-adaptive feedback generation. The effectiveness of those extensions is underlined by a significant increase in model performance of 3.3 METEOR points. We go beyond traditional surface form-based metrics to assess two important dimensions of feedback quality, i.e., faithfulness and informativeness. Hereby, we are the first to propose an automatic metric measuring the degree to which feedback divulges the correct answer, that we call Informativeness Index I^2 . We verify in how far each metric captures feedback quality.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CodeAgent: Autonomous Communicative Agents for Code Review

Xunzhu Tang, KISUB KIM, Yewei Song, Cedric Lothritz, Bei Li, Saad Ezzini, Haoye Tian, Jacques Klein, Tegawendé F. Bissyandé

Code review, which aims at ensuring the overall quality and reliability of software, is a cornerstone of software development. Unfortunately, while crucial, Code review is a labor-intensive process that the research community is looking to automate. Existing automated methods rely on single input-output generative models and thus generally struggle to emulate the collaborative nature of code review. This work introduces CodeAgent, a novel multi-agent Large Language Model (LLM) system for code review automation. CodeAgent incorporates a supervisory agent, QA-Checker, to ensure that all the agents' contributions address the initial review question. We evaluated CodeAgent on critical code review tasks: (1) detect inconsistencies between code changes and commit messages, (2) identify vulnerability introductions, (3) validate code style adherence, and (4) suggest code revisions. The results demonstrate CodeAgent's effectiveness, contributing to a new state-of-the-art in code review automation. Our data and code are publicly available (<https://github.com/Daniel4SE/codeagent>).

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

DECOR: Improving Coherence in L2 English Writing with a Novel Benchmark for Incoherence Detection, Reasoning, and Rewriting

Xiuming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, Zhou Yu

Coherence in writing, an aspect that L2 English learners often struggle with, is crucial in assessing L2 English writing. Existing automated writing evaluation systems primarily use basic surface linguistic features to detect coherence in writing. However, little effort has been made to correct the detected incoherence, which could significantly benefit L2 language learners seeking to improve their writing. To bridge this gap, we introduce DECOR, a novel benchmark that includes expert annotations for detecting incoherence in L2 English writing, identifying the underlying reasons, and rewriting the incoherent sentences. To our knowledge, DECOR is the first coherence assessment dataset specifically designed for improving L2 English writing, featuring pairs of original incoherent sentences alongside their expert-rewritten counterparts. Additionally, we fine-tuned models to automatically detect and rewrite incoherence in student essays. We find that incorporating specific reasons for incoherence during fine-tuning consistently improves the quality of the rewrites, achieving a level that is favored in both automatic and human evaluations.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Large Language Models Can Self-Correct with Key Condition Verification

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, Meng Jiang

Intrinsic self-correct was a method that instructed large language models (LLMs) to verify and correct their responses without external feedback. Unfortunately, the study concluded that the LLMs could not self-correct reasoning yet. We find that a simple yet effective prompting method enhances LLM performance in identifying and correcting inaccurate answers without external feedback. That is to mask a key condition in the question, add the current response to construct a verification question, and predict the condition to verify the response. The condition can be an entity in an open-domain question or a numerical value in an arithmetic question, which requires minimal effort (via prompting) to identify. We propose an iterative verify-then-correct framework to progressively identify and correct (probably) false responses, named ProCo. We conduct experiments on three reasoning tasks. On average, ProCo, with GPT-3.5-Turbo-1106 as the backend LLM, yields +6.8 exact match on four open-domain question answering datasets, +14.1 accuracy on three arithmetic reasoning datasets, and +9.6 accuracy on a commonsense reasoning dataset, compared to Self-Correct. Our implementation is made publicly available at <https://wzy6642.github.io/proco.github.io>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Large Language Models Are Involuntary Truth-Tellers: Exploiting Fallacy Failure for Jailbreak Attacks

Yue Zhou, Henry Peng Zou, Barbara Di Eugenio, Yang Zhang

We find that language models have difficulties generating fallacious and deceptive reasoning. When asked to generate deceptive outputs, language models tend to leak honest counterparts but believe them to be false. Exploiting this deficiency, we propose a jailbreak attack method that elicits an aligned language model for malicious output. Specifically, we query the model to generate a fallacious yet deceptively real procedure for the harmful behavior. Since a fallacious procedure is generally considered fake and thus harmless by LLMs, it helps bypass the safeguard mechanism. Yet the output is factually harmful since the LLM cannot fabricate fallacious solutions but proposes truthful ones. We evaluate our approach over five safety-aligned large language models, comparing four previous jailbreak methods, and show that our approach achieves competitive performance with more harmful outputs. We believe the findings could be extended beyond model safety, such as self-verification and hallucination.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MolTRES: Improving Chemical Language Representation Learning for Molecular Property Prediction

Jin-Hyung Park, Yeachan Kim, Mingyu Lee, Hyuntae Park, SangKeun Lee

Chemical representation learning has gained increasing interest due to the limited availability of supervised data in fields such as drug and

materials design. This interest particularly extends to chemical language representation learning, which involves pre-training Transformers on SMILES sequences - textual descriptors of molecules. Despite its success in molecular property prediction, current practices often lead to overfitting and limited scalability due to early convergence. In this paper, we introduce a novel chemical language representation learning framework, called MolTRES, to address these issues. MolTRES incorporates generator-discriminator training, allowing the model to learn from more challenging examples that require structural understanding. In addition, we enrich molecular representations by transferring knowledge from scientific literature by integrating external materials embedding. Experimental results show that our model outperforms existing state-of-the-art models on popular molecular property prediction tasks.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

SecCoder: Towards Generalizable and Robust Secure Code Generation

Boyu Zhang, Tianyu Du, Junkai Tong, Xuhong Zhang, Kingsun Chow, Sheng Cheng, Xun Wang, Jianwei Yin

After large models (LMs) have demonstrated widespread acceptance in code-related tasks, their superior generative capacity has greatly promoted the application of the code LM. Nevertheless, the security of the generated code has raised attention to its potential damage. Existing secure code generation methods have limited generalizability to unseen test cases and poor robustness against the attacked model, leading to safety failures in code generation. In this paper, we propose a generalizable and robust secure code generation method SecCoder by using in-context learning (ICL) and the safe demonstration. The dense retriever is also used to select the most helpful demonstration to maximize the improvement of the generated codes security. Experimental results show the superior generalizability of the proposed model SecCoder compared to the current secure code generation method, achieving a significant security improvement of an average of 7.20% on unseen test cases. The results also show the better robustness of SecCoder compared to the current attacked code LM, achieving a significant security improvement of an average of 7.74%. Our analysis indicates that SecCoder enhances the security of LMs in generating code, and it is more generalizable and robust.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

DynaThink: Fast or Slow? A Dynamic Decision-Making Framework for Large Language Models

Jiaobao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, Haizhou Li

Large language models (LLMs) have demonstrated emergent capabilities across diverse reasoning tasks via popular Chains-of-Thought (COT) prompting. However, such a simple and fast COT approach often encounters limitations in dealing with complicated problems, while a thorough method, which considers multiple reasoning pathways and verifies each step carefully, results in slower inference. This paper addresses the challenge of enabling LLMs to autonomously select between fast and slow inference methods, thereby optimizing both efficiency and effectiveness. We introduce a dynamic decision-making framework that categorizes tasks into two distinct pathways: 'Fast,' designated for tasks where the LLM quickly identifies a high-confidence solution, and 'Slow,' allocated for tasks that the LLM perceives as complex and for which it has low confidence in immediate solutions as well as requiring more reasoning paths to verify. Experiments on five popular reasoning benchmarks demonstrated the superiority of the DynaThink over baselines. For example, when we compared it to strong COT with self-consistency baseline on the complicated MATH dataset, DynaThink achieved more than 3% increase in accuracy with lower cost. The code will be made available upon publication.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

RaTEScore: A Metric for Entity-Aware Radiology Text Similarity

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Weidi Xie

This paper introduces a novel, entity-aware metric, termed as Radiological Report (Text) Evaluation (RaTEScore), to assess the quality of medical reports generated by AI models. RaTEScore emphasizes crucial medical entities such as diagnostic outcomes and anatomical details, and is robust against complex medical synonyms and sensitive to negation expressions. Technically, we developed a comprehensive medical NER dataset, RaTE-NER, and trained an NER model specifically for this purpose. This model enables the decomposition of complex radiological reports into constituent medical entities. The metric itself is derived by comparing the similarity of entity embeddings, obtained from a language model, based on their types and relevance to clinical significance. Our evaluations demonstrate that RaTEScore aligns more closely with human preference than existing metrics, validated both on established public benchmarks and our newly proposed RaTE-Eval benchmark.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

ERVQA: A Dataset to Benchmark the Readiness of Large Vision Language Models in Hospital Environments

Souryadip Ray, Kushal Gupta, Soumi Kundu, Dr Payal Arvind Kasat, Somak Aditya, Pawan Goyal

The global shortage of healthcare workers has demanded the development of smart healthcare assistants, which can help monitor and alert healthcare workers when necessary. We examine the healthcare knowledge of existing Large Vision Language Models (LVLMs) via the Visual Question Answering (VQA) task in hospital settings through expert annotated open-ended questions. We introduce the Emergency Room Visual Question Answering (ERVQA) dataset, consisting of <image, question, answer> triplets covering diverse emergency room scenarios, a seminal benchmark for LVLMs. By developing a detailed error taxonomy and analyzing answer trends, we reveal the nuanced nature of the task. We benchmark state-of-the-art open-source and closed LVLMs using traditional and adapted VQA metrics: Entailment Score and CLIPScore Confidence. Analyzing errors across models, we infer trends based on properties like decoder type, model size, and in-context solutions. Our findings suggest the ERVQA dataset presents a highly complex task, highlighting the need for specialized, domain-specific solutions.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

SciAgent: Tool-augmented Language Models for Scientific Reasoning

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuang Wang, Liangming Pan, Yujin Yang, Yixin Cao, Aixin Sun

Scientific reasoning poses an excessive challenge for even the most advanced Large Language Models (LLMs). To make this task more practical and solvable for LLMs, we introduce a new task setting named tool-augmented scientific reasoning. This setting supplements LLMs with scalable toolsets, and shifts the focus from pursuing an omniscient problem solver to a proficient tool-user. To facilitate the research of such setting, we construct a tool-augmented training corpus named MathFunc which encompasses over 30,000 samples and roughly 6,000 tools. Building on MathFunc, we develop SciAgent to retrieve, understand and, if necessary, use tools for scientific problem solving. Additionally, we craft a benchmark, SciToolBench, spanning five scientific domains to evaluate LLMs' abilities with tool assistance. Extensive experiments on SciToolBench confirm the effectiveness of SciAgent. Notably, SciAgent-Llama3-8B surpasses other LLMs with the comparable size by more than 8.0% in absolute accuracy. Furthermore, SciAgent-DeepMath-7B shows much superior performance than ChatGPT.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

IM-BERT: Enhancing Robustness of BERT through the Implicit Euler Method

MiHyeon Kim, Juhyoung Park, YoungBin Kim

Pre-trained Language Models (PLMs) have achieved remarkable performance on diverse NLP tasks through pre-training and fine-tuning. However, fine-tuning the model with a large number of parameters on limited downstream datasets often leads to vulnerability to adversarial attacks, causing overfitting of the model on standard datasets. To address these issues, we propose IM-BERT from the perspective of a dynamic system by conceptualizing a layer of BERT as a solution of Ordinary Differential Equations (ODEs). Under the situation of initial value perturbation, we analyze the numerical stability of two main numerical ODE solvers: *the explicit and implicit Euler approaches.* Based

on these analyses, we introduce a numerically robust IM-connection incorporating BERTs layers. This strategy enhances the robustness of PLMs against adversarial attacks, even in low-resource scenarios, without introducing additional parameters or adversarial training strategies. Experimental results on the adversarial GLUE (AdvGLUE) dataset validate the robustness of IM-BERT under various conditions. Compared to the original BERT, IM-BERT exhibits a performance improvement of approximately 8.3%p on the AdvGLUE dataset. Furthermore, in low-resource scenarios, IM-BERT outperforms BERT by achieving 5.9%p higher accuracy.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

LM2: A Simple Society of Language Models Solves Complex Reasoning

Gurushka Juneja, Subhabrata Dutta, Tammoz Chakraborty

Despite demonstrating emergent reasoning abilities, Large Language Models (LLMs) often lose track of complex, multi-step reasoning. Existing studies show that providing guidance via decomposing the original question into multiple subproblems elicits more robustness in LLM reasoning – a decomposer generates the subproblems, and a solver solves each of these subproblems. However, these techniques fail to accommodate coordination between the decomposer and the solver modules (either in a single model or different specialized ones) – the decomposer does not keep track of the ability of the solver to follow the decomposed reasoning. In this paper, we propose LM2 to address these challenges. LM2 modularizes the decomposition, solution, and verification into three different language models. The decomposer module identifies the key concepts necessary to solve the problem and generates step-by-step subquestions according to the reasoning requirement. The solver model generates the solution to the subproblems that are then checked by the verifier module; depending upon the feedback from the verifier, the reasoning context is constructed using the subproblems and the solutions. These models are trained to coordinate using policy learning. Exhaustive experimentation suggests the superiority of LM2 over existing methods on in- and out-domain reasoning problems, outperforming the best baselines by 8.1% on MATH, 7.71% on JEEBench, and 9.7% on MedQA problems (code available at https://github.com/LCS2-IIITD/Language_Model_Multiplex).

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Symbiotic Working Memory Enhances Language Models for Complex Rule Application

Siyan Wang, zhongyu wei, Yejin Choi, Xiang Ren

Large Language Models (LLMs) have shown remarkable reasoning performance but struggle with multi-step deductive reasoning involving a series of rule application steps, especially when rules are presented non-sequentially. Our preliminary analysis shows that while LLMs excel in single-step rule application, their performance drops significantly in multi-step scenarios due to the challenge in rule grounding. It requires anchoring the applicable rule and supporting facts at each step, amidst multiple input rules, facts, and inferred facts. To address this, we propose augmenting LLMs with external working memory and introduce a neurosymbolic framework for rule application. The memory stores facts and rules in both natural language and symbolic forms, enabling precise tracking. Utilizing this memory, our framework iteratively performs symbolic rule grounding and LLM-based rule implementation. The former matches predicates and variables of symbolic rules and facts to ground applicable rules at each step. Experiments indicate our framework's effectiveness in rule application and its robustness across various steps and settings.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

LLMEdgeRefine: Enhancing Text Clustering with LLM-Based Boundary Point Refinement

Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, Kam-Fai Wong

Text clustering is a fundamental task in natural language processing with numerous applications. However, traditional clustering methods often struggle with domain-specific fine-tuning and the presence of outliers. To address these challenges, we introduce LLMEdgeRefine, an iterative clustering method enhanced by large language models (LLMs), focusing on edge points refinement. LLMEdgeRefine enhances current clustering methods by creating super-points to mitigate outliers and iteratively refining clusters using LLMs for improved semantic coherence. Our method demonstrates superior performance across multiple datasets, outperforming state-of-the-art techniques, and offering robustness, adaptability, and cost-efficiency for diverse text clustering applications.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Recurrent Alignment with Hard Attention for Hierarchical Text Rating

Chenxi Lin, Ren Jiaya, Guoxiu He, Zhaoren Jiang, Haiyan Yu, Xiaomin Zhu

While large language models (LLMs) excel at understanding and generating plain text, they are not tailored to handle hierarchical text structures or directly predict task-specific properties such as text rating. In fact, selectively and repeatedly grasping the hierarchical structure of large-scale text is pivotal for deciphering its essence. To this end, we propose a novel framework for hierarchical text rating utilizing LLMs, which incorporates Recurrent Alignment with Hard Attention (RAHA). Particularly, hard attention mechanism prompts a frozen LLM to selectively focus on pertinent leaf texts associated with the root text and generate symbolic representations of their relationships. Inspired by the gradual stabilization of the Markov Chain, recurrent alignment strategy involves feeding predicted ratings iteratively back into the prompts of another trainable LLM, aligning it to progressively approximate the desired target. Experimental results demonstrate that RAHA outperforms existing state-of-the-art methods on three hierarchical text rating datasets. Theoretical and empirical analysis confirms RAHAs ability to gradually converge towards the underlying target through multiple inferences. Additional experiments on plain text rating datasets verify the effectiveness of this Markov-like alignment. Our data and code can be available in <https://github.com/ECNU-Text-Computing/Markov-LLM>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

DetoxLLM: A Framework for Detoxification with Explanations

Md Tawakal Islam Khondaker, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan

Prior works on detoxification are scattered in the sense that they do not cover all aspects of detoxification needed in a real-world scenario. Notably, prior works restrict the task of developing detoxification models to only a seen subset of platforms, leaving the question of how the models would perform on unseen platforms unexplored. Additionally, these works do not address non-detoxifiability, a phenomenon whereby the toxic text cannot be detoxified without altering the meaning. We propose DetoxLLM, the first comprehensive end-to-end detoxification framework, which attempts to alleviate the aforementioned limitations. We first introduce a cross-platform pseudo-parallel corpus applying multi-step data processing and generation strategies leveraging ChatGPT. We then train a suite of detoxification models with our cross-platform corpus. We show that our detoxification models outperform the SoTA model trained with human-annotated parallel corpus. We further introduce explanation to promote transparency and trustworthiness. DetoxLLM additionally offers a unique paraphrase detector especially dedicated for the detoxification task to tackle the non-detoxifiable cases. Through experimental analysis, we demonstrate the effectiveness of our cross-platform corpus and the robustness of DetoxLLM against adversarial toxicity.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CodeJudge: Evaluating Code Generation with Large Language Models

Weixi Tong, Tianyi Zhang

Large Language Models (LLMs) have shown promising performance in code generation. However, how to reliably evaluate code generated by LLMs remains an unresolved problem. This paper presents CodeJudge, a code evaluation framework that leverages LLMs to evaluate the semantic correctness of generated code without the need for test cases. We investigate different ways to guide the LLM in performing slow

thinking to arrive at an in-depth and reliable evaluation. We experimented with four LLMs as evaluators on four code generation datasets and five programming languages. The results show that CodeJudge significantly outperformed existing methods in most settings. Furthermore, compared with a SOTA GPT-3.5-based code evaluation method, CodeJudge achieved better results even when using a much smaller model, Llama-3-8B-Instruct. Our code and datasets are available on GitHub <https://github.com/VichyTong/CodeJudge>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Task Oriented In-Domain Data Augmentation

Xiao Liang, Xinyu Hu, Simiao Zuo, Yeyun Gong, Qiang Lou, Yi Liu, Shao-Lun Huang, Jian Jiao

Large Language Models (LLMs) have shown superior performance in various applications and fields. To achieve better performance on specialized domains such as law and advertisement, LLMs are often continue pre-trained on in-domain data. However, existing approaches suffer from two major issues. First, in-domain data are scarce compared with general domain-agnostic data. Second, data used for continual pre-training are not task-aware, such that they may not be helpful to downstream applications. We propose TRAIT, a task-oriented in-domain data augmentation framework. Our framework is divided into two parts: in-domain data selection and task-oriented synthetic passage generation. The data selection strategy identifies and selects a large amount of in-domain data from general corpora, and thus significantly enriches domain knowledge in the continual pre-training data. The synthetic passages contain guidance on how to use domain knowledge to answer questions about downstream tasks. By training on such passages, the model aligns with the need of downstream applications. We adapt LLMs to two domains: advertisement and math. On average, TRAIT improves LLM performance by 8% in the advertisement domain and 7.5% in the math domain.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Generalizing Clinical De-identification Models by Privacy-safe Data Augmentation using GPT-4

Woojin Kim, Sungjeon Hahn, Jaejin Lee

De-identification (de-ID) refers to removing the association between a set of identifying data and the data subject. In clinical data management, the de-ID of Protected Health Information (PHI) is critical for patient confidentiality. However, state-of-the-art de-ID models show poor generalization on a new dataset. This is due to the difficulty of retaining training corpora. Additionally, labeling standards and the formats of patient records vary across different institutions. Our study addresses these issues by exploiting GPT-4 for data augmentation through one-shot and zero-shot prompts. Our approach effectively circumvents the problem of PHI leakage, ensuring privacy by redacting PHI before processing. To evaluate the effectiveness of our proposal, we conduct cross-dataset testing. The experimental result demonstrates significant improvements across three types of F1 scores.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers

Ran Xu, Wengi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, Carl Yang

Developing effective biomedical retrieval models is important for excelling at knowledge-intensive biomedical tasks but still challenging due to the lack of sufficient publicly annotated biomedical data and computational resources. We present BMRetriever, a series of dense retrievers for enhancing biomedical retrieval via unsupervised pre-training on large biomedical corpora, followed by instruction fine-tuning on a combination of labeled datasets and synthetic pairs. Experiments on 5 biomedical tasks across 11 datasets verify BMRetriever's efficacy on various biomedical applications. BMRetriever also exhibits strong parameter efficiency, with the 410M variant outperforming baselines up to 11.7 times larger, and the 2B variant matching the performance of models with over 5B parameters. The training data and model checkpoints are released at <https://huggingface.co/BMRetriever> to ensure transparency, reproducibility, and application to new domains.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MedAdapter: Efficient Test-Time Adaptation of Large Language Models Towards Medical Reasoning

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, May Dongmei Wang

Despite their improved capabilities in generation and reasoning, adapting large language models (LLMs) to the biomedical domain remains challenging due to their immense size and privacy concerns. In this study, we propose MedAdapter, a unified post-hoc adapter for test-time adaptation of LLMs towards biomedical applications. Instead of fine-tuning the entire LLM, MedAdapter effectively adapts the original model by fine-tuning only a small BERT-sized adapter to rank candidate solutions generated by LLMs. Experiments on four biomedical tasks across eight datasets demonstrate that MedAdapter effectively adapts both white-box and black-box LLMs in biomedical reasoning, achieving average performance improvements of 18.24% and 10.96%, respectively, without requiring extensive computational resources or sharing data with third parties. MedAdapter also yields enhanced performance when combined with train-time adaptation, highlighting a flexible and complementary solution to existing adaptation methods. Faced with the challenges of balancing model performance, computational resources, and data privacy, MedAdapter provides an efficient, privacy-preserving, cost-effective, and transparent solution for adapting LLMs to the biomedical domain.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Coffee-Gym: An Environment for Evaluating and Improving Natural Language Feedback on Erroneous Code

Hyungjoo Chae, Taeyoon Kwon, Seungjun Moon, Yongho Song, Dongjin Kang, Kai Tzu-iunn Ong, Beong-woo Kwak, Seonghyeon Bae, seung-won hwang, Jinyoung Yeo

This paper presents Coffee-Gym, a comprehensive RL environment for training models that provide feedback on code editing. Coffee-Gym includes two major components: (1) Coffee, a dataset containing humans' code edit traces for coding questions and human-written feedback for editing erroneous code; (2) CoffeeEval, a reward function that faithfully reflects the helpfulness of feedback by assessing the performance of the revised code in unit tests. With them, Coffee-Gym addresses the unavailability of high-quality datasets for training feedback models with RL, and provides more accurate rewards than the SOTA reward model (i.e., GPT-4). By applying Coffee-Gym, we elicit feedback models that outperform baselines in enhancing open-source code LLMs' code editing, making them comparable with closed-source LLMs. We make the dataset and the model checkpoint publicly available in <https://huggingface.co/spaces/Coffee-Gym/Project-Coffee-Gym>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Entity Insertion in Multilingual Linked Corpora: The Case of Wikipedia

Tomas Feith, Akhil Arora, Martin Gerlach, Debjit Paul, Robert West

Links are a fundamental part of information networks, turning isolated pieces of knowledge into a network of information that is much richer than the sum of its parts. However, adding a new link to the network is not trivial: it requires not only the identification of a suitable pair of source and target entities but also the understanding of the content of the source to locate a suitable position for the link in the text. The latter problem has not been addressed effectively, particularly in the absence of text spans in the source that could serve as anchors to insert a link to the target entity. To bridge this gap, we introduce and operationalize the task of entity insertion in information networks. Focusing on the case of Wikipedia, we empirically show that this problem is, both, relevant and challenging for editors. We compile a benchmark dataset in 105 languages and develop a framework for entity insertion called LocEL (Localized Entity Insertion) and its multilingual variant XLocEL. We show that XLocEL outperforms all baseline models (including state-of-the-art prompt-based ranking with LLMs such as GPT-4) and that it can be applied in a zero-shot manner on languages not seen during training with minimal performance drop. These findings are important for

applying entity insertion models in practice, e.g., to support editors in adding links across the more than 300 language versions of Wikipedia.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CodeFort: Robust Training for Code Generation Models

Yuhao Zhang, Shiqi Wang, Haifeng Qian, Zijian Wang, Mingyue Shang, Linbo Liu, Sanjay Krishna Gouda, Baishakhi Ray, Murali Krishna Ramamathan, Xiaofei Ma, Anoop Deoras

Code generation models are not robust to small perturbations, which often lead to incorrect generations and significantly degrade the performance of these models. Although improving the robustness of code generation models is crucial to enhancing user experience in real-world applications, existing research efforts do not address this issue. To fill this gap, we propose CodeFort, a framework to improve the robustness of code generation models, generalizing a large variety of code perturbations to enrich the training data and enabling various robust training strategies, mixing data augmentation, batch augmentation, adversarial logits pairing, and contrastive learning, all carefully designed to support high-throughput training. Extensive evaluations show that we increase the average robust pass rates of baseline CodeGen models from 14.79 to 21.74. We notably decrease the robustness drop rate from 95.02% to 54.95% against code-syntax perturbations.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Exploring Automated Keyword Mnemonics Generation with Large Language Models via Overage-and-Rank

Jaewook Lee, Hunter McNichols, Andrew Lan

In this paper, we study an under-explored area of language and vocabulary learning: keyword mnemonics, a technique for memorizing vocabulary through memorable associations with a target word via a verbal cue. Typically, creating verbal cues requires extensive human effort and is quite time-consuming, necessitating an automated method that is more scalable. We propose a novel overgenerate-and-rank method via prompting large language models (LLMs) to generate verbal cues and then ranking them according to psycholinguistic measures and takeaways from a pilot user study. To assess cue quality, we conduct both an automated evaluation of imageability and coherence, as well as a human evaluation involving English teachers and learners. Results show that LLM-generated mnemonics are comparable to human-generated ones in terms of imageability, coherence, and perceived usefulness, but there remains plenty of room for improvement due to the diversity in background and preference among language learners.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

ESG-Kor: A Korean Dataset for ESG-related Information Extraction and Practical Use Cases

Jaeyoung Lee, Geonyeong Son, Misuk Kim

With the expansion of pre-trained language model usage in recent years, the importance of datasets for performing tasks in specialized domains has significantly increased. Therefore, we have built a Korean dataset called ESG-Kor to automatically extract Environmental, Social, and Governance (ESG) information, which has recently gained importance. ESG-Kor is a dataset consisting of a total of 118,946 sentences that extracted information on each ESG component from Korean companies' sustainability reports and manually labeled it according to objective rules provided by ESG evaluation agencies. To verify the effectiveness and applicability of the ESG-Kor dataset, classification performance was confirmed using several Korean pre-trained language models, and significant performance was obtained. Additionally, by extending the ESG classification model to documents of small and medium enterprises and extracting information based on ESG key issues and in-depth analysis, we demonstrated potential and practical use cases in the ESG field.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Explaining Graph Neural Networks with Large Language Models: A Counterfactual Perspective on Molecule Graphs

Yinhan He, Zaiyi Zheng, Patrick Soga, Yaochen Zhu, Yushun Dong, Jundong Li

In recent years, Graph Neural Networks (GNNs) have become successful in molecular property prediction tasks such as toxicity analysis. However, due to the black-box nature of GNNs, their outputs can be concerning in high-stakes decision-making scenarios, e.g., drug discovery. Facing such an issue, Graph Counterfactual Explanation (GCE) has emerged as a promising approach to improve GNN transparency. However, current GCE methods usually fail to take domain-specific knowledge into consideration, which can result in outputs that are not easily comprehensible by humans. To address this challenge, we propose a novel GCE method, LLM-GCE, to unleash the power of large language models (LLMs) in explaining GNNs for molecular property prediction. Specifically, we utilize an autoencoder to generate the counterfactual graph topology from a set of counterfactual text pairs (CTPs) based on an input graph. Meanwhile, we also incorporate a CTP dynamic feedback module to mitigate LLM hallucination, which provides intermediate feedback derived from the generated counterfactuals as an attempt to give more faithful guidance. Extensive experiments demonstrate the superior performance of LLM-GCE. Our code is released on https://github.com/YinhanHe123/new_LLM4GNNEExplanation.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

HoLLMwood: Unleashing the Creativity of Large Language Models in Screenwriting via Role Playing

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, Tian Feng

Generative AI has demonstrated unprecedented creativity in the field of computer vision, yet such phenomena have not been observed in natural language processing. In particular, large language models (LLMs) can hardly produce written works at the level of human experts due to the extremely high complexity of literature writing. In this paper, we present HoLLMwood, an automated framework for unleashing the creativity of LLMs and exploring their potential in screenwriting, which is a highly demanding task. Mimicking the human creative process, we assign LLMs to different roles involved in the real-world scenario. In addition to the common practice of treating LLMs as *Writer*, we also apply LLMs as *Editor*, who is responsible for providing feedback and revision advice to *Writer*. Besides, to enrich the characters and deepen the plots, we introduce a role-playing mechanism and adopt LLMs as *Actors* that can communicate and interact with each other. Evaluations on automatically generated screenplays show that HoLLMwood substantially outperforms strong baselines in terms of coherence, relevance, interestingness and overall quality.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, Mingyu Lee, Kang-Min Kim

Recent advancements in educational platforms have emphasized the importance of personalized education. Accurately estimating question difficulty based on the ability of the student group is essential for personalized question recommendations. Several studies have focused on predicting question difficulty using student question-solving records or textual information about the questions. However, these approaches require a large amount of student question-solving records and fail to account for the subjective difficulties perceived by different student groups. To address these limitations, we propose the LLaSA framework that utilizes large language models to represent students at various levels. Our proposed method, LLaSA and the zero-shot LLaSA, can estimate question difficulty both with and without students question-solving records. In evaluations on the DBE-KT22 and ASSISTments 2005/2006 benchmarks, the zero-shot LLaSA demonstrated a performance comparable to those of strong baseline models even without any training. When evaluated using the classification method, LLaSA outperformed the baseline models, achieving state-of-the-art performance. In addition, the zero-shot LLaSA showed a high correlation with the regressed IRT curve when compared to question difficulty derived from students question-solving records, highlighting its potential for real-world applications.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Weak-to-Strong Reasoning

Yuqing Yang, Yan Ma, Pengfei Liu

When large language models (LLMs) surpass human capabilities, supervising them effectively becomes difficult. *Weak-to-strong learning*, where a less capable model enhances a stronger one, proves valuable in this context. Yet, the efficacy of this paradigm for complex reasoning tasks is still unexplored. In this paper, we introduce a progressive weak-to-strong reasoning framework that enables the strong model to autonomously refine its training data, maximizing the use of weak signals and unlocking its latent abilities. This framework begins with supervised fine-tuning on a selective small but high-quality dataset, followed by preference optimization on contrastive samples identified by the strong model itself. Experiments on the GSM8K and MATH datasets verify that our method can effectively improve the reasoning capabilities of Llama2-70b using three separate weak models. This work paves the way for a more scalable and sophisticated strategy to enhance AI reasoning powers. All relevant code and resources are available in <https://github.com/GAIR-NLP/weak-to-strong-reasoning>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, Xiang Li

In recent years, the rapid increase in scientific papers has overwhelmed traditional review mechanisms, resulting in varying quality of publications. Although existing methods have explored the capabilities of Large Language Models (LLMs) for automated scientific reviewing, their generated contents are often generic or partial. To address the issues above, we introduce an automated paper reviewing framework SEA. It comprises of three modules: Standardization, Evaluation, and Analysis, which are represented by models SEA-S, SEA-E, and SEA-A, respectively. Initially, SEA-S distills data standardization capabilities of GPT-4 for integrating multiple reviews for a paper. Then, SEA-E utilizes standardized data for fine-tuning, enabling it to generate constructive reviews. Finally, SEA-A introduces a new evaluation metric called mismatch score to assess the consistency between paper contents and reviews. Moreover, we design a self-correction strategy to enhance the consistency. Extensive experimental results on datasets collected from eight venues show that SEA can generate valuable insights for authors to improve their papers.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Light-weight Fine-tuning Method for Defending Adversarial Noise in Pre-trained Medical Vision-Language Models

Xu Han, Linghao Jin, Xuezhe Ma, Xiaofeng Liu

Fine-tuning pre-trained Vision-Language Models (VLMs) has shown remarkable capabilities in medical image and textual depiction synergy. Nevertheless, many pre-training datasets are restricted by patient privacy concerns, potentially containing noise that can adversely affect downstream performance. Moreover, the growing reliance on multi-modal generation exacerbates this issue because of its susceptibility to adversarial attacks. To investigate how VLMs trained on adversarial noisy data perform on downstream medical tasks, we first craft noisy upstream datasets using multi-modal adversarial attacks. Through our comprehensive analysis, we unveil that moderate noise enhances model robustness and transferability, but increasing noise levels negatively impact downstream task performance. To mitigate this issue, we propose rectify adversarial noise (RAN) framework, a recipe designed to effectively defend adversarial attacks and rectify the influence of upstream noise during fine-tuning.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

BSharedRAG: Backbone Shared Retrieval-Augmented Generation for the E-commerce Domain

Kaisi Guan, Qian Cao, Yuchong Sun, Xiting Wang, Ruihua Song

Retrieval-Augmented Generation (RAG) system is important in domains such as e-commerce, which has many long-tail entities and frequently updated information. Most existing works adopt separate modules for retrieval and generation, which may be suboptimal since the retrieval task and the generation task cannot benefit from each other to improve performance. We propose a novel Backbone Shared RAG framework (BSharedRAG). It first uses a domain-specific corpus to continually pre-train a base model as a domain-specific backbone model and then trains two plug-and-play Low-Rank Adaptation (LoRA) modules based on the shared backbone to minimize retrieval and generation losses respectively. Experimental results indicate that our proposed BSharedRAG outperforms baseline models by 5% and 13% in Hit@3 upon two datasets in retrieval evaluation and by 23% in terms of BLEU-3 in generation evaluation. Our codes, models, and dataset are available at <https://bsharedrag.github.io>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Data-driven Coreference-based Ontology Building

Shir Ashury Tahan, Amir David Nissan Cohen, Nadav Cohen, Yoram Louzoun, Yoav Goldberg

While coreference resolution is traditionally used as a component in individual document understanding, in this work we take a more global view and explore what we can learn about a domain from the set of all document-level coreference relations that are present in a large corpus. We derive coreference chains from a corpus of 30 million biomedical abstracts and construct a graph based on the string phrases within these chains, establishing connections between phrases if they co-occur within the same coreference chain. We then use the graph structure and the betweenness centrality measure to distinguish between edges denoting hierarchy, identity and noise, assign directionality to edges denoting hierarchy, and split nodes (strings) that correspond to multiple distinct concepts. The result is a rich, data-driven ontology over concepts in the biomedical domain, parts of which overlaps significantly with human-authored ontologies. We release the coreference chains and resulting ontology under a creative-commons license.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Zero-Shot Fact Verification via Natural Logic and Large Language Models

Marek Strong, Rami Aly, Andreas Vlachos

The recent development of fact verification systems with natural logic has enhanced their explainability by aligning claims with evidence through set-theoretic operators, providing faithful justifications. Despite these advancements, such systems often rely on a large amount of training data annotated with natural logic. To address this issue, we propose a zero-shot method that utilizes the generalization capabilities of instruction-tuned large language models. To comprehensively assess the zero-shot capabilities of our method and other fact verification systems, we evaluate all models on both artificial and real-world claims, including multilingual datasets. We also compare our method against other fact verification systems in two setups. First, in the zero-shot generalization setup, we demonstrate that our approach outperforms other systems that were not specifically trained on natural logic data, achieving an average accuracy improvement of 8.96 points over the best-performing baseline. Second, in the zero-shot transfer setup, we show that current systems trained on natural logic data do not generalize well to other domains, and our method outperforms these systems across all datasets with real-world claims.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Privacy Evaluation Benchmarks for NLP Models

Wei Huang, Yinggui Wang, Cen Chen

By inducing privacy attacks on NLP models, attackers can obtain sensitive information such as training data and model parameters, etc. Although researchers have studied, in-depth, several kinds of attacks in NLP models, they are non-systematic analyses. It lacks a comprehensive understanding of the impact caused by the attacks. For example, we must consider which scenarios can apply to which attacks, what the common factors are that affect the performance of different attacks, the nature of the relationships between different attacks, and the influence of various datasets and models on the effectiveness of the attacks, etc. Therefore, we need a benchmark to holistically assess the privacy risks faced by NLP models. In this paper, we present a privacy attack and defense evaluation benchmark in the field of NLP, which includes the conventional/small models and large language models (LLMs). This benchmark supports a variety of models, datasets, and protocols, along with standardized modules for comprehensive evaluation of attacks and defense strategies. Based on the above framework, we present a study on the association between auxiliary data from different domains and the strength of privacy attacks. And we provide an improved attack method in this scenario with the help of Knowledge Distillation (KD). Furthermore, we propose a chained framework for privacy attacks. Allowing a practitioner to chain multiple attacks to achieve a higher-level attack objective. Based on this, we provide some defense and enhanced attack strategies. The code for reproducing the results can be found at https://anonymous.4open.science/r/nlp_doctor-AF48

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

HoneyComb: A Flexible LLM-Based Agent System for Materials Science

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, Bang Liu

The emergence of specialized large language models (LLMs) has shown promise in addressing complex tasks in materials science. Many LLMs, however, often struggle with the distinct complexities of materials science tasks, such as computational challenges, and rely heavily on outdated implicit knowledge, leading to inaccuracies and hallucinations. To address these challenges, we introduce HoneyComb, the first LLM-based agent system specifically designed for materials science. HoneyComb leverages a reliable, high-quality materials science knowledge base (MatSciKB) and a sophisticated tool hub (ToolHub) tailored specifically for materials science to enhance its reasoning and computational capabilities. MatSciKB is a curated, structured knowledge collection based on reliable literature, while ToolHub employs an Inductive Tool Construction method to generate, decompose, and refine API tools for materials science. Additionally, HoneyComb leverages a retriever module that adaptively selects the appropriate knowledge source or tools for specific tasks, thereby ensuring accuracy and relevance. Our results demonstrate that HoneyComb significantly outperforms baseline models across various tasks in materials science, effectively bridging the gap between current LLM capabilities and the specialized needs of this domain. Furthermore, our adaptable framework can be easily extended to other scientific domains, highlighting its potential for broad applicability in advancing scientific research and applications.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

BookWorm: A Dataset for Character Description and Analysis

Argyrios Papoudakis, Mirella Lapata, Frank Keller

Characters are at the heart of every story, driving the plot and engaging readers. In this study, we explore the understanding of characters in full-length books, which contain complex narratives and numerous interacting characters. We define two tasks: character description, which generates a brief factual profile, and character analysis, which offers an in-depth interpretation, including character development, personality, and social context. We introduce the BookWorm dataset, pairing books from the Gutenberg Project with human-written descriptions and analyses. Using this dataset, we evaluate state-of-the-art long-context models in zero-shot and fine-tuning settings, utilizing both retrieval-based and hierarchical processing for book-length inputs. Our findings show that retrieval-based approaches outperform hierarchical ones in both tasks. Additionally, fine-tuned models using coreference-based retrieval produce the most factual descriptions, as measured by fact- and entailment-based metrics. We hope our dataset, experiments, and analysis will inspire further research in character-based narrative understanding.

Resources and Evaluation 3

Nov 13 (Wed) 10:30-12:00 - Room: Riverfront Hall

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CLEAR: Can Language Models Really Understand Causal Graphs?

Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Chaochao Lu

Causal reasoning is a cornerstone of how humans interpret the world. To model and reason about causality, causal graphs offer a concise yet effective solution. Given the impressive advancements in language models, a crucial question arises: can they really understand causal graphs? To this end, we pioneer an investigation into language models' understanding of causal graphs. Specifically, we develop a framework to define causal graph understanding, by assessing language models' behaviors through four practical criteria derived from diverse disciplines (e.g., philosophy and psychology). We then develop CLEAR, a novel benchmark that defines three complexity levels and encompasses 20 causal graph-based tasks across these levels. Finally, based on our framework and benchmark, we conduct extensive experiments on six leading language models and summarize five empirical findings. Our results indicate that while language models demonstrate a preliminary understanding of causal graphs, significant potential for improvement remains.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Evaluating Diversity in Automatic Poetry Generation

Yanran Chen, Hannes Gröner, Sina ZarieSS, Steffen Eger

Natural Language Generation (NLG), and more generally generative AI, are among the currently most impactful research fields. Creative NLG, such as automatic poetry generation, is a fascinating niche in this area. While most previous research has focused on forms of the Turing test when evaluating automatic poetry generation can humans distinguish between automatic and human generated poetry we evaluate the diversity of automatically generated poetry (with a focus on quatrains), by comparing distributions of generated poetry to distributions of human poetry along structural, lexical, semantic and stylistic dimensions, assessing different model types (word vs. character-level, general purpose LLMs vs. poetry-specific models), including the very recent LLaMA3-8B, and types of fine-tuning (conditioned vs. unconditioned). We find that current automatic poetry systems are considerably underdiverse along multiple dimensions they often do not rhyme sufficiently, are semantically too uniform and even do not match the length distribution of human poetry. Our experiments reveal, however, that style-conditioning and character-level modeling clearly increases diversity across virtually all dimensions we explore. Our identified limitations may serve as the basis for more genuinely diverse future poetry generation models.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

GLAPE: Gold Label-agnostic Prompt Evaluation for Large Language Models

Xuanchang Zhang, Zhuosheng Zhang, hai zhao

Despite the rapid progress of large language models (LLMs), their task performance remains sensitive to prompt design. Recent studies have

explored leveraging the LLM itself as an optimizer to identify optimal prompts that maximize task accuracy. However, when evaluating prompts, such approaches heavily rely on elusive manually annotated gold labels to calculate task accuracy for each candidate prompt, which hinders its generality. To overcome the limitation, this work proposes GLaPE, a gold-label-agnostic prompt evaluation method to alleviate dependence on gold labels. GLaPE is composed of two critical aspects: self-consistency evaluation of a single prompt and mutual-consistency refinement across multiple prompts. Experimental results on 8 widely-recognized reasoning tasks demonstrate that GLaPE can produce more effective prompts, achieving performance comparable to those derived from manually annotated gold labels. Analysis shows that GLaPE provides reliable evaluations aligned with accuracy, even in the absence of gold labels. Code is publicly available at [**Anonymous**](#).

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

RepEval: Effective Text Evaluation with LLM Representation

Shuguai Sheng, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, Chenghu Zhou

The era of Large Language Models (LLMs) raises needs for automatic evaluation metrics, which should be adaptable to various application scenarios while maintaining low cost and effectiveness. Traditional metrics for automatic text evaluation are often tailored to specific scenarios, while LLM-based evaluation metrics are costly, requiring fine-tuning or rely heavily on the generation capabilities of LLMs. Besides, previous LLM-based metrics ignore the fact that, within the space of LLM representations, there exist direction vectors that indicate the estimation of text quality. To this end, we introduce RepEval, a metric that leverages the projection of LLM representations for evaluation. Through simple prompt modifications, RepEval can easily transition to various tasks, requiring only minimal sample pairs for direction vector construction. Results on fourteen datasets across two evaluation tasks demonstrate the high effectiveness of our method, which exhibits a higher correlation with human judgments than previous methods, even in complex evaluation scenarios involving pair-wise selection under nuanced aspects. Our work underscores the richness of information regarding text quality embedded within LLM representations, offering insights for the development of new metrics.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Liar, Liar, Logical Mire: A Benchmark for Suppositional Reasoning in Large Language Models

Philipp Mondorf, Barbara Plank

Knights and knaves problems represent a classic genre of logical puzzles where characters either tell the truth or lie. The objective is to logically deduce each character's identity based on their statements. The challenge arises from the truth-telling or lying behavior, which influences the logical implications of each statement. Solving these puzzles requires not only direct deductions from individual statements, but the ability to assess the truthfulness of statements by reasoning through various hypothetical scenarios. As such, knights and knaves puzzles serve as compelling examples of suppositional reasoning. In this paper, we introduce *TruthQuest*, a benchmark for suppositional reasoning based on the principles of knights and knaves puzzles. Our benchmark presents problems of varying complexity, considering both the number of characters and the types of logical statements involved. Evaluations on *TruthQuest* show that large language models like Llama 3 and Mixtral-8x7B exhibit significant difficulties solving these tasks. A detailed error analysis of the models' output reveals that lower-performing models exhibit a diverse range of reasoning errors, frequently failing to grasp the concept of truth and lies. In comparison, more proficient models primarily struggle with accurately inferring the logical implications of potentially false statements.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MagIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, Jiashi Feng

Large Language Models (LLMs) have significantly advanced natural language processing, demonstrating exceptional reasoning, tool usage, and memory capabilities. As their applications expand into multi-agent environments, there arises a need for a comprehensive evaluation framework that captures LLMs' reasoning, planning, collaboration, and other social abilities. This work introduces a novel competition-based benchmark framework specifically designed to assess LLMs within multi-agent settings, providing quantitative metrics to evaluate their judgment, reasoning, deception, self-awareness, cooperation, coordination, and rationality. We utilize two social deduction games alongside three game-theory scenarios to create diverse environments. Our frame is fortified with the probabilistic graphic modeling (PGM) method, enhancing the LLMs' capabilities in navigating complex social and cognitive dimensions. We evaluate seven LLMs, quantitatively highlighting a significant capability gap of over threefold between the strongest, GPT 0.1, and the weakest, Llama-2-70B. It also confirms that our PGM enhancement boosts the abilities of all selected models by an average of 37%. Our data and code can be found here <https://github.com/-cathy/MaGIC>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Do Text-to-Vis Benchmarks Test Real Use of Visualizations?

Hy Nguyen, Xuefei He, Andrew Reeson, Cecile Paris, Josiah Poon, Jonathan K. Kummerfeld

Large language models are able to generate code for visualisations in response to simple user requests. This is a useful application and an appealing one for NLP research because plots of data provide grounding for language. However, there are relatively few benchmarks, and those that exist may not be representative of what users do in practice. This paper investigates whether benchmarks reflect real-world use through an empirical study comparing benchmark datasets with code from public repositories. Our findings reveal a substantial gap, with evaluations not testing the same distribution of chart types, attributes, and actions as real-world examples. One dataset is representative, but requires extensive modification to become a practical end-to-end benchmark. This shows that new benchmarks are needed to support the development of systems that truly address users' visualisation needs. These observations will guide future data creation, highlighting which features hold genuine significance for users.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhengqiang Gong, Bhuvan Dhingra

The rapid scaling of large language models (LLMs) has raised concerns about the transparency and fair use of the data used in their pretraining. Detecting such content is challenging due to the scale of the data and limited exposure of each instance during training. We propose ReCaLL (Relative Conditional Log-Likelihood), a novel membership inference attack (MIA) to detect LLMs' pretraining data by leveraging their conditional language modeling capabilities. ReCaLL examines the relative change in conditional log-likelihoods when prefixing target data points with non-member context. Our empirical findings show that conditioning member data on non-member prefixes induces a larger decrease in log-likelihood compared to non-member data. We conduct comprehensive experiments and show that ReCaLL achieves state-of-the-art performance on the WikiMIA dataset, even with random and synthetic prefixes, and can be further improved using an ensemble approach. Moreover, we conduct an in-depth analysis of LLMs' behavior with different membership contexts, providing insights into how LLMs leverage membership information for effective inference at both the sequence and token level.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents

Liyang Tang, Philippe Laban, Greg Durrett

Recognizing if LLM output can be grounded in evidence is central to many tasks in NLP: retrieval-augmented generation, summarization,

document-grounded dialogue, and more. Current approaches to this kind of fact-checking are based on verifying each piece of a model generation against potential evidence using an LLM. However, this process can be very computationally expensive, requiring many calls to a model to check a single response. In this work, we show how to build small fact-checking models that have GPT-4-level performance but for 400x lower cost. We do this by constructing synthetic training data with GPT-4, which involves creating realistic yet challenging instances of factual errors via a structured generation procedure. Training on this data teaches models to check each fact in the claim and recognize synthesis of information across sentences. For evaluation, we unify datasets from recent work on fact-checking and grounding LLM generations into a new benchmark, LLM-AggrFact. Our best system MiniCheck-FT5 (770M parameters) outperforms all systems of comparable size and reaches GPT-4 accuracy. We release LLM-AggrFact code for data synthesis, and models.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks?

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, Jonathan Berant

Language agents, built on top of language models (LMs), are systems that can interact with complex environments, such as the open web. In this work, we examine whether such agents can perform realistic and time-consuming tasks on the web, e.g., monitoring real-estate markets or locating relevant nearby businesses. We introduce AssistantBench, a challenging new benchmark consisting of 214 realistic tasks that can be automatically evaluated, covering different scenarios and domains. We find that AssistantBench exposes the limitations of current systems, including language models and retrieval-augmented language models, as no model reaches an accuracy of more than 25 points. While closed-book LMs perform well in terms of accuracy, they exhibit low precision and tend to hallucinate facts. State-of-the-art web agents reach a score of near zero. Additionally, we introduce SeePlanAct (SPA), a new web agent that significantly outperforms previous agents, and an ensemble of SPA and closed-book models reaches the best overall performance. Moreover, we analyze failures of current systems and highlight that open web navigation remains a major challenge.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

APPLS: Evaluating Evaluation Metrics for Plain Language Summarization

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, Lucy Lu Wang

While there has been significant development of models for Plain Language Summarization (PLS), evaluation remains a challenge. PLS lacks a dedicated assessment metric, and the suitability of text generation evaluation metrics is unclear due to the unique transformations involved (e.g., adding background explanations, removing jargon). To address these questions, our study introduces a granular meta-evaluation testbed, APPLS, designed to evaluate metrics for PLS. We identify four PLS criteria from previous work—informativeness, simplification, coherence, and faithfulness—and define a set of perturbations corresponding to these criteria that sensitive metrics should be able to detect. We apply these perturbations to extractive hypotheses for two PLS datasets to form our testbed. Using APPLS, we assess performances of 14 metrics, including automated scores, lexical features, and LLM prompt-based evaluations. Our analysis reveals that while some current metrics show sensitivity to specific criteria, no single method captures all four criteria simultaneously. We therefore recommend a suite of automated metrics be used to capture PLS quality along all relevant criteria. This work contributes the first meta-evaluation testbed for PLS and a comprehensive evaluation of existing metrics.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Ontologically Faithful Generation of Non-Player Character Dialogues

Nathaniel Weir, Ryan Thomas, Randolph d'Amore, Kellie Hill, Benjamin Van Durme, Harsh Jhamtani

We introduce a language generation dataset grounded in a popular video game, KNUUDGE (**KN**owledge Constrained **U**ser-NPC **D**ialogue **G**eneration) requires models to produce trees of dialogue between video game characters that accurately reflect quest and entity specifications stated in natural language. KNUUDGE is constructed from side quest dialogues drawn directly from game data of Obsidian Entertainments’ *The Outer Worlds*, leading to real-world complexities in generation: (1) utterances must remain faithful to the game lore, including character personas and backstories; (2) a dialogue must accurately reveal new quest details to the human player; and (3) dialogues are large trees as opposed to linear chains of utterances. We report results for a set of neural generation models using supervised and in-context learning techniques; we find competent performance but room for future work addressing the challenges of creating realistic, game-quality dialogues.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Belief Revision: The Adaptability of Large Language Models Reasoning

Bryan Wible, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, Pascale Fung

The capability to reason from text is crucial for real-world NLP applications. Real-world scenarios often involve incomplete or evolving data. In response, individuals update their beliefs and understandings accordingly. However, most existing evaluations assume that language models (LMs) operate with consistent information. We introduce Belief-R, a new dataset designed to test LMs’ belief revision ability when presented with new evidence. Inspired by how humans suppress prior inferences, this task assesses LMs within the newly proposed delta reasoning (ΔR) framework. Belief-R features sequences of premises designed to simulate scenarios where additional information could necessitate prior conclusions drawn by LMs. We evaluate ~ 30 LMs across diverse prompting strategies and found that LMs generally struggle to appropriately revise their beliefs in response to new information. Further, models adept at updating often underperformed in scenarios without necessary updates, highlighting a critical trade-off. These insights underscore the importance of improving LMs’ adaptiveness to changing information, a step toward more reliable AI systems.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Speechworthy Instruction-tuned Language Models

Hyundong Justin Cho, Nicolaas Paul Jedema, Leonardo F. R. Ribeiro, Karishma Sharma, Pedro Szekely, Alessandro Moschitti, Ruben Janssen, Jonathan May

Current instruction-tuned language models are exclusively trained with textual preference data and thus may not be aligned to the unique requirements of other modalities, such as speech. To better align language models with the speech domain, we explore i) prompting strategies based on radio-industry best practices and ii) preference learning using a novel speech-based preference data of 20K samples collected by annotators who listen to response pairs. Both human and automatic evaluation show that both prompting and preference learning increase the speech-suitability of popular instruction tuned LLMs. More interestingly, we show that these methods are additive; combining them achieves the best win rates in head-to-head comparison, resulting in responses that are preferred or tied to the base model in 76.2% of comparisons on average. Lastly, we share lexical, syntactical, and qualitative analyses that elicit how our studied methods differ with baselines in generating more speech-suitable responses.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation

Weipeng Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, Jun Liu

Automatically generated questions often suffer from problems such as unclear expression or factual inaccuracies, requiring a reliable and comprehensive evaluation of their quality. Human evaluation is widely used in the field of question generation (QG) and serves as the gold

standard for automatic metrics. However, there is a lack of unified human evaluation criteria, which hampers consistent and reliable evaluations of both QG models and automatic metrics. To address this, we propose ****QGEval****, a multi-dimensional ****Eval**-uation** benchmark for ****Q**-uestion ****G**-eneration****, which evaluates both generated questions and existing automatic metrics across 7 dimensions: fluency, clarity, conciseness, relevance, consistency, answerability, and answer consistency. We demonstrate the appropriateness of these dimensions by examining their correlations and distinctions. Through consistent evaluations of QG models and automatic metrics with QGEval, we find that 1) most QG models perform unsatisfactorily in terms of answerability and answer consistency, and 2) existing metrics fail to align well with human judgments when evaluating generated questions across the 7 dimensions. We expect this work to foster the development of both QG technologies and their evaluation.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents

Qi Zhang, Zhihua Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, Edward Dragut

Scientific information extraction (SciIE) is critical for converting unstructured knowledge from scholarly articles into structured data (entities and relations). Several datasets have been proposed for training and validating SciIE models. However, due to the high complexity and cost of annotating scientific texts, those datasets restrict their annotations to specific parts of paper, such as abstracts, resulting in the loss of diverse entity mentions and relations in context. In this paper, we release a new entity and relation extraction dataset for entities related to datasets, methods, and tasks in scientific articles. Our dataset contains 106 manually annotated full-text scientific publications with over 24k entities and 12k relations. To capture the intricate use and interactions among entities in full texts, our dataset contains a fine-grained tag set for relations. Additionally, we provide an out-of-distribution test set to offer a more realistic evaluation. We conduct comprehensive experiments, including state-of-the-art supervised models and our proposed LLM-based baselines, and highlight the challenges presented by our dataset, encouraging the development of innovative models to further the field of SciIE.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Learning Personalized Alignment for Evaluating Open-ended Text Generation

Danding Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, Yuandong Tian

Recent research has increasingly focused on evaluating large language models' (LLMs) alignment with diverse human values and preferences, particularly for open-ended tasks like story generation. Traditional evaluation metrics rely heavily on lexical similarity with human-written references, often showing poor correlation with human judgments and failing to account for alignment with the diversity of human preferences. To address these challenges, we introduce PerSE, an interpretable evaluation framework designed to assess alignment with specific human preferences. It is tuned to infer specific preferences from an in-context personal profile and evaluate the alignment between the generated content and personal preferences. PerSE enhances interpretability by providing detailed comments and fine-grained scoring, facilitating more personalized content generation. Our 13B LLaMA-2-based PerSE shows a 15.8% increase in Kendall correlation and a 13.7% rise in accuracy with zero-shot reviewers compared to GPT-4. It also outperforms GPT-4 by 46.01% in Kendall correlation on new domains, indicating its transferability.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

More DWUGs: Extending and Evaluating Word Usage Graph Datasets in Multiple Languages

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulze im Walde, Nina Tahmasebi

Word Usage Graphs (WUGs) represent human semantic proximity judgments for pairs of word uses in a weighted graph, which can be clustered to infer word sense clusters from simple pairwise word use judgments, avoiding the need for word sense definitions. SemEval-2020 Task 1 provided the first and to date largest manually annotated, diachronic WUG dataset. In this paper, we check the robustness and correctness of the annotations by continuing the SemEval annotation algorithm for two more rounds and comparing against an established annotation paradigm. Further, we test the reproducibility by resampling a new, smaller set of word uses from the SemEval source corpora and annotating them. Our work contributes to a better understanding of the problems and opportunities of the WUG annotation paradigm and points to future improvements.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Building Resources for Emakhuwa: Machine Translation and News Classification Benchmarks

Felermimo D. M. A. Ali, Henrique Lopes Cardoso, Rui Sousa-Silva

This paper introduces a comprehensive collection of NLP resources for Emakhuwa, Mozambique's most widely spoken language. The resources include the first manually translated news bitext corpus between Portuguese and Emakhuwa, news topic classification datasets, and monolingual data. We detail the process and challenges of acquiring this data and present benchmark results for machine translation and news topic classification tasks. Our evaluation examines the impact of different data types—originally clean text, post-corrected OCR, and back-translated data—and the effects of fine-tuning from pre-trained models, including those focused on African languages. Our benchmarks demonstrate good performance in news topic classification and promising results in machine translation. We fine-tuned multilingual encoder-decoder models using real and synthetic data and evaluated them on our test set and the FLORES evaluation sets. The results highlight the importance of incorporating more data and potential for future improvements. All models, code, and datasets are available in the <https://huggingface.co/LIACC> repository under the CC BY 4.0 license.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation

Tong Chen, Akari Asai, Nilofar Miresghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, Pang Wei Koh

Evaluating the degree of reproduction of copyright-protected content by language models (LMs) is of significant interest to the AI and legal communities. Although both literal and non-literal similarities are considered by courts when assessing the degree of reproduction, prior research has focused only on literal similarities. To bridge this gap, we introduce CopyBench, a benchmark designed to measure both literal and non-literal copying in LM generations. Using copyrighted fiction books as text sources, we provide automatic evaluation protocols to assess literal and non-literal copying, balanced against the model utility in terms of the ability to recall facts from the copyrighted works and generate fluent completions. We find that, although literal copying is relatively rare, two types of non-literal copying—event copying and character copying—occur even in models as small as 7B parameters. Larger models demonstrate significantly more copying, with literal copying rates increasing from 0.2% to 10.5% and non-literal copying from 2.3% to 5.9% when comparing Llama3-8B and 70B models, respectively. We further evaluate the effectiveness of current strategies for mitigating copying and show that (1) training-time alignment can reduce literal copying but may increase non-literal copying, and (2) current inference-time mitigation methods primarily reduce literal but not non-literal copying.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Can LLMs replace Neil deGrasse Tyson? Evaluating the Reliability of LLMs as Science Communicators

Prasoon Bajpai, Niladri Chatterjee, Subhabrata Dutta, Tanmoy Chakraborty

Large Language Models (LLMs) and AI assistants driven by these models are experiencing exponential growth in usage among both ex-

pert and amateur users. In this work, we focus on evaluating the reliability of current LLMs as science communicators. Unlike existing benchmarks, our approach emphasizes assessing these models on scientific question-answering tasks that require a nuanced understanding and awareness of answerability. We introduce a novel dataset, SCiP5-QA, comprising 742 Yes/No queries embedded in complex scientific concepts, along with a benchmarking suite that evaluates LLMs for correctness and consistency across various criteria. We benchmark three proprietary LLMs from the OpenAI GPT family and 13 open-access LLMs from the Meta Llama-2, Llama-3, and Mistral families. While most open-access models significantly underperform compared to GPT-4 Turbo, our experiments identify Llama-3-70B as a strong competitor, often surpassing GPT-4 Turbo in various evaluation aspects. We also find that even the GPT models exhibit a general incompetence in reliably verifying LLM responses. Moreover, we observe an alarming trend where human evaluators are deceived by incorrect responses from GPT-4 Turbo.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Exploring the Compositional Deficiency of Large Language Models in Mathematical Reasoning Through Trap Problems

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, Xuanjing Huang

Human cognition exhibits systematic compositionality, the algebraic ability to generate infinite novel combinations from finite learned components, which is the key to understanding and reasoning about complex logic. In this work, we investigate the compositionality of large language models (LLMs) in mathematical reasoning. Specifically, we construct a new dataset MATHTRAP by introducing carefully designed logical traps into the problem descriptions of MATH and GSM8K. Since problems with logical flaws are quite rare in the real world, these represent “unseen” cases to LLMs. Solving these requires the models to systematically compose (1) the mathematical knowledge involved in the original problems with (2) knowledge related to the introduced traps. Our experiments show that while LLMs possess both components of requisite knowledge, they do not **spontaneously** combine them to handle these novel cases. We explore several methods to mitigate this deficiency, such as natural language prompts, few-shot demonstrations, and fine-tuning. We find that LLMs’ performance can be improved through the above external intervention. Overall, systematic compositionality remains an open challenge for large language models.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

The Odyssey of Commonsense Causality: From Foundational Benchmarks to Cutting-Edge Reasoning

Shaobo Cui, Zhiqing Jin, Bernhard Schölkopf, Boi Faltings

Understanding commonsense causality is a unique mark of intelligence for humans. It helps people understand the principles of the real world better and benefits the decision-making process related to causation. For instance, commonsense causality is crucial in judging whether a defendant’s action causes the plaintiff’s loss in determining legal liability. Despite its significance, a systematic exploration of this topic is notably lacking. Our comprehensive survey bridges this gap by focusing on taxonomies, benchmarks, acquisition methods, qualitative reasoning, and quantitative measurements in commonsense causality, synthesizing insights from over 200 representative articles. Our work aims to provide a systematic overview, update scholars on recent advancements, provide a practical guide for beginners, and highlight promising future research directions in this vital field. A summary of the related literature is available at <https://github.com/cui-shaobo/causality-papers>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

LLM-Evolve: Evaluation for LLMs Evolving Capability on Benchmarks

Jiaxuan You, Mingjie Liu, Shrimai Prabhunoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro

The advancement of large language models (LLMs) has extended their use to dynamic and interactive real-world applications, where models engage continuously with their environment and potentially enhance their performance over time. Most existing LLM benchmarks evaluate LLMs on i.i.d. tasks, overlooking their ability to learn iteratively from past experiences. Our paper bridges this evaluation gap by proposing a novel framework, LLM-Evolve, which extends established benchmarks to sequential problem-solving settings. LLM-Evolve evaluates LLMs over multiple rounds, providing feedback after each round to build a demonstration memory that the models can query in future tasks. We applied LLM-Evolve to the MMLU, GSM8K, and AgentBench benchmarks, testing 8 state-of-the-art open-source and closed-source models. Results show that LLMs can achieve performance improvements of up to 17% by learning from past interactions, with the quality of retrieval algorithms and feedback significantly influencing this capability. These insights advocate for more understanding and benchmarks for LLMs’ performance in evolving interactive scenarios.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

NoiseBench: Benchmarking the Impact of Real Label Noise on Named Entity Recognition

Elena Merdjanova, Ansar Aynetdinov, Alan Akbik

Available training data for named entity recognition (NER) often contains a significant percentage of incorrect labels for entity types and entity boundaries. Such label noise poses challenges for supervised learning and may significantly deteriorate model quality. To address this, prior work proposed various noise-robust learning approaches capable of learning from data with partially incorrect labels. These approaches are typically evaluated using simulated noise where the labels in a clean dataset are automatically corrupted. However, as we show in this paper, this leads to unrealistic noise that is far easier to handle than real noise caused by human error or semi-automatic annotation. To enable the study of the impact of various types of real noise, we introduce NoiseBench, a NER benchmark consisting of clean training data corrupted with 6 types of real noise, including expert errors, crowdsourcing errors, automatic annotation errors and LLM errors. We present an analysis that shows that real noise is significantly more challenging than simulated noise, and show that current state-of-the-art models for noise-robust learning fall far short of their achievable upper bound. We release NoiseBench for both English and German to the research community.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

NeuroTrialNER: An Annotated Corpus for Neurological Diseases and Therapies in Clinical Trial Registries

Simona Emilia Doneva, Tilia Ellendorff, Jean-Philippe Goldman, Amelia Elaine Cannon, Gerold Schneider, Beate Sick, Benjamin Victor Ineichen

Extracting and aggregating information from clinical trial registries could provide invaluable insights into the drug development landscape and advance the treatment of neurologic diseases. However, achieving this at scale is hampered by the volume of available data and the lack of an annotated corpus to assist in the development of automation tools. Thus, we introduce NeuroTrialNER, a new and fully open corpus for named entity recognition (NER). It comprises 1093 clinical trial summaries sourced from ClinicalTrials.gov, annotated for neurological diseases, therapeutic interventions, and control treatments. We describe our data collection process and the corpus in detail. We demonstrate its utility for NER using large language models and achieve a close-to-human performance. By bridging the gap in data resources, we hope to foster the development of text processing tools that help researchers navigate clinical trials data more easily.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Mitigating the Impact of Reference Quality on Evaluation of Summarization Systems with Reference-Free Metrics

Théo Gigant, Camille Guinaudeau, Marc decombis, Frédéric Dufaux

Automatic metrics are used as proxies to evaluate abstractive summarization systems when human annotations are too expensive. To be useful, these metrics should be fine-grained, show a high correlation with human annotations, and ideally be independent of reference quality; however, most standard evaluation metrics for summarization are reference-based, and existing reference-free metrics correlate poorly with

relevance, especially on summaries of longer documents. In this paper, we introduce a reference-free metric that correlates well with human evaluated relevance, while being very cheap to compute. We show that this metric can also be used along reference-based metrics to improve their robustness in low quality reference settings.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs

Mihir Parmar, Hanieh Deilamzadeh, Franck Dernoncourt, Seunghyun Yoon, Ryan A. Rossi, Trung Bui

Extractive summarization plays a pivotal role in natural language processing due to its wide-range applications in summarizing diverse content efficiently, while also being faithful to the original content. Despite significant advancement achieved in extractive summarization by Large Language Models (LLMs), these summaries frequently exhibit incoherence. An important aspect of the coherent summary is its readability for intended users. Although there have been many datasets and benchmarks proposed for creating coherent extractive summaries, none of them currently incorporate user intent to improve coherence in extractive summarization. Motivated by this, we propose a systematically created human-annotated dataset consisting of coherent summaries for five publicly available datasets and natural language user feedback, offering valuable insights into how to improve coherence in extractive summaries. We utilize this dataset for aligning LLMs through supervised fine-tuning with natural language human feedback to enhance the coherence of their generated summaries. Preliminary experiments with Falcon-40B and Llama-2-13B show significant performance improvements (10% Rouge-L) in terms of producing coherent summaries. We further utilize human feedback to benchmark results over instruction-tuned models such as FLAN-T5 which resulted in several interesting findings.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Revealing Personality Traits: A New Benchmark Dataset for Explainable Personality Recognition on Dialogues

Lei Sun, Jinning Zhao, Qin Jin

Personality recognition aims to identify the personality traits implied in user data such as dialogues and social media posts. Current research predominantly treats personality recognition as a classification task, failing to reveal the supporting evidence for the recognized personality. In this paper, we propose a novel task named Explainable Personality Recognition, aiming to reveal the reasoning process as supporting evidence of the personality trait. Inspired by personality theories, personality traits are made up of stable patterns of personality state, where the states are short-term characteristic patterns of thoughts, feelings, and behaviors in a concrete situation at a specific moment in time. We propose an explainable personality recognition framework called Chain-of-Personality-Evidence (CoPE), which involves a reasoning process from specific contexts to short-term personality states to long-term personality traits. Furthermore, based on the CoPE framework, we construct an explainable personality recognition dataset from dialogues, PersonalityEvd. We introduce two explainable personality state recognition and explainable personality trait recognition tasks, which require models to recognize the personality state and trait labels and their corresponding support evidence. Our extensive experiments based on Large Language Models on the two tasks show that revealing personality traits is very challenging and we present some insights for future research. We will release our dataset and source code to facilitate further studies in this direction.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Is This a Bad Table? A Closer Look at the Evaluation of Table Generation from Text

Pritika Ramu, Aparna Garimella, Sambaran Bandyopadhyay

Understanding whether a generated table is of good quality is important to be able to use it in creating or editing documents using automatic methods. In this work, we underline that existing measures for table quality evaluation fail to capture the overall semantics of the tables, and sometimes unfairly penalize good tables and reward bad ones. We propose TabEval, a novel table evaluation strategy that captures table semantics by first breaking down a table into a list of natural language atomic statements and then compares them with ground truth statements using entailment-based measures. To validate our approach, we curate a dataset comprising of text descriptions for 1,250 diverse Wikipedia tables, covering a range of topics and structures, in contrast to the limited scope of existing datasets.⁶ We compare TabEval with existing metrics using unsupervised and supervised text-to-table generation methods, demonstrating its stronger correlation with human judgments of table quality across four datasets.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts

Aditya Sharma, Michael Saxon, William Yang Wang

We present LoCoVQA, a dynamic benchmark generator for evaluating long-context reasoning in vision language models (VLMs). LoCoVQA augments test examples for mathematical reasoning, VQA, and character recognition tasks with increasingly long visual contexts composed of both in-distribution and out-of-distribution distractor images. Across these tasks, a diverse set of VLMs rapidly lose performance as the visual context length grows, often exhibiting a striking logarithmic decay trend. This test assesses how well VLMs can ignore irrelevant information when answering queries—a task that is quite easy for language models (LMs) in the text domain—demonstrating that current state-of-the-art VLMs lack this essential capability for many long-context applications.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Evaluating Language Model Math Reasoning via Grounding in Educational Curricula

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, Kyle Lo

To ensure that math curriculum is grade-appropriate and aligns with critical skills or concepts in accordance with educational standards, pedagogical experts can spend months carefully reviewing published math problems. Drawing inspiration from this process, our work presents a novel angle for evaluating language models' (LMs) mathematical abilities, by investigating whether they can discern skills and concepts enabled by math content. We contribute two datasets: one consisting of 385 fine-grained descriptions of K-12 math skills and concepts, or *standards*, from Achieve the Core (*ATC*), and another of 9.9K math problems labeled with these standards (*MathFish*). We develop two tasks for evaluating LMs' abilities to assess math problems: (1) verifying whether a problem aligns with a given standard, and (2) tagging a problem with all aligned standards. Working with experienced teachers, we find that LMs struggle to tag and verify standards linked to problems, and instead predict labels that are close to ground truth, but differ in subtle ways. We also show that LMs often generate problems that do not fully align with standards described in prompts, suggesting the need for careful scrutiny on use cases involving LMs for generating curricular materials. Finally, we categorize problems in GSM8k using math standards, allowing us to better understand why some problems are more difficult to solve for models than others.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MalayMMLU: A Multitask Benchmark for the Low-Resource Malay Language

Soon Chang Poh, Sze Jue Yang, Jeraelyn Ming Li Tan, Lawrence Leroy Tze Yao Chieng, Jia Xuan Tan, Zhenyu Yu, Foong Chee Mun, Chee Seng Chan

⁶We will release the dataset upon acceptance.

Large Language Models (LLMs) and Large Vision Language Models (LVLMs) exhibit advanced proficiency in language reasoning and comprehension across a wide array of languages. While their performance is notably robust in well-resourced languages, their capabilities in low-resource languages, such as Bahasa Melayu (hereinafter referred to as *Malay*), remain less explored due to a scarcity of dedicated studies and benchmarks. To enhance our understanding of LLMs/LVLMs performance in Malay, we introduce the first multi-task language understanding benchmark specifically for this language, named MalayMMLU. This benchmark comprises 24,213 questions spanning both primary (Year 1-6) and secondary (Form 1-5) education levels in Malaysia, encompassing 5 broad topics that further divided into 22 subjects. We conducted an empirical evaluation of 44 LLMs/LVLMs, assessing their proficiency in both Malay and the nuanced contexts of Malaysian culture using this benchmark. The benchmark and evaluation code are available at <https://github.com/UMxYTL-AI-Labs/MalayMMLU>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Authorship Obfuscation in Multilingual Machine-Generated Text Detection

Dominik Macko, Robert Mori, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, Maria Bielikova

High-quality text generation capability of latest Large Language Models (LLMs) causes concerns about their misuse (e.g., in massive generation/spread of disinformation). Machine-generated text (MGT) detection is important to cope with such threats. However, it is susceptible to authorship obfuscation (AO) methods, such as paraphrasing, which can cause MGTs to evade detection. So far, this was evaluated only in monolingual settings. Thus, the susceptibility of recently proposed multilingual detectors is still unknown. We fill this gap by comprehensively benchmarking the performance of 10 well-known AO methods, attacking 37 MGT detection methods against MGTs in 11 languages (i.e., $10 \times 37 \times 11 = 4,070$ combinations). We also evaluate the effect of data augmentation on adversarial robustness using obfuscated texts. The results indicate that all tested AO methods can cause evasion of automated detection in all tested languages, where homoglyph attacks are especially successful. However, some of the AO methods severely damaged the text, making it no longer readable or easily recognizable by humans (e.g., changed language, weird characters).

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

DyKnow: Dynamically Verifying Time-Sensitive Factual Knowledge in LLMs

Seyed Mahed Mousavi, Simone Alghisi, giuseppe riccardi

LLMs acquire knowledge from massive data snapshots collected at different timestamps. Their knowledge is then commonly evaluated using static benchmarks. However, factual knowledge is generally subject to time-sensitive changes, and static benchmarks cannot address those cases. We present an approach to dynamically evaluate the knowledge in LLMs and their time-sensitivity against Wikidata, a publicly available up-to-date knowledge graph. We evaluate the time-sensitive knowledge in twenty-four private and open-source LLMs, as well as the effectiveness of four editing methods in updating the outdated facts. Our results show that 1) outdatedness is a critical problem across state-of-the-art LLMs; 2) LLMs output inconsistent answers when prompted with slight variations of the question prompt; and 3) the performance of the state-of-the-art knowledge editing algorithms is very limited, as they can not reduce the cases of outdatedness and output inconsistency.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Forecasting Future International Events: A Reliable Dataset for Text-Based Event Modeling

Daehoon Gwak, Junwo Park, Minho Park, ChaeHun Park, Hyunchan Lee, Edward Choi, Jaegul Choo

Predicting future international events from textual information, such as news articles, has tremendous potential for applications in global policy, strategic decision-making, and geopolitics. However, existing datasets available for this task are often limited in quality, hindering the progress of related research. In this paper, we introduce a novel dataset designed to address these limitations by leveraging the advanced reasoning capabilities of large-language models (LLMs). Our dataset features high-quality scoring labels generated through advanced prompt modeling and rigorously validated by domain experts in political science. We showcase the quality and utility of our dataset for real-world event prediction tasks, demonstrating its effectiveness through extensive experiments and analysis. Furthermore, we publicly release our dataset along with the full automation source code for data collection, labeling, and benchmarking, aiming to support and advance research in text-based event prediction.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models

Kedai Chen, Qin Chen, Jie Zhou, He Yishen, Liang He

Though large language models (LLMs) achieve significant success in recent years, the hallucination issue remains a challenge, and numerous benchmarks are proposed for hallucination detection. Nevertheless, some of these benchmarks are not naturally generated by LLMs but are intentionally induced. Also, many merely focus on the factuality hallucination while ignoring the faithfulness hallucination. Additionally, although dialogue pattern is more widely utilized in the era of LLMs, current benchmarks only concentrate on sentence-level and passage-level hallucination. In this study, we propose DiaHalu, the first dedicated dialogue-level hallucination evaluation benchmark for LLMs to our knowledge. Initially, we integrate the collected topics into system prompts and facilitate a dialogue between two LLMs. Subsequently, we manually modify the contents that do not adhere to human language conventions and then have LLMs re-generate, simulating authentic human-machine interaction scenarios. Finally, professional scholars annotate all the samples in the dataset. DiaHalu covers four common multi-turn dialogue domains and five hallucination subtypes, extended from factuality and faithfulness hallucination. Experiments through some well-known LLMs and detection methods on the dataset show that DiaHalu is a challenging benchmark, holding significant value for further research.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

STARD: A Chinese Statute Retrieval Dataset Derived from Real-life Queries by Non-professionals

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, quezibing, Ning Zheng, Yun Liu, Weixing Shen, Yiqun LIU

Statute retrieval aims to find relevant statutory articles for specific queries. This process is the basis of a wide range of legal applications such as legal advice, automated judicial decisions, legal document drafting, etc. Existing statute retrieval benchmarks emphasize formal and professional queries from sources like bar exams and legal case documents, thereby neglecting non-professional queries from the general public, which often lack precise legal terminology and references. To address this gap, we introduce the STAtute Retrieval Dataset (STARD), a Chinese dataset comprising 1,543 query cases collected from real-world legal consultations and 55,348 candidate statutory articles. Unlike existing statute retrieval datasets, which primarily focus on professional legal queries, STARD captures the complexity and diversity of real queries from the general public. Through a comprehensive evaluation of various retrieval baselines, we reveal that existing retrieval approaches all fall short of these real queries issued by non-professional users. The best method only achieves a Recall@100 of 0.907, suggesting the necessity for further exploration and additional research in this area.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Ask the experts: sourcing a high-quality nutrition counseling dataset through Human-AI collaboration

Simone Ballucco, Ehud Reiter, Karen Jia-Hui Li, Rafael Sargsyan, Vivek Kumar, Diego Reforgiato, Daniele Riboni, Ondrej Dusek

Large Language Models (LLMs) are being employed by end-users for various tasks, including sensitive ones such as health counseling, disregarding potential safety concerns. It is thus necessary to understand how adequately LLMs perform in such domains. We conduct a case

study on ChatGPT in nutrition counseling, a popular use-case where the model supports a user with their dietary struggles. We crowd-source real-world diet-related struggles, then work with nutrition experts to generate supportive text using ChatGPT. Finally, experts evaluate the safety and text quality of ChatGPT's output. The result is the HAI-coaching dataset, containing 2.4K crowdsourced dietary struggles and 97K corresponding ChatGPT-generated and expert-annotated supportive texts. We analyse ChatGPT's performance, discovering potentially harmful behaviours, especially for sensitive topics like mental health. Finally, we use HAI-coaching to test open LLMs on various downstream tasks, showing that even the latest models struggle to achieve good performance. HAI-coaching is available at <https://github.com/uccollab/hai-coaching/>

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

"Vorbești Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlateanu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalescu, Traian Rebenda

In recent years, Large Language Models (LLMs) have achieved almost human-like performance on various tasks. While some LLMs have been trained on multilingual data, most of the training data is in English; hence, their performance in English greatly exceeds other languages. To our knowledge, we are the first to collect and translate a large collection of texts, instructions, and benchmarks and train, evaluate, and release open-source LLMs tailored for Romanian. We evaluate our methods on four different categories, including academic benchmarks, MT-Bench (manually translated), and a professionally built historical, cultural, and social benchmark adapted to Romanian. We argue for the usefulness and high performance of RoLLMs by obtaining state-of-the-art results across the board. We publicly release all resources (i.e., data, training and evaluation code, models) with the goal of supporting and encouraging research on Romanian LLMs while concurrently creating a generalizable recipe adequate for other low or less-resourced languages.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks

Jack Gallifant, Shan Chen, Pedro José Ferreira Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, Danielle Bitterman

Medical knowledge is context-dependent and requires consistent reasoning across various natural language expressions of semantically equivalent phrases. This is particularly crucial for drug names, where patients often use brand names like Advil or Tylenol instead of their generic equivalents. To study this, we create a new robustness dataset, **RABBITS**, to evaluate performance differences on medical benchmarks after swapping brand and generic drug names using physician expert annotations. We assess both open-source and API-based LLMs on MedQA and MedMCQA, revealing a consistent performance drop ranging from 1-10%. Furthermore, we identify a potential source of this fragility as the contamination of test data in widely used pre-training datasets.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP

Zonghai Yao, Nandiyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, SUNJAE KWON, Zhichao Yang, hong yu

The advancement in healthcare has shifted focus toward patient-centric approaches, particularly in self-care and patient education, facilitated by access to Electronic Health Records (EHR). However, medical jargon in EHRs poses significant challenges in patient comprehension. To address this, we introduce a new task of automatically generating lay definitions, aiming to simplify complex medical terms into patient-friendly lay language. We first created the README dataset, an extensive collection of over 50,000 unique (medical term, lay definition) pairs and 300,000 mentions, each offering context-aware lay definitions manually annotated by domain experts. We have also engineered a data-centric Human-AI pipeline that synergizes data filtering, augmentation, and selection to improve data quality. We then used README as the training data for models and leveraged a Retrieval-Augmented Generation method to reduce hallucinations and improve the quality of model outputs. Our extensive automatic and human evaluations demonstrate that open-source mobile-friendly models, when fine-tuned with high-quality data, are capable of matching or even surpassing the performance of state-of-the-art closed-source large language models like ChatGPT. This research represents a significant stride in closing the knowledge gap in patient education and advancing patient-centric healthcare solutions.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MVP-Bench: Can Large Visual-Language Models Conduct Multi-level Visual Perception Like Humans?

Guanzhen Li, Yuxi Xie, Min-Yen Kan

Humans perform visual perception at multiple levels, including low-level object recognition and high-level semantic interpretation such as behavior understanding. Subtle differences in low-level details can lead to substantial changes in high-level perception. For example, substituting the shopping bag held by a person with a gun suggests violent behavior, implying criminal or violent activity. Despite significant advancements in various multimodal tasks, Large Visual Language Models (LVLMs) remain unexplored in their capabilities to conduct such multi-level visual perceptions. To investigate the perception gap between LVLMs and humans, we introduce MVP-Bench, the first visual-language benchmark systematically evaluating both low- and high-level visual perception of LVLMs. We construct MVP-Bench across natural and synthetic images to investigate how manipulated content influences model perception. Using MVP-Bench, we diagnose the visual perception of 10 open-source and 2 closed-source LVLMs, showing that high-level perception tasks significantly challenge existing LVLMs. The state-of-the-art GPT-4 only achieves an accuracy of 56% on Yes/No questions, compared with 74% in low-level scenarios. Furthermore, the performance gap between natural and manipulated images indicates that current LVLMs do not generalize in understanding the visual semantics of synthetic images as humans do.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Knowledge-Aware Reasoning over Multimodal Semi-structured Tables

Sivash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Srivastava, Vivek Gupta, Mohit Bansal, Dan Roth

Existing datasets for tabular question answering typically focus exclusively on text within cells. However, real-world data is inherently multimodal, often blending images such as symbols, faces, icons, patterns, and charts with textual content in tables. With the evolution of AI models capable of multimodal reasoning, it is pertinent to assess their efficacy in handling such structured data. This study investigates whether current AI models can perform knowledge-aware reasoning on multimodal structured data. We explore their ability to reason on tables that integrate both images and text, introducing MMTabQA, a new dataset designed for this purpose. Our experiments highlight substantial challenges for current AI models in effectively integrating and interpreting multiple text and image inputs, understanding visual context, and comparing visual content across images. These findings establish our dataset as a robust benchmark for advancing AI's comprehension and capabilities in analyzing multimodal structured data.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

MM-MATH: Advancing Multimodal Math Evaluation with Process Evaluation and Fine-grained Classification

Kai Sun, Yushu Bai, Ji Qi, Lei Hou, Juanzi Li

To advance the evaluation of multimodal math reasoning in large multimodal models (LMMs), this paper introduces a novel benchmark, MM-MATH. MM-MATH consists of 5,929 open-ended middle school math problems with visual contexts, with fine-grained classification across

difficulty, grade level, and knowledge points. Unlike existing benchmarks relying on binary answer comparison, MM-MATH incorporates both outcome and process evaluations. Process evaluation employs LMM-as-a-judge to automatically analyze solution steps, identifying and categorizing errors into specific error types. Extensive evaluation of ten models on MM-MATH reveals significant challenges for existing LMs, highlighting their limited utilization of visual information and struggles with higher-difficulty problems. The best-performing model achieves only 31% accuracy on MM-MATH, compared to 82% for humans. This highlights the challenging nature of our benchmark for existing models and the significant gap between the multimodal reasoning capabilities of current models and humans. Our process evaluation reveals that diagram misinterpretation is the most common error, accounting for more than half of the total error cases, underscoring the need for improved image comprehension in multimodal reasoning.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM

Jielin Qiu, Andrea Madotto, Zhaoqiang Lin, Paul A. Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, Seungwhan Moon

Vision-extended LLMs have made significant strides in Visual Question Answering (VQA). Despite these advancements, VLLMs still encounter substantial difficulties in handling queries involving long-tail entities, with a tendency to produce erroneous or hallucinated responses. In this work, we introduce a novel evaluative benchmark named **SnapNTell**, specifically tailored for entity-centric VQA. This task aims to test the models' capabilities in identifying entities and providing detailed, entity-specific knowledge. We have developed the **SnapNTell Dataset**, distinct from traditional VQA datasets: (1) It encompasses a wide range of categorized entities, each represented by images and explicitly named in the answers; (2) It features QA pairs that require extensive knowledge for accurate responses. The dataset is organized into 22 major categories, containing 7,568 unique entities in total. For each entity, we curated 10 illustrative images and crafted 10 knowledge-intensive QA pairs. To address this novel task, we devised a scalable, efficient, and transparent retrieval-augmented multimodal LLM. Our approach markedly outperforms existing methods on the SnapNTell dataset, achieving a 66.5% improvement in the BELURT score.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

The Base-Rate Effect on LLM Benchmark Performance: Disambiguating Test-Taking Strategies from Benchmark Performance

Kyle Moore, Jesse Roberts, Thao Pham, Oseremhen Ewalejoh, Douglas Fisher

Cloze testing is a common method for measuring the behavior of large language models on a number of benchmark tasks. Using the MMLU dataset, we show that the base-rate probability (BRP) differences across answer tokens are significant and affect task performance i.e. guess A if uncertain. We find that counterfactual prompting does sufficiently mitigate the BRP effect. The BRP effect is found to have a similar effect to test taking strategies employed by humans leading to the conflation of task performance and test-taking ability. We propose the Nvr-X-MMLU task, a variation of MMLU, which helps to disambiguate test-taking ability from task performance and reports the latter.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models

Yuqing Zhou, Ruxiang Tang, Ziyu Yao, Ziwei Zhu

Language models (LMs), despite their advances, often depend on spurious correlations, undermining their accuracy and generalizability. This study addresses the overlooked impact of subtler, more complex shortcuts that compromise model reliability beyond oversimplified shortcuts. We introduce a comprehensive benchmark that categorizes shortcuts into occurrence, style, and concept, aiming to explore the nuanced ways in which these shortcuts influence the performance of LMs. Through extensive experiments across traditional LMs, large language models, and state-of-the-art robust models, our research systematically investigates models' resilience and susceptibilities to sophisticated shortcuts. Our benchmark and code can be found at: <https://github.com/yuqing-zhou/shortcut-learning-in-text-classification>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Characterizing LLM Abstention Behavior in Science QA with Context Perturbations

Bingbing Wen, Bill Howe, Lucy Lu Wang

The correct model response in the face of uncertainty is to abstain from answering a question so as not to mislead the user. In this work, we study the ability of LLMs to abstain from answering context-dependent science questions when provided insufficient or incorrect context. We probe model sensitivity in several settings: removing gold context, replacing gold context with irrelevant context, and providing additional context beyond what is given. In experiments on four QA datasets with six LLMs, we show that performance varies greatly across models, across the type of context provided, and also by question type; in particular, many LLMs seem unable to abstain from answering boolean questions using standard QA prompts. Our analysis also highlights the unexpected impact of abstention performance on QA task accuracy. Counter-intuitively, in some settings, replacing gold context with irrelevant context or adding irrelevant context to gold context can improve abstention performance in a way that results in improvements in task performance. Our results imply that changes are needed in QA dataset design and evaluation to more effectively assess the correctness and downstream impacts of model abstention.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Counter Turing Test (CT^2): Investigating AI-Generated Text Detection for Hindi - Ranking LLMs based on Hindi AI Detectability Index (ADI_{-hi})

Ishan Kavathekar, Anku Rani, Ashmit Chamoli, Ponnurangam Kumaraguru, Amit P. Sheth, Amitava Das

The widespread adoption of Large Language Models (LLMs) and awareness around multilingual LLMs have raised concerns regarding the potential risks and repercussions linked to the misapplication of AI-generated text, necessitating increased vigilance. While these models are primarily trained for English, their extensive training on vast datasets covering almost the entire web, equips them with capabilities to perform well in numerous other languages. AI-Generated Text Detection (AGTD) has emerged as a topic that has already received immediate attention in research, with some initial methods having been proposed, soon followed by the emergence of techniques to bypass detection. In this paper, we report our investigation on AGTD for an indic language Hindi. Our major contributions are in four folds: i) examined 26 LLMs to evaluate their proficiency in generating Hindi text, ii) introducing the AI-generated news article in Hindi (AG_{-hi}) dataset, iii) evaluated the effectiveness of five recently proposed AGTD techniques: ConDA, J-Guard, RADAR, RAIDAR and Intrinsic Dimension Estimation for detecting AI-generated Hindi text, iv) proposed Hindi AI Detectability Index (ADI_{-hi}) which shows a spectrum to understand the evolving landscape of eloquence of AI-generated text in Hindi. The code and dataset is available at https://github.com/ishank31/Counter_Turing_Test

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Holmes Benchmark Linguistic Knowledge in Language Models

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, Iryna Gurevych

We introduce Holmes, a benchmark to assess the linguistic competence of language models (LMs) - their ability to grasp linguistic phenomena. Unlike prior prompting-based evaluations, Holmes assesses the linguistic competence of LMs via their internal representations using classifier-based probing. In doing so, we disentangle specific phenomena (e.g., part-of-speech of words) from other cognitive abilities, like following textual instructions, and meet recent calls to assess LMs' linguistic competence in isolation. Composing Holmes, we review over 250 probing studies and feature more than 200 datasets to assess syntax, morphology, semantics, reasoning, and discourse phenomena. An-

alyzing over 50 LMs reveals that, aligned with known trends, their linguistic competence correlates with model size. However, surprisingly, model architecture and instruction tuning also significantly influence performance, particularly in morphology and syntax. Finally, we propose FlashHolmes, a streamlined version of Holmes designed to lower the high computation load while maintaining high-ranking precision.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Remember This Event That Year? Assessing Temporal Information and Understanding in Large Language Models

Himanshu Benival, Dishant Patel, Kowsik Nandagopan D, Hritik Ladha, Ankit Yadav, Mayank Singh

Large Language Models (LLMs) are increasingly ubiquitous, yet their ability to retain and reason about temporal information remains limited, hindering their application in real-world scenarios where understanding the sequential nature of events is crucial. Our study experiments with 12 state-of-the-art models (ranging from 2B to 70B+ parameters) on a novel numerical-temporal dataset, TempUN, spanning from 10,000 BCE to 2100 CE, to uncover significant temporal retention and comprehension limitations. We propose six metrics to assess three learning paradigms to enhance temporal knowledge acquisition. Our findings reveal that open-source models exhibit knowledge gaps more frequently, suggesting a trade-off between limited knowledge and incorrect responses. Additionally, various fine-tuning approaches significantly improved performance, reducing incorrect outputs and impacting the identification of 'information not available' in the generations. The associated dataset and code are available at the [URL](<https://anonymous.4open.science/r/TempUN-ARR/>).

Speech Processing and Spoken Language Understanding 1

Nov 13 (Wed) 10:30-12:00 - Room: Riverfront Hall

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Speaking in Wavelet Domain: A Simple and Efficient Approach to Speed up Speech Diffusion Model

Xiangyu Zhang, Dayiao Liu, Hexiong Liu, Qiqian Zhang, Hanyu Meng, Leibny Paola García Pérez, EngStong Cheng, Lina Yao

Recently, Denoising Diffusion Probabilistic Models (DDPMs) have attained leading performances across a diverse range of generative tasks. However, in the field of speech synthesis, although DDPMs exhibit impressive performance, their prolonged training duration and substantial inference costs hinder practical deployment. Existing approaches primarily focus on enhancing inference speed, while approaches to accelerate training a key factor in the costs associated with adding or customizing voices often necessitate complex modifications to the model, compromising their universal applicability. To address the aforementioned challenges, we propose an inquiry: is it possible to enhance the training/inference speed and performance of DDPMs by modifying the speech signal itself? In this paper, we double the training and inference speed of Speech DDPMs by simply redirecting the generative target to the wavelet domain. This method not only achieves comparable or superior performance to the original model in speech synthesis tasks but also demonstrates its versatility. By investigating and utilizing different wavelet bases, our approach proves effective not just in speech synthesis, but also in speech enhancement.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Scaling Properties of Speech Language Models

Santiago Cuervo, Ricard Marxer

Speech Language Models (SLMs) aim to learn language from raw audio, without textual resources. Despite significant advances, our current models exhibit weak syntax and semantic abilities. However, if the scaling properties of neural language models hold for the speech modality, these abilities will improve as the amount of compute used for training increases. In this paper, we use models of this scaling behavior to estimate the scale at which our current methods will yield a SLM with the English proficiency of text-based Large Language Models (LLMs). We establish a strong correlation between pre-training loss and downstream syntactic and semantic performance in SLMs and LLMs, which results in predictable scaling of linguistic performance. We show that the linguistic performance of SLMs scales up to three orders of magnitude more slowly than that of text-based LLMs. Additionally, we study the benefits of synthetic data designed to boost semantic understanding and the effects of coarser speech tokenization.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

EmphAssess : a Prosodic Benchmark on Assessing Emphasis Transfer in Speech-to-Speech Models

Maureen de Seyssel, Antony D'Avirro, Adina Williams, Emmanuel Dupoux

We introduce EmphAssess, a prosodic benchmark designed to evaluate the capability of speech-to-speech models to encode and reproduce prosodic emphasis. We apply this to two tasks: speech resynthesis and speech-to-speech translation. In both cases, the benchmark evaluates the ability of the model to encode emphasis in the speech input and accurately reproduce it in the output, potentially across a change of speaker and language. As part of the evaluation pipeline, we introduce EmphaClass, a new model that classifies emphasis at the frame or word level.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Improving Spoken Language Modeling with Phoneme Classification: A Simple Fine-tuning Approach

Maxime Poli, Emmanuel Chemla, Emmanuel Dupoux

Recent progress in Spoken Language Modeling has shown that learning language directly from speech is feasible. Generating speech through a pipeline that operates at the text level typically loses nuances, intonations, and non-verbal vocalizations. Modeling directly from speech opens up the path to more natural and expressive systems. On the other hand, speech-only systems require up to three orders of magnitude more data to catch up with their text-based counterparts in terms of their semantic abilities. We show that fine-tuning speech representation models on phoneme classification leads to more context-invariant representations, and language models trained on these units achieve comparable lexical comprehension to ones trained on hundred times more data.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

On Mitigating Performance Disparities in Multilingual Speech Recognition

Monorama Swain, Anna Katrine van Zee, Anders Søgaard

How far have we come in mitigating performance disparities across genders in multilingual speech recognition? We compare the impact on gender disparity of different fine-tuning algorithms for automated speech recognition across model sizes, languages and gender. We look at both performance-focused and fairness-promoting algorithms. Across languages, we see slightly better performance for female speakers for larger models regardless of the fine-tuning algorithm. The best trade-off between performance and parity is found using adapter fusion. Fairness-promoting fine-tuning algorithms (Group-DRO and Spectral Decoupling) hurt performance compared to adapter fusion with only slightly better performance parity. LoRA increases disparities slightly. Fairness-mitigating fine-tuning techniques led to slightly higher variance in performance across languages, with the exception of adapter fusion.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models

Vyas Raina, Rao Ma, Charles McGhee, Kate Knill, Mark Gales

Recent developments in large speech foundation models like Whisper have led to their widespread use in many automatic speech recognition (ASR) applications. These systems incorporate ‘special tokens’ in their vocabulary, such as <|endoftext|>, to guide their language generation process. However, we demonstrate that these tokens can be exploited by adversarial attacks to manipulate the model’s behavior. We propose a simple yet effective method to learn a universal acoustic realization of Whisper’s <|endoftext|> token, which, when prepended to any speech signal, encourages the model to ignore the speech and only transcribe the special token, effectively ‘muting’ the model. Our experiments demonstrate that the same, universal 0.64-second adversarial audio segment can successfully mute a target Whisper ASR model for over 97% of speech samples. Moreover, we find that this universal adversarial audio segment often transfers to new datasets and tasks. Overall this work demonstrates the vulnerability of Whisper models to ‘muting’ adversarial attacks, where such attacks can pose both risks and potential benefits in real-world settings: for example the attack can be used to bypass speech moderation systems, or conversely the attack can also be used to protect private speech data.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control

Haizhou Chen, Run Chen, Julia Hirschberg

While recent advances in Text-to-Speech (TTS) technology produce natural and expressive speech, they lack the option for users to select emotion and control intensity. We propose EmoKnob, a framework that allows fine-grained emotion control in speech synthesis with few-shot demonstrative samples of arbitrary emotion. Our framework leverages the expressive speaker representation space made possible by recent advances in foundation voice cloning models. Based on the few-shot capability of our emotion control framework, we propose two methods to apply emotion control on emotions described by open-ended text, enabling an intuitive interface for controlling a diverse array of nuanced emotions. To facilitate a more systematic emotional speech synthesis field, we introduce a set of evaluation metrics designed to rigorously assess the faithfulness and recognizability of emotion control frameworks. Through objective and subjective evaluations, we show that our emotion control framework effectively embeds emotions into speech and surpasses emotion expressiveness of commercial TTS services.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

ESC: Efficient Speech Coding with Cross-Scale Residual Vector Quantized Transformers

Yuzhe Gu, Enmao Diao

Neural speech codecs aim to compress input signals into minimal bits while maintaining content quality in a low-latency manner. However, existing neural codecs often trade model complexity for reconstruction performance. These codecs primarily use convolutional blocks for feature transformation, which are not inherently suited for capturing the local redundancies in speech signals. To compensate, they require either adversarial discriminators or a large number of model parameters to enhance audio quality. In response to these challenges, we introduce the Efficient Speech Codec (ESC), a lightweight, parameter-efficient speech codec based on a cross-scale residual vector quantization scheme and transformers. Our model employs mirrored hierarchical window transformer blocks and performs step-wise decoding from coarse-to-fine feature representations. To enhance bitrate efficiency, we propose a novel combination of vector quantization techniques along with a pre-training paradigm. Extensive experiments demonstrate that ESC can achieve high-fidelity speech reconstruction with significantly lower model complexity, making it a promising alternative to existing convolutional audio codecs.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Towards Robust Speech Representation Learning for Thousands of Languages

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiataong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, Shinji Watanabe

Self-supervised learning (SSL) has helped extend speech technologies to more languages by reducing the need for labeled data. However, models are still far from supporting the world’s 7000+ languages. We propose XEUS, a Cross-lingual Encoder for Universal Speech, trained on over 1 million hours of data across 4057 languages, extending the language coverage of SSL models 4-fold. We combine 1 million hours of speech from existing publicly accessible corpora with a newly created corpus of 7400+ hours from 4057 languages, which will be publicly released. To handle the diverse conditions of multilingual speech data, we augment the typical SSL masked prediction approach with a novel dereverberation objective, increasing robustness. We evaluate XEUS on several benchmarks, and show that it consistently outperforms or achieves comparable results to state-of-the-art (SOTA) SSL models across a variety of tasks. XEUS sets a new SOTA on the ML-SUPERB benchmark: it outperforms MMS 1B and w2v-BERT 2.0 v2 by 0.8% and 4.4% respectively, despite having less parameters or pre-training data. Checkpoints, code, and data are found in <https://www.wavlab.org/activities/2024/xeus/>.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Unveiling the Role of Pretraining in Direct Speech Translation

Belen Alastruey, Gerard I. Cállego, Marta R. Costa-jussà

Direct speech-to-text translation systems encounter an important drawback in data scarcity. A common solution consists on pretraining the encoder on automatic speech recognition, hence losing efficiency in the training process. In this study, we compare the training dynamics of a system using a pretrained encoder, the conventional approach, and one trained from scratch. We observe that, throughout the training, the randomly initialized model struggles to incorporate information from the speech inputs for its predictions. Hence, we hypothesize that this issue stems from the difficulty of effectively training an encoder for direct speech translation. While a model trained from scratch needs to learn acoustic and semantic modeling simultaneously, a pretrained one can just focus on the latter. Based on these findings, we propose a subtle change in the decoder cross-attention to integrate source information from earlier steps in training. We show that with this change, the model trained from scratch can achieve comparable performance to the pretrained one, while reducing the training time.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Optimizing Rare Word Accuracy in Direct Speech Translation with a Retrieval-and-Demonstration Approach

Sigui Li, Danni Liu, Jan Niehues

Direct speech translation (ST) models often struggle with rare words. Incorrect translation of these words can have severe consequences, impacting translation quality and user trust. While rare word translation is inherently challenging for neural models due to sparse learning signals, real-world scenarios often allow access to translations of past recordings on similar topics. To leverage these valuable resources, we propose a retrieval-and-demonstration approach to enhance rare word translation accuracy in direct ST models. First, we adapt existing ST models to incorporate retrieved examples for rare word translation, which allows the model to benefit from prepended examples, similar to in-context learning. We then develop a cross-modal (speech-to-speech, speech-to-text, text-to-text) retriever to locate suitable examples. We demonstrate that standard ST models can be effectively adapted to leverage examples for rare word translation, improving rare word translation accuracy over the baseline by 17.6% with gold examples and 8.5% with retrieved examples. Moreover, our speech-to-speech retrieval approach outperforms other modalities and exhibits higher robustness to unseen speakers. Our code is publicly available.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

VHASR: A Multimodal Speech Recognition System With Vision Hotwords

Jiliang Hu, Zuchao Li, Ping Wang, Haojun Ai, Lefei Zhang, hai zhao

The image-based multimodal automatic speech recognition (ASR) model enhances speech recognition performance by incorporating audio-related image. However, some works suggest that introducing image information to model does not help improving ASR performance. In this paper, we propose a novel approach effectively utilizing audio-related image information and set up VHASN, a multimodal speech recognition system that uses vision as hotwords to strengthen the model's speech recognition capability. Our system utilizes a dual-stream architecture, which firstly transcribes the text on the two streams separately, and then combines the outputs. We evaluate the proposed model on four datasets: Flickr8k, ADE20k, COCO, and OpenImages. The experimental results show that VHASN can effectively utilize key information in images to enhance the model's speech recognition ability. Its performance not only surpasses unimodal ASR, but also achieves SOTA among existing image-based multimodal ASR.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

TokenVerse: Unifying Speech and NLP Tasks via Transducer-based ASR

Shashi Kumar, Srikanth Madikeri, Juan Pablo Zuluaga Gomez, Iuliia Thorbecke, Esaú VILLATORO-TELLO, Sergio Burdisso, Petr Motlick, Karthik Pandia D S, Aravind Ganapathiraju

In traditional conversational intelligence from speech, a cascaded pipeline is used, involving tasks such as voice activity detection, diarization, transcription, and subsequent processing with different NLP models for tasks like semantic endpointing and named entity recognition (NER). Our paper introduces TokenVerse, a single Transducer-based model designed to handle multiple tasks. This is achieved by integrating task-specific tokens into the reference text during ASR model training, streamlining the inference and eliminating the need for separate NLP models. In addition to ASR, we conduct experiments on 3 different tasks: speaker change detection, endpointing, and NER. Our experiments on a public and a private dataset show that the proposed method improves ASR by up to 7.7% in relative WER while outperforming the cascaded pipeline approach in individual task performance. Our code is publicly available: <https://github.com/idiap/tokenverse-unifying-speech-nlp>

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents

Beyond Partisan Interfacing. Sykes Shobs LEWIS as Paul Duplex Dialogue Age Bandhay Veluri, Benjamin N Peloquin, Bokai YU, Hongyu Gong, Shyamnath Gollakote

Despite broad interest in modeling spoken dialogue agents, most approaches are inherently "half-duplex" – restricted to turn-based interaction with responses requiring explicit prompting by the user or implicit tracking of interruption or silence events. Human dialogue, by contrast, is "full-duplex" allowing for rich synchronicity in the form of quick and dynamic turn-taking, overlapping speech, and backchanneling. Technically, the challenge of achieving full-duplex dialogue with LLMs lies in modeling synchrony as pre-trained LLMs do not have a sense of "time". To bridge this gap, we propose Synchronous LLMs for full-duplex spoken dialogue modeling. We design a novel mechanism to integrate time information into Llama3-8b so that they run synchronously with the real-world clock. We also introduce a training recipe that uses 21k hours of synthetic spoken dialogue data generated from text dialogue data to create a model that generates meaningful and natural spoken dialogue, with just 2k hours of real-world spoken dialogue data. Synchronous LLMs outperform state-of-the-art in dialogue meaningfulness while maintaining naturalness. Finally, we demonstrate the model's ability to participate in full-duplex dialogue by simulating interactions between two agents trained on different datasets, while considering Internet-scale latencies of up to 240 ms.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Casablanca: Data and Models for Multidialectal Arabic Speech Recognition

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou cheikh tourad, Rahaf Alhamouri, Rwa Assi, Aisha Alraeesi, Hoor Mohamed, Fakhraddin Aljawhy, Abdelrahman Mohamed, Abdellah EL MEKKI, El Moatze Billah Nagoudi, Beneladjy Dlala Mama Saadia, Hamzah Al. Alsaidy, Walid Al-Dhabanyah, Sarai Shamatwi, Yasir ECH-CHAMMAKY, AMAL MAKOUAR, Yousra Berrachedi, Mustafa Jarrar, Shady Shehata, Ismail Berrada, Muhammad Abdul-Majeed

In spite of the recent progress in speech processing, the majority of world languages and dialects remain uncovered. This situation only furthers an already wide technological divide, thereby hindering technological and socioeconomic inclusion. This challenge is largely due to the absence of datasets that can empower diverse speech systems. In this paper, we seek to mitigate this obstacle for a number of Arabic dialects by presenting Casablanca, a large-scale community-driven effort to collect and transcribe a multi-dialectal Arabic dataset. The dataset covers eight dialects: Algerian, Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni, and includes annotations for transcription, gender, dialect, and code-switching. We also develop a number of strong baselines exploiting Casablanca. The project page for Casablanca is accessible at: www.dlnlp.ai/speech/casablanca.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Nov 15 (Wed) 10:30-12:00 Riverfront Hall
MultiVerse: Efficient and Expressive Zero-Shot Multi-Task Text-to-Speech

Taejun Bak, Youngsik Eom, SeungJae Choi, Young-Sun Joo

Text-to-speech (TTS) systems that scale up the amount of training data have achieved significant improvements in zero-shot speech synthesis. However, these systems have certain limitations: they require a large amount of training data, which increases costs, and often overlook prosody similarity. To address these issues, we propose MultiVerse, a zero-shot multi-task TTS system that is able to perform TTS or speech style transfer in zero-shot and cross-lingual conditions. MultiVerse requires much less training data than traditional data-driven approaches. To ensure zero-shot performance even with limited data, we leverage source-filter theory-based disentanglement, utilizing the prompt for modeling filter-related and source-related representations. Additionally, to further enhance prosody similarity, we adopt a prosody modeling approach combining prompt-based autoregressive and non-autoregressive methods. Evaluations demonstrate the remarkable zero-shot multi-task TTS performance of MultiVerse and show that MultiVerse not only achieves zero-shot TTS performance comparable to data-driven TTS systems with much less data, but also significantly outperforms other zero-shot TTS systems trained with the same small amount of data. In particular, our novel prosody modeling technique significantly contributes to MultiVerse's ability to generate speech with high prosody similarity to the given prompts.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Where Visual Speech Meets Language: VSP-LLM Framework for Efficient and Context-Aware Visual Speech Processing

Where Visual Speech Meets Language: VSI LEMENT
Jeonghun Yeo, Seunghee Han, Minsu Kim, Yong Man Ro

In visual speech processing, context modeling capability is one of the most important requirements due to the ambiguous nature of lip movements. For example, homophenes, words that share identical lip movements but produce different sounds, can be distinguished by considering the context. In this paper, we propose a novel framework, namely Visual Speech Processing incorporated with LLMs (VSP-LLM), to maximize the context modeling ability by bringing the overwhelming power of LLMs. Specifically, VSP-LLM is designed to perform multi-tasks of visual speech recognition and translation, where the given instructions control the type of task. The input video is mapped to the input latent space of an LLM by employing a self-supervised visual speech model. Focused on the fact that there is redundant information in input frames, we propose a novel deduplication method that reduces the embedded visual features by employing visual speech units. Through the proposed deduplication and low rank adaptation, VSP-LLM can be trained in a computationally efficient manner. In the translation dataset, the MuAvIC benchmark, we demonstrate that VSP-LLM trained on just 30 hours of labeled data can more effectively translate compared to the recent model trained with 433 hours of data.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Phonetic and Lexical Discovery of Canine Vocalization

Sinong Wang, Xingyuan Li, Chunhao Zhang, Mengyue Wu, Kenny Q. Zhu

This paper attempts to discover communication patterns automatically within dog vocalizations in a data-driven approach, which breaks the barrier previous approaches that rely on human prior knowledge on limited data. We present a self-supervised approach with HuBERT, enabling the accurate classification of phones, and an adaptive grammar induction method that identifies phone sequence patterns that suggest a preliminary vocabulary within dog vocalizations. Our results show that a subset of this vocabulary has substantial causality relations with certain canine activities, suggesting signs of stable semantics associated with these “words”.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Audio-Based Linguistic Feature Extraction for Enhancing Multi-lingual and Low-Resource Text-to-Speech

Yongjae Kim, Yejin Jeon, Gary Lee

The difficulty of acquiring abundant, high-quality data, especially in multi-lingual contexts, has sparked interest in addressing low-resource scenarios. Moreover, current literature rely on fixed expressions from language IDs, which results in the inadequate learning of language representations, and the failure to generate speech in unseen languages. To address these challenges, we propose a novel method that directly extracts linguistic features from audio input while effectively filtering out miscellaneous acoustic information including speaker-specific attributes like timbre. Subjective and objective evaluations affirm the effectiveness of our approach for multi-lingual text-to-speech, and highlight its superiority in low-resource transfer learning for previously unseen language.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Modeling Gender and Dialect Bias in Automatic Speech Recognition

Camille Harris, Chijioke Mgbahurike, Neha Kumar, Diyi Yang

Dialect and gender-based biases have become an area of concern in language-dependent AI systems including around automatic speech recognition (ASR) which processes speech audio into text. These potential biases raise concern for discriminatory outcomes with AI systems depending on demographic—particularly gender discrimination against women, and racial discrimination against minorities with ethnic or cultural English dialects. As such we aim to evaluate the performance of ASR systems across different genders and across dialects of English. Concretely, we take a deep dive of the performance of ASR systems on men and women across four US-based English dialects: Standard American English (SAE), African American Vernacular English (AAVE), Chicano English, and Spanglish. To do this, we construct a labeled dataset of 13 hours of podcast audio, transcribed by speakers of the represented dialects. We then evaluate zero-shot performance of different automatic speech recognition models on our dataset, and further finetune models to better understand how finetuning can impact performance. Our work fills the gap of investigating possible gender disparities within underrepresented dialects.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Textless Speech-to-Speech Translation With Limited Parallel Data

Anuj Diwan, Anirudh Srinivasan, David Harwath, Eunsol Choi

Existing speech-to-speech translation (S2ST) models fall into two camps: they either leverage text as an intermediate step or require hundreds of hours of parallel speech data. Both approaches are incompatible with textless languages or language pairs with limited parallel data. We present PFB, a framework for training textless S2ST models that require just dozens of hours of parallel speech data. We first pretrain a model on large-scale monolingual speech data, finetune it with a small amount of parallel speech data (20-60 hours), and lastly train with an unsupervised backtranslation objective. We train and evaluate our models for English-to-German, German-to-English and Marathi-to-English translation on three different domains (European Parliament, Common Voice, and All India Radio) with single-speaker synthesized speech. Evaluated using the ASR-BLEU metric, our models achieve reasonable performance on all three domains, with some being within 1-2 points of our higher-resourced topline.

Nov 13 (Wed) 10:30-12:00 - Riverfront Hall

Fast Streaming Transducer ASR Prototyping via Knowledge Distillation with Whisper

Julia Thorbecke, Juan Pablo Zuluaga Gomez, Esau VILLATORO-TELLO, Shashi Kumar, Pradeep Rangappa, Sergio Burdisso, Petr Motlicek, Karthik Pandia D S, Aravind Ganapathiraju

The training of automatic speech recognition (ASR) with little to no supervised data remains an open question. In this work, we demonstrate that streaming Transformer-Transducer (TT) models can be trained from scratch in consumer and accessible GPUs in their entirety with pseudo-labeled (PL) speech from foundational speech models (FSM). This allows training a robust ASR model just in one stage and does not require large data and computational budget compared to the two-step scenario with pre-training and fine-tuning. We perform a comprehensive ablation on different aspects of PL-based streaming TT models such as the impact of (1) shallow fusion of n-gram LMs, (2) contextual biasing with named entities, (3) chunk-wise decoding for low-latency streaming applications, and (4) TT overall performance as the function of the FSM size. Our results demonstrate that TT can be trained from scratch without supervised data, even with very noisy PLs. We validate the proposed framework on 6 languages from CommonVoice and propose multiple heuristics to filter out hallucinated PLs.

Session 09 - Nov 13 (Wed) 16:00-17:30

Demo

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Schema-Guided Culture-Aware Complex Event Simulation with Multi-Agent Role-Play

Chi Han, Clare R. Voss, Heng Ji, Jiawei Han, Kartik Natarajan, Khanh Duy Nguyen, Qingyun Wang, Revanth Gangi Reddy, Sha Li, Yi Fung
Complex news events, such as natural disasters and socio-political conflicts, require swift responses from the government and society. Relying on historical events to project the future is insufficient as such events are sparse and do not cover all possible conditions and nuanced situations. Simulation of these complex events can help better prepare and reduce the negative impact. We develop a controllable complex news event simulator guided by both the event schema representing domain knowledge about the scenario and user-provided assumptions representing case-specific conditions. As event dynamics depend on the fine-grained social and cultural context, we further introduce a geo-diverse commonsense and cultural norm-aware knowledge enhancement component. To enhance the coherence of the simulation, apart from the global timeline of events, we take an agent-based approach to simulate the individual character states, plans, and actions. By incorporating

the schema and cultural norms, our generated simulations achieve much higher coherence and appropriateness and are received favorably by participants from a humanitarian assistance organization.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

MIMIR: A Customizable Agent Tuning Platform for Enhanced Scientific Applications

Arman Cohan, Chunyuan Deng, Haoran Wang, Heng Ji, Jiannan Cao, Wangchunshu Zhou, Wengi Shi, Blind Name, Yi Fung, Yilun Zhao, hanminwang, Mark Gerstein

Recently, large language models (LLMs) have evolved into interactive agents, proficient in planning, tool use, and task execution across various tasks. However, without agent-tuning, open-source models like LLaMA2 currently struggle to match the efficiency of larger models such as GPT-4 in scientific applications due to a lack of agent-tuning datasets. In response, we introduce MIMIR, a streamlined platform offering a customizable pipeline that enables users to leverage both private knowledge and publicly available, legally compliant datasets at scale for agent tuning. Additionally, MIMIR supports the generation of general instruction-tuning datasets from the same input. This dual capability ensures LLM agents developed through the platform possess specific agent abilities and general competencies. MIMIR integrates these features into an end-to-end platform, facilitating everything from the uploading of scientific data to one-click agent fine-tuning. MIMIR is publicly released and actively maintained at <https://github.com/gersteinlab/MIMIR>, along with a demo video for a quick start, calling for broader development.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

TinyAgent: Function Calling at the Edge

Coleman Richard Charles Hooper, Gopala Anumanchipalli, Kurt Keutzer, Lutfi Eren Erdogan, Nicholas Lee, Ryan Tabrizi, Sehoon Kim, Siddharth Jha, Suhong Moon, Amir Gholaminejad

Recent large language models (LLMs) have enabled the development of advanced agentic systems that can integrate various tools and APIs to fulfill user queries through function calling. However, the deployment of these LLMs on the edge has not been explored since they typically require cloud-based infrastructure due to their substantial model size and computational demands. To this end, we present TinyAgent, an end-to-end framework for training and deploying task-specific small language model agents capable of function calling for driving agentic systems at the edge. We first show how to enable accurate function calling for open-source models via the LLMCompiler framework. We then systematically curate a high-quality dataset for function calling, which we use to fine-tune two small language models, TinyAgent-1.1B and 7B. For efficient inference, we introduce a novel tool retrieval method to reduce the input prompt length and utilize quantization to further accelerate the inference speed. As a driving application, we demonstrate a local Siri-like system for Apple's MacBook that can execute user commands through text or voice input. Our results show that our models can achieve, and even surpass, the function-calling capabilities of larger models like GPT-4-Turbo, while being fully deployed at the edge. We open-source our [dataset, models, and installable package] (<https://github.com/SqueezeAI/Lab/TinyAgent>) and provide a [demo video] (<https://www.youtube.com/watch?v=OGvaGL9IDpQ>) for our MacBook assistant agent.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ReDel: A Toolkit for LLM-Powered Recursive Multi-Agent Systems

Andrew Zhu, Chris Callison-Burch, Liam Dugan

Recently, there has been increasing interest in using Large Language Models (LLMs) to construct complex multi-agent systems to perform tasks such as compiling literature reviews, drafting consumer reports, and planning vacations. Many tools and libraries exist for helping create such systems, however none support *recursive* multi-agent systems—where the models themselves flexibly decide when to delegate tasks and how to organize their delegation structure. In this work, we introduce ReDel: a toolkit for recursive multi-agent systems that supports custom tool-use, delegation schemes, event-based logging, and interactive replay in an easy-to-use web interface. We show that, using ReDel, we are able to achieve significant performance gains on agentic benchmarks and easily identify potential areas of improvements through the visualization and debugging tools. Our code, documentation, and PyPI package are open-source at <https://github.com/zhudotex/redel>, and free to use under the MIT license.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

MATSA: Multi-Agent Table Structure Attribution

Nedim Lipka, Puneet Mathur, Tong Sun, Alexa F Siu

Large Language Models (LLMs) have significantly advanced QA tasks through in-context learning but often suffer from hallucinations. Attributing supporting evidence grounded in source documents has been explored for unstructured text in the past. However, tabular data present unique challenges for attribution due to ambiguities (e.g., abbreviations, domain-specific terms), complex header hierarchies, and the difficulty in interpreting individual table cells without row and column context. We introduce a new task, Fine-grained Structured Table Attribution (FAST-Tab), to generate row and column-level attributions supporting LLM-generated answers. We present MATSA, a novel LLM-based Multi-Agent system capable of post-hoc Table Structure Attribution to help users visually interpret factual claims derived from tables. MATSA augments tabular entities with descriptive context about structure, metadata, and numerical trends to semantically retrieve relevant rows and columns corresponding to facts in an answer. Additionally, we propose TabCite, a diverse benchmark designed to evaluate the FAST-Tab task on tables with complex layouts sourced from Wikipedia and business PDF documents. Extensive experiments demonstrate that MATSA significantly outperforms SOTA baselines on TabCite, achieving an 8-13% improvement in F1 score. Qualitative user studies show that MATSA helps increase user trust in Generative AI by providing enhanced explainability for LLM-assisted table QA and enables professionals to be more productive by saving time on fact-checking LLM-generated answers.

Dialogue and Interactive Systems 2

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Mitigating Hallucination in Fictional Character Role-Play

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, Julian McAuley

Role-playing has wide-ranging applications in customer support, embodied agents, and computational social science. The influence of parametric world knowledge of large language models (LLMs) often causes role-playing characters to act out of character and to hallucinate about things outside the scope of their knowledge. In this work, we focus on the evaluation and mitigation of hallucination in fictional character role-play. We introduce a dataset with over 2,000 characters and 72,000 interviews, including 18,000 adversarial questions. We propose RoleFact, a role-playing method that mitigates hallucination by modulating the influence of parametric knowledge using a pre-calibrated confidence threshold. Experiments show that the proposed method improves the factual precision of generated responses by 18% for adversarial questions with a 44% reduction in temporal hallucination for time-sensitive interviews. The code and the dataset are available at <https://github.com/NafisSadeq/rolefact.git>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

MSI-Agent: Incorporating Multi-Scale Insight into Embodied Agents for Superior Planning and Decision-Making

Dayuan Fu, Biging Qi, Yihuai Gao, Che Jiang, Guanting Dong, Bowen Zhou

Insight gradually becomes a crucial form of long-term memory for an agent. However, the emergence of irrelevant insight and the lack of general insight can greatly undermine the effectiveness of insight. To solve this problem, in this paper, we introduce **M**ulti-**S**cale **I**nsight Agent (MSI-Agent), an embodied agent designed to improve LLMs' planning and decision-making ability by summarizing and utilizing insight effectively across different scales. MSI achieves this through the experience selector, insight generator, and insight selector. Leveraging a three-part pipeline, MSI can generate task-specific and high-level insight, store it in a database, and then use relevant insight from it to aid in decision-making. Our experiments show that MSI outperforms another insight strategy when planning by GPT3.5. Moreover, We delve into the strategies for selecting seed experience and insight, aiming to provide LLM with more useful and relevant insight for better decision-making. Our observations also indicate that MSI exhibits better robustness when facing domain-shifting scenarios.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Be Helpful but Dont Talk too Much - Enhancing Helpfulness in Conversations through Relevance in Multi-Turn Emotional Support

Li Junlin, Bo Peng, Yu-Yin Hsu, Chu-Ren Huang

For a conversation to help and support, speakers should maintain an "effort-reward" tradeoff. As outlined in the gist of Cognitive Relevance Principle", helpful speakers should optimize the cognitive relevance" through maximizing the cognitive effects" and minimizing the processing effort" imposed on listeners. Although preference learning methods have given rise a boon of studies in pursuit of effect-optimization", none have delved into the critical effort-optimization" to fully cultivate the awareness of optimal relevance" into the cognition of conversation agents. To address this gap, we integrate the "Cognitive Relevance Principle" into emotional support agents in the environment of multi-turn conversation. The results demonstrate a significant and robust improvement against the baseline systems with respect to response quality, human-likeness and supportiveness. This study offers compelling evidence for the effectiveness of the "Relevance Principle" in generating human-like, helpful, and harmless emotional support conversations. The source code will be available at <https://github.com/CN-Eyek/VLESA-ORL.git>

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Bootstrapped Policy Learning for Task-oriented Dialogue through Goal Shaping

Yangyang Zhao, Ben Niu, Mehdi Dastani, Shihua Wang

Reinforcement learning shows promise in optimizing dialogue policies, but addressing the challenge of reward sparsity remains crucial. While curriculum learning offers a practical solution by strategically training policies from simple to complex, it hinges on the assumption of a gradual increase in goal difficulty to ensure a smooth knowledge transition across varied complexities. In complex dialogue environments without intermediate goals, achieving seamless knowledge transitions becomes tricky. This paper proposes a novel Bootstrapped Policy Learning (BPL) framework, which adaptively tailors progressively challenging subgoal curriculum for each complex goal through goal shaping, ensuring a smooth knowledge transition. Goal shaping involves goal decomposition and evolution, decomposing complex goals into subgoals with solvable maximum difficulty and progressively increasing difficulty as the policy improves. Moreover, to enhance BPL's adaptability across various environments, we explore various combinations of goal decomposition and evolution within BPL, and identify two universal curriculum patterns that remain effective across different dialogue environments, independent of specific environmental constraints. By integrating the summarized curriculum patterns, our BPL has exhibited efficacy and versatility across four publicly available datasets with different difficulty levels.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Retrospect: Language Agent Meets Offline Reinforcement Learning Critic

Yufei Xiang, Yiqun Shen, Yeqin Zhang, Nguyen Cam-Tu

Large language models (LLMs) possess extensive knowledge and commonsense reasoning capabilities, making them valuable for creating powerful agents. However, existing LLM agent frameworks have not fully utilized past experiences for improvement. This work introduces a new LLM-based agent framework called Retrospect, which addresses this challenge by analyzing past experiences in depth. Unlike previous approaches, Retrospect does not directly integrate experiences into the LLMs context. Instead, it combines the LLMs action likelihood with action values estimated by a Reinforcement Learning (RL) Critic, which is trained on past experiences through an offline retrospective process. Additionally, Retrospect employs a dynamic action rescorer mechanism that increases the importance of experience-based values for tasks that require more interaction with the environment. We evaluate Retrospect in ScienceWorld, ALFWORLD and Webshop environments, demonstrating its advantages over strong baselines.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Synergizing In-context Learning with Hints for End-to-end Task-oriented Dialog Systems

Vishal Vivek Saley, Rocktin Jyoti Das, Dinesh Raghu, Mausam

End-to-end Task-Oriented Dialog (TOD) systems typically require extensive training datasets to perform well. In contrast, large language model (LLM) based TOD systems can excel even with limited data due to their ability to learn tasks through in-context exemplars. However, these models lack alignment with the style of responses in training data and often generate comprehensive responses, making it difficult for users to grasp the information quickly. In response, we propose SyncTOD that synergizes LLMs with task-specific hints to improve alignment in low-data settings. SyncTOD employs small auxiliary models to provide hints and select exemplars for in-context prompts. With ChatGPT, SyncTOD achieves superior performance compared to LLM-based baselines and SoTA models in low-data settings, while retaining competitive performance in full-data settings.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Reconstruct Your Previous Conversations! Comprehensively Investigating Privacy Leakage Risks in Conversations with GPT Models

Junjie Chu, Zeyang Sha, Michael Backes, Yang Zhang

Significant advancements have recently been made in large language models, represented by GPT models. Users frequently have multi-round private conversations with cloud-hosted GPT models for task optimization. Yet, this operational paradigm introduces additional attack surfaces, particularly in custom GPTs and hijacked chat sessions. In this paper, we introduce a straightforward yet potent Conversation Reconstruction Attack. This attack targets the contents of previous conversations between GPT models and benign users, i.e., the benign users' input contents during their interaction with GPT models. The adversary could induce GPT models to leak such contents by querying them with designed malicious prompts. Our comprehensive examination of privacy risks during the interactions with GPT models under this attack reveals GPT-4's considerable resilience. We present two advanced attacks targeting improved reconstruction of past conversations, demonstrating significant privacy leakage across all models under these advanced techniques. Evaluating various defense mechanisms, we find them ineffective against these attacks. Our findings highlight the ease with which privacy can be compromised in interactions with GPT models, urging the community to safeguard against potential abuses of these models' capabilities.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Middleware for LLMs: Tools Are Instrumental for Language Agents in Complex Environments*Yu Gu, Yiheng Shu, Hao Yu, Xiao Liu, Yuxiao Dong, Jie Tang, Jayanth Srinivas, Hugo Latapie, Yu Su*

The applications of large language models (LLMs) have expanded well beyond the confines of text processing, signaling a new era where LLMs are envisioned as generalist agents capable of operating within complex environments. These environments are often highly expansive, making it impossible for the LLM to process them within its short-term memory. Motivated by recent research on extending the capabilities of LLMs with tools, we seek to investigate the intriguing potential of tools to augment LLMs in handling such complexity by introducing a novel class of tools, termed *middleware*, to aid in the proactive exploration within these massive environments. Such specialized tools can serve as a middleware layer shielding the LLM from environmental complexity. In two representative complex environments—knowledge bases (KBs) and databases—we demonstrate the significant potential of augmenting language agents with tools in complex environments. Notably, equipped with the middleware, GPT-4 achieves ***2.8**X the performance of the best baseline in tasks requiring access to database content and ***2.2**X in KB tasks. Our findings illuminate the path for advancing language agents in real-world applications.

*Nov 13 (Wed) 16:00-17:30 - Riverfront Hall***Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors***Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, Mrimmaya Sachan*

Large language models (LLMs) offer many opportunities to scale high-quality personalized tutoring. A promising approach is to build dialog tutoring models to scaffold students' problem-solving. However, even though existing models perform well in solving reasoning questions, they can struggle to precisely detect student's errors and tailor their feedback to these errors. Inspired by real-world teaching practice where teachers identify student errors and customize their response based on them, we focus on verifying student solutions and show how grounding to such verification improves the overall quality of tutor response generation. We collect a dataset of 1,002 stepwise math reasoning chains with the first error step annotated by teachers. We show empirically that finding the mistake in a student solution is challenging for current models. We propose and evaluate several verifiers for detecting these errors. Using both automatic and human evaluation we show that the student solution verifiers steer the generation model towards highly targeted responses to student error which are more often correct with less hallucinations compared to existing baselines. The benchmark dataset and code will be released openly.

*Nov 13 (Wed) 16:00-17:30 - Riverfront Hall***Zero-shot Cross-domain Dialogue State Tracking via Context-aware Auto-prompting and Instruction-following Contrastive Decoding***Xiaoyu DONG, Yufje Feng, ZEXIN LU, Guangyuan SHI, Xiao-Ming Wu*

Zero-shot cross-domain dialogue state tracking (DST) enables us to manage task-oriented dialogues in new, unseen domains without the cost of collecting in-domain data. Previous studies have implemented slot-based input improvements, such as schema-driven descriptions and question-answering formats, but still suffer from negative transfer for seen slots and inefficient transfer for unseen slots due to the significant source-target domain gap. To address these issues, we introduce a novel framework called Context-aware Auto-prompting and Instruction-following Contrastive Decoding (CAPID). This framework generates dynamic, context-aware slot queries, effectively improving the model's transferability. Our context-aware auto-prompting approach tailors slot queries to the current dialogue context, increasing flexibility and reducing ambiguities. Additionally, an instruction-following contrastive decoding strategy helps reduce errors related to off-topic slots by penalizing deviations from the provided instructions. Extensive experiments on two datasets, with varying model sizes (from 60M to 7B), demonstrate the superior performance of CAPID. The source code is provided for reproducibility.

*Nov 13 (Wed) 16:00-17:30 - Riverfront Hall***Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding***Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, Justine Cassell*

Conversational grounding, vital for building dependable dialog systems, involves ensuring a mutual understanding of shared information. Despite its importance, there has been limited research on this aspect of conversation in recent years, especially after the advent of Large Language Models (LLMs). Previous studies have highlighted the shortcomings of pre-trained language models in conversational grounding. However, most testing for conversational grounding capabilities involves human evaluations that are costly and time-consuming. This has led to a lack of testing across multiple models of varying sizes, a critical need given the rapid rate of new model releases. This gap in research becomes more significant considering recent advances in language models, which have led to new emergent capabilities. In this paper, we aim to evaluate the performance of LLMs in various aspects of conversational grounding and analyze why some models perform better than others. We demonstrate a direct correlation between the size of the pre-training data and conversational grounding abilities, meaning that they have independently acquired a specific form of pragmatic capabilities from larger pre-training datasets. Finally, we propose ways to enhance the capabilities of the models that lag in this aspect.

*Nov 13 (Wed) 16:00-17:30 - Riverfront Hall***"In-DIALOGUES WE LEARN": Towards Personalized Dialogue Without Pre-defined Profiles through In-Dialogue Learning***Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, Rui Yan*

Personalized dialogue systems have gained significant attention in recent years for their ability to generate responses in alignment with different personas. However, most existing approaches rely on pre-defined personal profiles, which are not only time-consuming and labor-intensive to create but also lack flexibility. We propose In-Dialogue Learning (IDL), a fine-tuning framework that enhances the ability of pre-trained large language models to leverage dialogue history to characterize persona for personalized dialogue generation tasks without pre-defined profiles. Our experiments on three datasets demonstrate that IDL brings substantial improvements, with BLEU and ROUGE scores increasing by up to 200% and 247%, respectively. Additionally, the results of human evaluations further validate the efficacy of our proposed method.

*Nov 13 (Wed) 16:00-17:30 - Riverfront Hall***Integrating Argumentation and Hate-Speech-based Techniques for Countering Misinformation***Sougata Saha, Rohini Srihari*

The proliferation of online misinformation presents a significant challenge, requiring scalable strategies for effective mitigation. While detection methods exist, current reactive approaches, like content flagging and banning, are short-term and insufficient. Additionally, advancements like large language models (LLMs) exacerbate the issue by enabling large-scale creation and dissemination of misinformation. Thus, sustainable, scalable solutions that encourage behavior change and broaden perspectives by persuading misinformants against their viewpoints or broadening their perspectives are needed. To this end, we propose persuasive LLM-based dialogue systems to tackle misinformation. However, challenges arise due to the lack of suitable datasets and formal frameworks for generating persuasive responses. Inspired by existing methods for countering online hate speech, we explore adapting counter-hate response strategies for misinformation. Since misinformation and hate speech often coexist despite differing intentions, we develop classifiers to identify and annotate response strategies from hate-speech counter-responses for use in misinformation scenarios. Human evaluations show a 91% agreement on the applicability of these strategies to misinformation. Next, as a scalable counter-misinformation solution, we create an LLM-based argument graph framework that generates persuasive responses, using the strategies as control codes to adjust the style and content. Human evaluations and case studies demonstrate that our framework generates expert-like responses and is 14% more engaging, 21% more natural, and 18% more factual than the best available alternatives.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Beyond the Turn-Based Game: Enabling Real-Time Conversations with Duplex Models

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zhiang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, Zhiyuan Liu

As large language models (LLMs) increasingly permeate daily lives, there is a growing demand for real-time interactions that mirror human conversations. Traditional turn-based chat systems driven by LLMs prevent users from verbally interacting with the system while generating responses. To overcome these limitations, we adapt existing LLMs to *duplex models* so that they can listen to users while generating output and dynamically adjust themselves to provide instant feedback. Specifically, we divide the queries and responses of conversations into several time slices and then adopt a time-division-multiplexing (TDM) encoding-decoding strategy to process these slices pseudo-simultaneously. Furthermore, to make LLMs proficient enough to handle real-time conversations, we build a fine-tuning dataset consisting of alternating time slices of queries and responses and covering typical feedback types in instantaneous interactions. Our experiments show that although the queries and responses of conversations are segmented into incomplete slices for processing, LLMs can preserve their original performance on standard benchmarks with a few fine-tuning steps on our dataset. Automatic and human evaluation indicate that duplex models make user-AI interactions more natural and human-like, and greatly improve user satisfaction compared to vanilla LLMs. Our duplex model and dataset will be released soon.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

RA2FD: Distilling Faithfulness into Efficient Dialogue Systems

Zhiyuan Zhu, Yusheng Liao, Chenxin Xu, Yunfeng Guan, Yanfeng Wang, Yu Wang

Generating faithful and fast responses is crucial in the knowledge-grounded dialogue. Retrieval Augmented Generation (RAG) strategies are effective but are inference inefficient, while previous Retrieval Free Generations (RFG) are more efficient but sacrifice faithfulness. To solve this faithfulness-efficiency trade-off dilemma, we propose a novel retrieval-free model training scheme named Retrieval Augmented to Retrieval Free Distillation (RA2FD) to build a retrieval-free model that achieves higher faithfulness than the previous RFG method while maintaining inference efficiency. The core idea of RA2FD is to use a teacher-student framework to distill the faithfulness capacity of a teacher, which is an oracle RAG model that generates multiple knowledge-infused responses. The student retrieval-free model learns how to generate faithful responses from these teacher labels through sequence-level distillation and contrastive learning. Experiment results show that RA2FD let the faithfulness performance of an RFG model surpass the previous SOTA RFG baseline on three knowledge-grounded dialogue datasets by an average of 33% and even matching an RAG model's performance while significantly improving inference efficiency. Our code is available at <https://github.com/zzysjtuwcl/RA2FD>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

TransferTOD: A Generalizable Chinese Multi-Domain Task-Oriented Dialogue System with Transfer Capabilities

Ming Zhang, Caishuang Huang, Yilong Wu, Shichun Liu, Huiyuan Zheng, Yurui Dong, Yuqiong Shen, Shihuan Dou, Jun Zhao, Junjie Ye, Qi Zhang, Tao Gui, Xuanjing Huang

Task-oriented dialogue (TOD) systems aim to efficiently handle task-oriented conversations, including information collection. How to utilize TOD accurately, efficiently and effectively for information collection has always been a critical and challenging task. Recent studies have demonstrated that Large Language Models (LLMs) excel in dialogue, instruction generation, and reasoning, and can significantly enhance the performance of TOD through fine-tuning. However, current datasets primarily cater to user-led systems and are limited to predefined specific scenarios and slots, thereby necessitating improvements in the proactiveness, diversity, and capabilities of TOD. In this study, we present a detailed multi-domain task-oriented data construction process for conversations, and a Chinese dialogue dataset generated based on this process, **TransferTOD***, which authentically simulates human-computer dialogues in 30 popular life service scenarios. Leveraging this dataset, we trained a model using full-parameter fine-tuning called **TransferTOD-7B***, showcasing notable abilities in slot filling and questioning. Our work has demonstrated its strong generalization capabilities in various downstream scenarios, significantly enhancing both data utilization efficiency and system performance. The data is released in <https://github.com/KongLongGeFDU/TransferTOD>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

DC-Instruct: An Effective Framework for Generative Multi-intent Spoken Language Understanding

In the realm of multi-intent spoken language understanding, recent advancements have leveraged the potential of prompt learning frameworks. However, critical gaps exist in these frameworks: the lack of explicit modeling of dual-task dependencies and the oversight of task-specific semantic differences among utterances. To address these shortcomings, we propose DC-Instruct, a novel generative framework based on Dual-task Inter-dependent Instructions (DII) and Supervised Contrastive Instructions (SCI). Specifically, DII guides large language models (LLMs) to generate labels for one task based on the other task's labels, thereby explicitly capturing dual-task inter-dependencies. Moreover, SCI leverages utterance semantics differences by guiding LLMs to determine whether a pair of utterances share the same or similar labels. This can improve LLMs on extracting and discriminating task-specific semantics, thus enhancing their SLU reasoning abilities. Extensive experiments on public benchmark datasets show that DC-Instruct markedly outperforms current generative models and state-of-the-art methods, demonstrating its effectiveness in enhancing dialogue language understanding and reasoning.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

GDPO: Learning to Align Language Models with Diversity Using GFlowNets

Oh Joon Kwon, Daiki E. Matsunaga, Kee-Eung Kim

A critical component of the current generation of language models is preference alignment, which aims to precisely control the model's behavior to meet human needs and values. The most notable among such methods is Reinforcement Learning with Human Feedback (RLHF) and its offline variant Direct Preference Optimization (DPO), both of which seek to maximize a reward model based on human preferences. In particular, DPO derives reward signals directly from the offline preference data, but in doing so overfits the reward signals and generates suboptimal responses that may contain human biases in the dataset. In this work, we propose a practical application of a diversity-seeking RL algorithm called GFlowNet-DPO (GDPO) in an offline preference alignment setting to curtail such challenges. Empirical results show GDPO can generate far more diverse responses than the baseline methods that are still relatively aligned with human values in dialog generation and summarization tasks.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Unsupervised Extraction of Dialogue Policies from Conversations

Makesh Narasimhan Sreedhar, Traian Rebedea, Christopher Parisien

Dialogue policies play a crucial role in developing task-oriented dialogue systems, yet their development and maintenance are challenging and typically require substantial effort from experts in dialogue modeling. While in many situations, large amounts of conversational data are available for the task at hand, people lack an effective solution able to extract dialogue policies from this data. In this paper, we address this gap by first illustrating how Large Language Models (LLMs) can be instrumental in extracting dialogue policies from datasets, through the conversion of conversations into a unified intermediate representation consisting of canonical forms. We then propose a novel method for generating dialogue policies utilizing a controllable and interpretable graph-based methodology. By combining canonical forms across conversations into a flow network, we find that running graph traversal algorithms helps in extracting dialogue flows. These flows are a

better representation of the underlying interactions than flows extracted by prompting LLMs. Our technique focuses on giving conversation designers greater control, offering a productivity tool to improve the process of developing dialogue policies.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Generative Subgraph Retrieval for Knowledge GraphGrounded Dialog Generation

Jinyoung Park, Minseok Joo, Joo-Kyung Kim, Hyunwoo J. Kim

Knowledge graphgrounded dialog generation requires retrieving a dialog-relevant subgraph from the given knowledge base graph and integrating it with the dialog history. Previous works typically represent the graph using an external encoder, such as graph neural networks, and retrieve relevant triplets based on the similarity between single-vector representations of triplets and the dialog history. However, these external encoders fail to leverage the rich knowledge of pretrained language models, and the retrieval process is also suboptimal due to the information bottleneck caused by the single-vector abstraction of the dialog history. In this work, we propose Dialog generation with Generative Subgraph Retrieval (DialogGSR), which retrieves relevant knowledge subgraphs by directly generating their token sequences on top of language models. For effective generative subgraph retrieval, we introduce two key methods: (i) structure-aware knowledge graph linearization with self-supervised graph-specific tokens and (ii) graph-constrained decoding utilizing graph structural proximity-based entity informativeness scores for valid and relevant generative retrieval. DialogGSR achieves state-of-the-art performance in knowledge graphgrounded dialog generation, as demonstrated on OpenDialKG and KOMODIS datasets.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs

Xuhui Zhou, Zhe Su, Tiwalayo Eisaope, Hyunwoo Kim, Maarten Sap

Recent advances in large language models (LLM) have enabled richer social simulations, allowing for the study of various social phenomena. However, most recent work has used a more omniscient perspective on these simulations (e.g., single LLM to generate all interlocutors), which is fundamentally at odds with the non-omniscient, information asymmetric interactions that involve humans and AI agents in the real world. To examine these differences, we develop an evaluation framework to simulate social interactions with LLMs in various settings (omniscient, non-omniscient). Our experiments show that LLMs perform better in unrealistic, omniscient simulation settings but struggle in ones that more accurately reflect real-world conditions with information asymmetry. Moreover, we illustrate the limitations inherent in learning from omniscient simulations. Our findings indicate that addressing information asymmetry remains a fundamental challenge for LLM-based agents.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ABLE: Personalized Disability Support with Politeness and Empathy Integration

Kshitij Mishra, Manisha Burja, Asif Ekbal

In today's dynamic world, providing inclusive and personalized support for individuals with physical disabilities is imperative. With diverse needs and preferences, tailored assistance according to user personas is crucial. In this paper, we introduce ABLE (Adaptive, Bespoke, Listen and Empathetic), a Conversational Support System for Physical Disabilities. By tracking user personas, including gender, age, and personality traits based on the OCEAN model, ABLE ensures that support interactions are uniquely tailored to each user's characteristics and preferences. Moreover, integrating politeness and empathy levels in responses enhances user satisfaction and engagement, fostering a supportive and respectful environment. The development of ABLE involves compiling a comprehensive conversational dataset enriched with user profile annotations. Leveraging reinforcement learning techniques and diverse reward mechanisms, ABLE trains a model to generate responses aligned with individual user profiles while maintaining appropriate levels of politeness and empathy. Based on rigorous empirical analysis encompassing automatic and human evaluation metrics based on persona-consistency, politeness accuracy, empathy accuracy, perplexity, and conversation coherence, the efficacy of ABLE is assessed. Our findings underscore ABLE's success in delivering tailored support to individuals grappling with physical disabilities. To the best of our knowledge, this is the very first attempt towards building a user's persona-oriented physical disability support system.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

A Fairness-Driven Method for Learning Human-Compatible Negotiation Strategies

Ryan Shea, Zhou Yu

Despite recent advancements in AI and NLP, negotiation remains a difficult domain for AI agents. Traditional game theoretic approaches that have worked well for two-player zero-sum games struggle in the context of negotiation due to their inability to learn human-compatible strategies. On the other hand, approaches that only use human data tend to be domain-specific and lack the theoretical guarantees provided by strategies grounded in game theory. Motivated by the notion of fairness as a criterion for optimality in general sum games, we propose a negotiation framework called FDHC which incorporates fairness into both the reward design and search to learn human-compatible negotiation strategies. Our method includes a novel, RL+search technique called LGM-Zero which leverages a pre-trained language model to retrieve human-compatible offers from large action spaces. Our results show that our method is able to achieve more egalitarian negotiation outcomes and improve negotiation quality.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

From Pixels to Persons: Investigating and Modeling Self-Anthropomorphism in Human-Robot Dialogues

Yu Li, Devamanyu Hazarika, Di Jin, Julia Hirschberg, Yang Liu

Self-anthropomorphism in robots manifests itself through their display of human-like characteristics in dialogue, such as expressing preferences and emotions. Our study systematically analyzes self-anthropomorphic expression within various dialogue datasets, outlining the contrasts between self-anthropomorphic and non-self-anthropomorphic responses in dialogue systems. We show significant differences in these two types of responses and propose transitioning from one type to the other. We also introduce Pix2Persona, a novel dataset aimed at developing ethical and engaging AI systems in various embodiments. This dataset preserves the original dialogues from existing corpora and enhances them with paired responses: self-anthropomorphic and non-self-anthropomorphic for each original bot response. Our work not only uncovers a new category of bot responses that were previously under-explored but also lays the groundwork for future studies about dynamically adjusting self-anthropomorphism levels in AI systems to align with ethical standards and user expectations.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Diverse and Effective Synthetic Data Generation for Adaptable Zero-Shot Dialogue State Tracking

James D. Finch, Jinho D. Choi

We demonstrate substantial performance gains in zero-shot dialogue state tracking (DST) by enhancing training data diversity through synthetic data generation. Existing DST datasets are severely limited in the number of application domains and slot types they cover due to the high costs of data collection, restricting their adaptability to new domains. This work addresses this challenge with a novel, fully automatic data generation approach that creates synthetic zero-shot DST datasets. Distinguished from previous methods, our approach can generate dialogues across a massive range of application domains, complete with silver-standard dialogue state annotations and slot descriptions. This technique is used to create the DOT dataset for training zero-shot DST models, encompassing an unprecedented 1,000+ domains. Experiments on the MultiWOZ benchmark show that training models on diverse synthetic data improves Joint Goal Accuracy by 6.7%, achieving results competitive with models 13.5 times larger than ours.

Posters and Demos

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Preference Tuning For Toxicity Mitigation Generalizes Across Languages

Xiaochen Li, Zheng Xin Yong, Stephen Bach

Detoxifying multilingual Large Language Models (LLMs) has become crucial due to their increasing global use. In this work, we explore zero-shot cross-lingual generalization of preference tuning in detoxifying LLMs. Unlike previous studies that show limited cross-lingual generalization for other safety tasks, we demonstrate that Direct Preference Optimization (DPO) training with only English data can significantly reduce toxicity in multilingual open-ended generations. For example, the probability of mGPT-1.3B generating toxic continuations drops from 46.8% to 3.9% across 17 different languages after training. Our results also extend to other multilingual LLMs, such as BLOOM, Llama3, and Aya-23. Using mechanistic interpretability tools like causal intervention and activation analysis, we identified the dual multilingual property of MLP layers in LLMs, which explains the cross-lingual generalization of DPO. Finally, we show that bilingual sentence retrieval can predict the cross-lingual transferability of DPO preference tuning.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Active Listening: Personalized Question Generation in Open-Domain Social Conversation with User Model Based Prompting

Kevin Bowden, Yue Fan, Winson Chen, Wen Cui, Davan Harrison, Marilyn Walker, Xin Eric Wang

Large language models (LLMs) capable of casual conversation have recently become widely available. We hypothesize that users of conversational systems want a more personalized experience, and existing work shows that users are highly receptive to personalized questions (PQs). Question Generation tasks, however, focus on factual questions from textual excerpts. To create a PQ generator, we first identify over 400 real user interests by anonymously aggregating 39K user models. We then populate prompt templates with these 400 interests and use an LLM to generate PQs customized to user interests. The result is PerQs, a novel corpus of 19K question/answer pairs. We evaluate PerQs at scale in the unique context of the Alexa Prize. Our results show significant positive effects on perceived conversation quality. We then fine-tune, deploy, and evaluate PerQy, a neural model that generates PQs in real-time. When evaluated against several competitive LLM baselines, PerQy produced the most natural and engaging responses.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

UrbanLLM: Autonomous Urban Activity Planning and Management with Large Language Models

YUE JIANG, Qin Chao, Yile Chen, Xuicheng Li, SHUAI LIU, Gao Cong

Location-based services play an critical role in improving the quality of our daily lives. Despite the proliferation of numerous specialized AI models within spatio-temporal context of location-based services, these models struggle to autonomously tackle problems regarding complex urban planing and management. To bridge this gap, we introduce UrbanLLM, a fine-tuned large language model (LLM) designed to tackle diverse problems in urban scenarios. UrbanLLM functions as a problem-solver by decomposing urban-related queries into manageable sub-tasks, identifying suitable spatio-temporal AI models for each sub-task, and generating comprehensive responses to the given queries. Our experimental results indicate that UrbanLLM significantly outperforms other established LLMs, such as Llama and the GPT series, in handling problems concerning complex urban activity planning and management. UrbanLLM exhibits considerable potential in enhancing the effectiveness of solving problems in urban scenarios, reducing the workload and reliance for human experts.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Multi-dimensional Evaluation of Empathetic Dialogue Responses

Zhihao Xu, Jiepu Jiang

Empathy is critical for effective and satisfactory conversational communication. Prior efforts to measure conversational empathy mostly focus on expressed communicative intent; that is, the way empathy is expressed. Yet, these works ignore the fact that conversation is also a collaboration involving both speakers and listeners. In contrast, we propose a multi-dimensional empathy evaluation framework to measure both expressed intents from the speakers perspective and perceived empathy from the listeners perspective. We apply our analytical framework to examine internal customer-service dialogues. We find the two dimensions (expressed intent types and perceived empathy) are interconnected, while perceived empathy has high correlations with dialogue satisfaction levels. To reduce the annotation cost, we explore different options to automatically measure conversational empathy: prompting LLMs and training language model-based classifiers. Our experiments show that prompting methods with even popular models like GPT-4 and Flan family models perform relatively poorly on both public and our internal datasets. In contrast, instruction-finetuned classifiers based on FlanT5 family models outperform prior works and competitive baselines. We conduct a detailed ablation study to give more insights into instruction finetuning methods strong performance.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Learning to Ask Informative Questions: Enhancing LLMs with Preference Optimization and Expected Information Gain

Davidide Muzzacara, Alberto Testoni, Raffaella Bernardi

Questions are essential tools for acquiring the necessary information to complete information-seeking tasks. However, large language models (LLMs), especially open-source models, often perform poorly in generating informative questions, as measured by expected information gain (EIG). In this paper, we propose a method to enhance the informativeness of LLM-generated questions in 20-question game dialogues. We sample multiple questions from the same model (LLaMA 2-Chat 7B) for each game and create pairs of low-EIG and high-EIG questions to apply a Direct Preference Optimization (DPO) algorithm. Our results show that this method produces more effective questions (in terms of EIG), even in domains different from those used to train the DPO model.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Can LLMs Understand the Implication of Emphasized Sentences in Dialogue?

Guan-Ting Lin, Hung-yi Lee

Emphasis is a crucial component in human communication, which indicates speaker's intention and implication beyond pure text in dialogue. While Large Language Models (LLMs) have revolutionized natural language processing, their ability to understand emphasis in dialogue remains uncertain. This paper introduces Emphasized-Talk, a benchmark dataset with annotated dialogue samples capturing the implications of emphasis. We evaluate various LLMs, both open-source and commercial, to assess their performance in understanding and generating emphasis. Additionally, we propose an automatic evaluation pipeline using GPT-4, which achieve high correlation with human scoring. Our findings reveal that although commercial LLMs generally perform better, there is still significant room for improvement in comprehending emphasized sentences.

Industry

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Moleco: Molecular Contrastive Learning with Chemical Language Models for Molecular Property Prediction

Jun-Hyung Park, Hyuntae Park, Yeachan Kim, Woossang Lim, Sangkeun Lee

Pre-trained chemical language models (CLMs) excel in the field of molecular property prediction, utilizing string-based molecular descriptors such as SMILES for learning universal representations. However, such string-based descriptors implicitly contain limited structural information, which is closely associated with molecular property prediction. In this work, we introduce Moleco, a novel contrastive learning framework to enhance the understanding of molecular structures within CLMs. Based on the similarity of fingerprint vectors among different molecules, we train CLMs to distinguish structurally similar and dissimilar molecules in a contrastive manner. Experimental results demonstrate that Moleco significantly improves the molecular property prediction performance of CLMs, outperforming state-of-the-art models. Moreover, our in-depth analysis with diverse Moleco variants verifies that fingerprint vectors are highly effective features in improving CLMs' understanding of the structural information of molecules.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Refining App Reviews: Dataset, Methodology, and Evaluation

Amrita Singh, Chirag Jain, Mohit Chaudhary, Preethu Rose Anish

With the growing number of mobile users, app development has become increasingly lucrative. Reviews on platforms such as Google Play and Apple App Store provide valuable insights to developers, highlighting bugs, suggesting new features, and offering feedback. However, many reviews contain typos, spelling errors, grammar mistakes, and complex sentences, hindering efficient interpretation and slowing down app improvement processes. To tackle this, we introduce RARE (Repository for App review REfinement), a benchmark dataset of 10,000 annotated pairs of original and refined reviews from 10 mobile applications. These reviews were collaboratively refined by humans and large language models (LLMs). We also conducted an evaluation of eight state-of-the-art LLMs for automated review refinement. The top-performing model (Plan-T5) was further used to refine an additional 10,000 reviews, contributing to RARE as a silver corpus.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

RRADistill: Distilling LLMs' Passage Ranking Ability for Long-Tail Queries Document Re-Ranking on a Search Engine

Nayoung Choi, Youngjune Lee, Gyu-Hwung Cho, Haeyu Jeong, Jungmin Kong, Saejun Kim, Keunchan Park, Jaeho Choi, Sarah Cho, Inchang Jeong, Gyohee Nam, Sunghoon Han, Wonil Yang

Large Language Models (LLMs) excel at understanding the semantic relationships between queries and documents, even with lengthy and complex long-tail queries. These queries are challenging for feedback-based rankings due to sparse user engagement and limited feedback, making LLMs' ranking ability highly valuable. However, the large size and slow inference of LLMs necessitate the development of smaller, more efficient models (sLLMs). Recently, integrating ranking label generation into distillation techniques has become crucial, but existing methods utilize sLLM's capabilities and are cumbersome. Our research, RRADistill: Re-Ranking Ability Distillation, propose an efficient label generation pipeline and novel sLLM training methods for both encoder and decoder models. We introduce an encoder-based method using a Term Control Layer to capture term matching signals and a decoder-based model with a ranking layer for enhanced understanding. A/B testing on a Korean-based search platform, validates the effectiveness of our approach in improving re-ranking for long-tail queries.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Time Matters: An End-to-End Solution for Temporal Claim Verification

Anab Maulana Barik, Wyme Hsu, Mong-Li Lee

Automated claim verification plays an essential role in fostering trust in the digital space. Despite the growing interest, the verification of temporal claims has not received much attention in the community. Temporal claim verification brings new challenges where cues of the temporal information need to be extracted, and temporal reasoning involving various temporal aspects of the text must be applied. In this work, we describe an end-to-end solution for temporal claim verification that considers the temporal information in claims to obtain relevant evidence sentences and harnesses the power of a large language model for temporal reasoning. We curate two datasets comprising a diverse range of temporal claims to learn time-sensitive representations that encapsulate not only the semantic relationships among the events, but also their chronological proximity. Experiment results demonstrate that the proposed approach significantly enhances the accuracy of temporal claim verification, thereby advancing current state-of-the-art in automated claim verification.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Improving Hierarchical Text Clustering with LLM-guided Multi-view Cluster Representation

Anup Pattnaik, Cijo George, Rishabh Kumar Tripathi, Sasanka Rani Vutla, Jithendra Vepa

In this work, we present an approach that introduces different perspectives or views to improve the quality of hierarchical clustering of interaction drivers in a contact center. Specifically, we present a multi-stage approach that introduces LLM-guided multi-view cluster representation that significantly improves the quality of generated clusters. Our approach improves average Silhouette Score by upto 70% and Human Preference Scores by 36.7% for top-level clusters compared to standard agglomerative clustering for the given business use-case. We also present how the proposed approach can be adapted to cater to a standard non-hierarchical clustering use-cases where it achieves state-of-the-art performance on public datasets based on NMI and ACC scores, with minimal number of LLM queries compared to the current state-of-the-art approaches. Moreover, we apply our technique to generate two new labeled datasets for hierarchical clustering. We open-source these labeled datasets, validated and corrected by domain experts, for the benefit of the research community.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

RAG4ITOps: A Supervised Fine-Tunable and Comprehensive RAG Framework for IT Operations and Maintenance

Tianyang Zhang, Zhuoxuan Jiang, Shengguang Bai, Tianrui Zhang, Lin Lin, Yang Liu, Jiawei Ren

With the ever-increasing demands on Question Answering (QA) systems for IT operations and maintenance, an efficient and supervised fine-tunable framework is necessary to ensure the data security, private deployment and continuous upgrading. Although Large Language Models (LLMs) have notably improved the open-domain QA's performance, how to efficiently handle enterprise-exclusive corpora and build domain-specific QA systems are still less-studied for industrial applications. In this paper, we propose a general and comprehensive framework based on Retrieval Augmented Generation (RAG) and facilitate the whole business process of establishing QA systems for IT operations and maintenance. In accordance with the prevailing RAG method, our proposed framework, named with RAG4ITOps, composes of two major stages: (1) Model Fine-tuning & Data Vectorization, and (2) Online QA System Process. At the Stage 1, we leverage a contrastive learning method with two negative sampling strategies to fine-tune the embedding model, and design the instruction templates to fine-tune the LLM with a Retrieval Augmented Fine-Tuning method. At the Stage 2, an efficient process of QA system is built for serving. We collect enterprise-exclusive corpora from the domain of cloud computing, and the extensive experiments show that our method achieves superior results than counterparts on two kinds of QA tasks. Our experiment also provide a case for applying the RAG4ITOps to real-world enterprise-level applications.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Probing the Depths of Language Models' Contact-Center Knowledge for Quality Assurance

Digvijay Anil Ingle, Aashraya Sachdeva, Surya Prakash Sahu, Mayank Sati, Cijo George, Jithendra Vepa

Recent advancements in large Language Models (LMs) have significantly enhanced their capabilities across various domains, including natural language understanding and generation. In this paper, we investigate the application of LMs to the specialized task of contact-center Quality Assurance (QA), which requires both sophisticated linguistic understanding and deep domain knowledge. We conduct a comprehensive assessment of eight LMs, revealing that larger models, such as Claude-3.5-Sonnet, exhibit superior performance in comprehending contact-center conversations. We introduce methodologies to transfer this domain-specific knowledge to smaller models, by leveraging evaluation plans generated by more knowledgeable models, with optional human-in-the-loop refinement to enhance the capabilities of smaller models. Notably, our experimental results demonstrate an improvement of up to 18.95% in Macro F1 on in-house QA dataset. Our findings emphasize the importance of evaluation plan in guiding reasoning and highlight the potential of AI-assisted tools to advance objective, consistent, and scalable agent evaluation processes in contact-centers.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Intelligent Predictive Maintenance RAG framework for Power Plants: Enhancing QA with StyleDFS and Domain Specific Instruction Tuning

Seongtae Hong, Shin Joong Min, Jaehyung Seo, Taemin Lee, Jeongbae Park, Cho Man Young, Byeongho Choi, Heuiseok Lim

Process plants are complex large-scale industrial facilities that convert raw materials or intermediate products into final products, requiring continuous processes with high safety and efficiency standards. In particular, in nuclear process plants, Predictive Maintenance Systems (PMS) play a critical role in predicting equipment anomalies and performing preventive maintenance. However, current PMS relies heavily on the experience of a few experts, leading to knowledge loss upon their retirement and difficulty in swift response. Existing off-premise Question-Answering (QA) systems based on Large Language Models (LLM) face issues such as data leakage and challenges in domain-specific tuning. To address these problems, this study proposes an on-premise intelligent PMS framework utilizing a new chunking method, *StyleDFS*, which effectively reflects the structural information of documents. Additionally, we demonstrate that Instruction tuning using relevant domain-specific data improves LLM performance even under limited data conditions.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

A Hassle-free Algorithm for Strong Differential Privacy in Federated Learning Systems

Hugh Brendan McMahan, Zheng Xu, Yanxiang Zhang

Differential privacy (DP) and federated learning (FL) are combined as advanced privacy-preserving methods when training on-device language models in production mobile key/board applications. DP-Follow-the-Regularized-Leader (DP-FTRL) algorithms, leveraging correlated noise mechanisms such as tree aggregation or matrix factorization, are widely used in practice for their superior privacy-utility trade-off and compatibility with FL systems. This paper presents a novel variant of DP-FTRL by adapting the recent theoretical advancements of the Buffered Linear Toepplitz (BLT) mechanism to multi-participant scenarios. In the FL setting, our BLT mechanism demonstrates enhanced privacy-utility trade-off and improved memory efficiency than the widely used tree aggregation mechanism. Moreover, BLT achieves comparable privacy and utility to the state-of-the-art banded matrix factorization mechanism, while significantly simplifying usage requirements and reducing memory. The flexibility of the BLT mechanism allows seamless integration with existing DP FL implementations in production environments. We evaluate the BLT-DP-FTRL algorithm on the StackOverflow dataset, serving as a research simulation benchmark, and across four on-device language model tasks in a production FL system. Our empirical results highlight the potential of the BLT mechanism to elevate the practicality and effectiveness of DP in real-world scenarios.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach

Zhuowen Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Michael Bendersky

Retrieval Augmented Generation (RAG) has been a powerful tool for Large Language Models (LLMs) to efficiently process overly lengthy contexts. However, recent LLMs like Gemini-1.5 and GPT-4 show exceptional capabilities to understand long contexts directly. We conduct a comprehensive comparison between RAG and long-context (LC) LLMs, aiming to leverage the strengths of both. We benchmark RAG and LC across various public datasets using three latest LLMs. Results reveal that when resource sufficiently, LC consistently outperforms RAG in terms of average performance. However, RAG's significantly lower cost remains a distinct advantage. Based on this observation, we propose Self-Route, a simple yet effective method that routes queries to RAG or LC based on model self-reflection. Self-Route significantly reduces the computation cost while maintaining a comparable performance to LC. Our findings provide a guideline for long-context applications of LLMs using RAG and LC.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Tell me what I need to know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization

Frederic Kirstein, Terry Ruas, Robert Kratel, Bela Gipp

Meeting summarization is crucial in digital communication, but existing solutions struggle with salience identification to generate personalized, workable summaries, and context understanding to fully comprehend the meetings' content. Previous attempts to address these issues by considering related supplementary resources (e.g., presentation slides) alongside transcripts are hindered by models' limited context sizes and handling the additional complexities of the multi-source tasks, such as identifying relevant information in additional files and seamlessly aligning it with the meeting content. This work explores multi-source meeting summarization considering supplementary materials through a three-stage large language model approach: identifying transcript passages needing additional context, inferring relevant details from supplementary materials and inserting them into the transcript, and generating a summary from this enriched transcript. Our multi-source approach enhances model understanding, increasing summary relevance by 9% and producing more content-rich outputs. We introduce a personalization protocol that extracts participant characteristics and tailors summaries accordingly, improving informativeness by 10%. This work further provides insights on performance-cost trade-offs across four leading model families, including edge-device capable options. Our approach can be extended to similar complex generative tasks benefitting from additional resources and personalization, such as dialogue systems and action planning.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution

Ankita Sinha, Wendi Cui, Jiaxin Zhang, Kamalika Das

Large language models (LLMs) have demonstrated remarkable capabilities; however, the optimization of their prompts has historically prioritized performance metrics at the expense of crucial safety and security considerations. To overcome this shortcoming, we introduce "Survival of the Safest" (sosinspace), an innovative multi-objective prompt optimization framework that enhances both performance and security in LLMs simultaneously. sos utilizes an interleaved multi-objective evolution strategy, integrating semantic, feedback, and crossover mutations to effectively traverse the prompt landscape. Differing from the computationally demanding Pareto front methods, sos provides a scalable solution that expedites optimization in complex, high-dimensional discrete search spaces while keeping computational demands low. Our approach accommodates flexible weighting of objectives and generates a pool of optimized candidates, empowering users to select prompts that optimally meet their specific performance and security needs. Experimental evaluations across diverse benchmark datasets affirm sosinspace's efficacy in delivering high performance and notably enhancing safety and security compared to single-objective methods. This advancement marks a significant stride towards the deployment of LLM systems that are both high-performing and secure across varied in-

dustrial applications

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

AmazonQAC: A Large-Scale, Naturalistic Query Autocomplete Dataset

Dante Everaert, Rohit Patki, Tianqi Zheng, Christopher Potts

Query Autocomplete (QAC) is a critical feature in modern search engines, facilitating user interaction by predicting search queries based on input prefixes. Despite its widespread adoption, the absence of large-scale, realistic datasets has hindered advancements in QAC system development. This paper addresses this gap by introducing AmazonQAC, a new QAC dataset sourced from Amazon Search logs, comprising 367M samples. The dataset includes actual sequences of user-typed prefixes leading to final search terms, as well as session IDs and timestamps that support modeling the context-dependent aspects of QAC. We assess Prefix Trees, semantic retrieval, and Large Language Models (LLMs) with and without finetuning. We find that finetuned LLMs perform best, particularly when incorporating contextual information. However, even our best system achieves less than half of what we calculate is theoretically possible on our test data, which implies QAC is a challenging problem that is far from solved with existing systems. This contribution aims to stimulate further research on QAC systems to better serve user needs in diverse environments.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Value Alignment from Unstructured Text

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Manish Nagireddy, Pierre Dognin, Kush R. Varshney

Aligning large language models (LLMs) to value systems has emerged as a significant area of research within the fields of AI and NLP. Currently, this alignment process relies on the availability of high-quality supervised and preference data, which can be both time-consuming and expensive to curate or annotate. In this paper, we introduce a systematic end-to-end methodology for aligning LLMs to the implicit and explicit values represented in unstructured text data. Our proposed approach leverages the use of scalable synthetic data generation techniques to effectively align the model to the values present in the unstructured data. Through two distinct use-cases, we demonstrate the efficiency of our methodology on the MISTRAL-7B-Instruct model. Our approach credibly aligns LLMs to the values embedded within documents, and shows improved performance against other approaches, as quantified through the use of automatic metrics and win rates.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ConvKGYarn: Spinning Configurable and Scalable Conversational Knowledge Graph QA Datasets with Large Language Models

Ronak Pradeep, Daniel Lee, Ali Mousavi, Jeffrey Pound, Yisi Sang, Jimmy Lin, Thab Ilyas, Saloni Potdar, Mostafa Arefyan, Yunyao Li

The rapid evolution of Large Language Models (LLMs) and conversational assistants necessitates dynamic, scalable, and configurable conversational datasets for training and evaluation. These datasets must accommodate diverse user interaction modes, including text and voice, each presenting unique modeling challenges. Knowledge Graphs (KGs), with their structured and evolving nature, offer an ideal foundation for current and precise knowledge. Although human-curated KG-based conversational datasets exist, they struggle to keep pace with the rapidly changing user information needs. We present ConvKGYarn, a scalable method for generating up-to-date and configurable conversational KGQA datasets. Qualitative psychometric analyses demonstrate ConvKGYarn's effectiveness in producing high-quality data comparable to popular conversational KGQA datasets across various metrics. ConvKGYarn excels in adhering to human interaction configurations and operating at a significantly larger scale. We showcase ConvKGYarn's utility by testing LLMs on diverse conversations - exploring model behavior on conversational KGQA sets with different configurations grounded in the same KG fact set. Our results highlight the ability of ConvKGYarn to improve KGQA foundations and evaluate parametric knowledge of LLMs, thus offering a robust solution to the constantly evolving landscape of conversational assistants.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Sequential LLM Framework for Fashion Recommendation

Han Liu, Xianfeng Tang, Tianlang Chen, Jiapeng Liu, Indu Indu, Henry Peng Zou, Peng Dai, Roberto Fernandez Galan, Michael D. Porter, Dongmei Jia, Ning Zhang, Lian Xiong

The fashion industry is one of the leading domains in the global e-commerce sector, prompting major online retailers to employ recommendation systems for product suggestions and customer convenience. While recommendation systems have been extensively studied, the majority have been tailored to general e-commerce problems. These approaches often struggle with the unique challenges of the fashion domain. To address these challenges, we present a sequential fashion recommendation framework. By harnessing the extensive knowledge of a pre-trained large language model (LLM), our framework adeptly handles millions of cold-start items using specialized prompts. We then employ parameter-efficient fine-tuning with extensive fashion data and introduce a novel mix-up-based retrieval technique for converting text into items. Extensive experiments show our proposed framework significantly enhances fashion recommendation performance.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Patentformer: A Novel Method to Automate the Generation of Patent Applications

Juanyan Wang, Sai Krishna Reddy Mudiganti, Manali Sharma

In recent years, Large Language Models (LLMs) have demonstrated impressive performances across various NLP tasks. However, their potential for automating the task of writing patent documents remains relatively unexplored. To address this gap, in this work, we propose a novel method, Patentformer, for generating patent specification by fine-tuning the generative models with diverse sources of information, e.g., patent claims, drawing text, and brief descriptions of the drawings. To enhance the generative models' comprehension of the complex task of writing patent specification, we introduce a new task, claim+drawing-to-specification, and release a new dataset. We evaluate our proposed method on thousands of patents from the USPTO and show that our method can generate good patent specification in legal writing style. Human evaluations by four patent experts further affirm that our proposed method can generate correct specification about 50% of the times, and the quality of generated specification may sometimes be better than the actual specification.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

RAC: Retrieval-augmented Conversation Dataset for Open-domain Question Answering in Conversational Settings

Bonggeun Choi, Jeongjae Park, Yoonsung Kim, Jae-Hyun Park, Younjoong Ko

In recent years, significant advancements in conversational question and answering (CQA) have been driven by the exponential growth of large language models and the integration of retrieval mechanisms that leverage external knowledge to generate accurate and contextually relevant responses. Consequently, the fields of conversational search and retrieval-augmented generation (RAG) have obtained substantial attention for their capacity to address two key challenges: query rewriting within conversational histories and generating responses by employing retrieved knowledge. However, both fields are often independently studied, and comprehensive study on entire systems remains underexplored. In this work, we present a novel retrieval-augmented conversation (RAC) dataset and develop a RAC system comprising query rewriting, retrieval, reranking, and response generation stages. Experimental results demonstrate the competitiveness of the system and extensive analyses are conducted to apprehend the impact of retrieval results to response generation.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Improving Retrieval in Sponsored Search by Leveraging Query Context Signals

Akash Kumar Mohankumar, Gururaj K, Gagan Madan, Amit S

Accurately retrieving relevant bid keywords for user queries is critical in Sponsored Search but remains challenging, particularly for short, ambiguous queries. Existing dense and generative retrieval models often fail to capture the nuanced user intent in these cases. To address this, we propose an approach to enhance query understanding by augmenting queries with rich contextual signals derived from web search results and large language models, stored in an online cache. Specifically, we use web search titles and snippets to ground queries in real-world information, and utilize GPT-4 to generate query rewrites and explanations that clarify user intent. These signals are efficiently integrated through a Fusion-in-Decoder based Unity architecture, enabling both dense and generative retrieval with serving costs on par with traditional context-free models. To address scenarios where context is unavailable in the cache, we introduce context glancing, a curriculum learning strategy that improves model robustness and performance even without contextual signals during inference. Extensive offline experiments demonstrate that our context-aware approach substantially outperforms context-free models. Furthermore, online A/B testing on a prominent search engine across 160+ countries shows significant improvements in user engagement and revenue.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Salient Information Prompting to Steer Content in Prompt-based Abstractive Summarization

Lei Xu, Mohammed Asad Karim, Saket Dingliwal, Aparna Elangovan

Large language models (LLMs) can generate fluent summaries across domains using prompting techniques, reducing the effort required for summarization applications. However, crafting effective prompts that guide LLMs to generate summaries with the appropriate level of detail and writing style remains a challenge. In this paper, we explore the use of salient information extracted from the source document to enhance summarization prompts. We show that adding keyphrases in prompts can improve ROUGE F1 and recall, making the generated summaries more similar to the reference and more complete. The number of keyphrases can control the precision-recall trade-off. Furthermore, our analysis reveals that incorporating phrase-level salient information is superior to word- or sentence-level. However, the impact on summary faithfulness is not universally positive across LLMs. To enable this approach, we introduce Keyphrase Signal Extractor (SigExt), a lightweight model that can be finetuned to extract salient keyphrases. By using SigExt, we achieve consistent ROUGE improvements across datasets and LLMs without any LLM customization. Our findings provide insights into leveraging salient information in building prompt-based summarization systems.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Don't Shoot The Breeze: Topic Continuity Model Using Nonlinear Naive Bayes With Attention

Shu-Ting Pi, Pradeep Bagavani, Yejia Li, Disha, Qun Liu

Utilizing Large Language Models (LLM) as chatbots in diverse business scenarios often presents the challenge of maintaining topic continuity. Abrupt shifts in topics can lead to poor user experiences and inefficient utilization of computational resources. In this paper, we present a topic continuity model aimed at assessing whether a response aligns with the initial conversation topic. Our model is built upon the expansion of the corresponding natural language understanding (NLU) model into quantifiable terms using a Naive Bayes approach. Subsequently, we have introduced an attention mechanism and logarithmic nonlinearity to enhance its capability to capture topic continuity. This approach allows us to convert the NLU model into an interpretable analytical formula. In contrast to many NLU models constrained by token limits, our proposed model can seamlessly handle conversations of any length with linear time complexity. Furthermore, the attention mechanism significantly improves the model's ability to identify topic continuity in complex conversations. According to our experiments, our model consistently outperforms traditional methods, particularly in handling lengthy and intricate conversations. This unique capability offers us an opportunity to ensure the responsible and interpretable use of LLMs.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Investigating the Personality Consistency in Quantized Role-Playing Dialogue Agents

Yixiao Wang, Homa Fashandi, Kevin Ferreira

This study explores the consistency of personality traits in quantized large language models (LLMs) for edge device role-playing scenarios. Using the Big Five personality traits model, we evaluate how stable assigned personalities are for Quantized Role-Playing Dialog Agents (QRPDA) during multi-turn interactions. We evaluate multiple LLMs with various quantization levels, combining binary indexing of personality traits, explicit self-assessments, and linguistic analysis of narratives. To address personality inconsistency, we propose a non-parametric method called Think2. Our multi-faceted evaluation framework demonstrates Think2's effectiveness in maintaining consistent personality traits for QRPDA. Moreover, we offer insights to help select the optimal model for QRPDA, improving its stability and reliability in real-world applications.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ReportGPT: Human-in-the-loop Verifiable Table-to-Text Generation

Lucas Cecchi, Petr Babkin

Recent developments in the quality and accessibility of large language models have precipitated a surge in user-facing tools for content generation. Motivated by a necessity for human quality control of these systems, we introduce ReportGPT: a pipeline framework for verifiable human-in-the-loop table-to-text generation. ReportGPT is based on a domain specific language, which acts as a proof mechanism for generating verifiable commentary. This allows users to quickly check the relevancy and factuality of model outputs. User selections then become few-shot examples for improving the performance of the pipeline. We configure 3 approaches to our pipeline, and find that usage of language models in ReportGPT's components trade off precision for more insightful downstream commentary. Furthermore, ReportGPT learns from human feedback in real-time, needing only a few samples to improve performance.

Information Retrieval and Text Mining 3

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Search if you don't know! Knowledge-Augmented Korean Grammatical Error Correction with Large Language Models

Seonmin Koo, Jinsung Kim, Chanjun Park, Heuiseok Lim

Grammatical error correction (GEC) system is a practical task used in the real world, showing high achievements alongside the development of large language models (LLMs). However, these achievements have been primarily obtained in English, and there is a relative lack of performance for non-English data, such as Korean. We hypothesize that this insufficiency occurs because relying solely on the parametric knowledge of LLMs makes it difficult to thoroughly understand the given context in the Korean GEC. Therefore, we propose a Knowledge-Augmented GEC (KAGEC) framework that incorporates evidential information from external sources into the prompt for the GEC task. KAGEC first extracts salient phrases from the given source and retrieves non-parametric knowledge based on these phrases, aiming to enhance the context-aware generation capabilities of LLMs. Furthermore, we conduct validations for fine-grained error types to identify those

requiring a retrieval-augmented manner when LLMs perform Korean GEC. According to experimental results, most LLMs, including ChatGPT, demonstrate significant performance improvements when applying KAGEC.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Towards Low-Resource Harmful Meme Detection with LMM Agents

Jianzhao Huang, Hongzhan Lin, Ziyang Luo, Guang Chen, Jing Ma

The proliferation of Internet memes in the age of social media necessitates effective identification of harmful ones. Due to the dynamic nature of memes, existing data-driven models may struggle in low-resource scenarios where only a few labeled examples are available. In this paper, we propose an agency-driven framework for low-resource harmful meme detection, employing both outward and inward analysis with few-shot annotated samples. Inspired by the powerful capacity of Large Multimodal Models (LMMs) on multimodal reasoning, we first retrieve relative memes with annotations to leverage label information as auxiliary signals for the LMM agent. Then, we elicit knowledge-revising behavior within the LMM agent to derive well-generalized insights into meme harmfulness. By combining these strategies, our approach enables dialectical reasoning over intricate and implicit harm-indicative patterns. Extensive experiments conducted on three meme datasets demonstrate that our proposed approach achieves superior performance than state-of-the-art methods on the low-resource harmful meme detection task.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Evaluating D-MERIT of Partial-annotation on Information Retrieval

Roij Rassin, Yaron Fairstein, Oren Kalinsky, Guy Kushlevitz, Nachshon Cohen, Alexander Libov, Yoav Goldberg

Retrieval models are often evaluated on partially-annotated datasets. Each query is mapped to a few relevant texts, and the remaining corpus is assumed to be irrelevant. As a result, models that successfully retrieve falsely labeled negatives are punished during evaluation. Unfortunately, completely annotating all texts for every query is not resource-efficient. In this work, we show that using partially-annotated datasets in evaluation can paint a distorted picture. We curate D-MERIT, a passage retrieval evaluation set from Wikipedia, aspiring to contain "all" relevant passages for each query. Queries describe a group (e.g., "journals about linguistics"), and relevant passages are evidence that entities belong to the group (e.g., a passage indicating that "Language" is a journal about linguistics). We show that evaluating on a dataset containing annotations for only a subset of the relevant passages might result in misleading ranking of the retrieval systems and that as more relevant texts are included in the evaluation set, the rankings converge. We propose our dataset as a resource for evaluation and our study as a recommendation for a balance between resource-efficiency and reliable evaluation when annotating evaluation sets for text retrieval. Our dataset can be downloaded from <https://D-MERIT.github.io>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

AGRaME: Any-Granularity Ranking with Multi-Vector Embeddings

Revanth Gangi Reddy, Omar Attia, Yunyao Li, Heng Ji, Saloni Potdar

Ranking is a fundamental problem in search, however, existing ranking algorithms usually restrict the granularity of ranking to full passages or require a specific dense index for each desired level of granularity. Such lack of flexibility in granularity negatively affects many applications that can benefit from more granular ranking, such as sentence-level ranking for open-domain QA, or proposition-level ranking for attribution. In this work, we introduce the idea of any-granularity ranking which leverages multi-vector embeddings to rank at varying levels of granularity while maintaining encoding at a single (coarser) level of granularity. We propose a multi-granular contrastive loss for training multi-vector approaches and validate its utility with both sentences and propositions as ranking units. Finally, we demonstrate the application of proposition-level ranking to post-hoc citation addition in retrieval-augmented generation, surpassing the performance of prompt-driven citation generation.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Exploring the Practicality of Generative Retrieval on Dynamic Corpora

Soyoung Yoon, Chaeun Kim, Hyunjoo Lee, Joel Jang, Sohee Yang, Minjoon Seo

Benchmarking the performance of information retrieval (IR) is mostly conducted with a fixed set of documents (static corpora). However, in realistic scenarios, this is rarely the case and the documents to be retrieved are constantly updated and added. In this paper, we focus on Generative Retrievals (GR), which apply autoregressive language models to IR problems, and explore their adaptability and robustness in dynamic scenarios. We also conduct an extensive evaluation of computational and memory efficiency, crucial factors for real-world deployment of IR systems handling vast and ever-changing document collections. Our results on the StreamingQA benchmark demonstrate that GR is more adaptable to evolving knowledge (411%), robust in learning knowledge with temporal information, and efficient in terms of inference FLOPs (x2), indexing time (x6), and storage footprint (x4) compared to Dual Encoders (DE), which are commonly used in retrieval systems. Our paper highlights the potential of GR for future use in practical IR systems within dynamic environments.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Open-world Multi-label Text Classification with Extremely Weak Supervision

Xintong Li, Jinya Jiang, Rit Dharmani, Jayanth Srinivas, Gaowen Liu, Jingbo Shang

We study open-world multi-label text classification under extremely weak supervision (XWS), where the user only provides a brief description for classification objectives without any labels or ground-truth label space. Similar single-label XWS settings have been explored recently, however, these methods cannot be easily adapted for multi-label. We observe that (1) most documents have a dominant class covering the majority of content and (2) long-tail labels would appear in some documents as a dominant class. Therefore, we first utilize the user description to prompt a large language model (LLM) for dominant keyphrases of a subset of raw documents, and then construct a (initial) label space via clustering. We further apply a zero-shot multi-label classifier to locate the documents with small top predicted scores, so we can revisit their dominant keyphrases for more long-tail labels. We iterate this process to discover a comprehensive label space and construct a multi-label classifier as a novel method, X-MLClass. X-MLClass exhibits a remarkable increase in ground-truth label space coverage on various datasets, for example, a 40% improvement on the AAPD dataset over topic modeling and keyword extraction methods. Moreover, X-MLClass achieves the best end-to-end multi-label classification accuracy.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Unleashing the Power of Emojis in Texts via Self-supervised Graph Pre-Training

Zhou Zhang, Dongzeng Tan, Juan Wang, Yilong Chen, Jiarong Xu

Emojis have gained immense popularity on social platforms, serving as a common means to supplement or replace text. However, existing data mining approaches generally either completely ignore or simply treat emojis as ordinary Unicode characters, which may limit the model's ability to grasp the rich semantic information in emojis and the interaction between emojis and texts. Thus, it is necessary to release the emoji's power in social media data mining. To this end, we first construct a heterogeneous graph consisting of three types of nodes, i.e. post, word and emoji nodes to improve the representation of different elements in posts. The edges are also well-defined to model how these three elements interact with each other. To facilitate the sharing of information among post, word and emoji nodes, we propose a graph pre-train framework for text and emoji co-modeling, which contains two graph pre-training tasks: node-level graph contrastive learning and edge-level link reconstruction learning. Extensive experiments on the Xiaohongshu and Twitter datasets with two types of downstream tasks

demonstrate that our approach proves significant improvement over previous strong baseline methods.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

LumberChunker: Long-Form Narrative Document Segmentation

André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, Arlindo L. Oliveira

Modern NLP tasks increasingly rely on dense retrieval methods to access up-to-date and relevant contextual information. We are motivated by the premise that retrieval benefits from segments that can vary in size such that a contents semantic independence is better captured. We propose LumberChunker, a method leveraging an LLM to dynamically segment documents, which iteratively prompts the LLM to identify the point within a group of sequential passages where the content begins to shift. To evaluate our method, we introduce GutenQA, a benchmark with 3000 "needle in a haystack" type of question-answer pairs derived from 100 public domain narrative books available on Project Gutenberg. Our experiments show that LumberChunker not only outperforms the most competitive baseline by 7.37% in retrieval performance (DCG@20) but also that, when integrated into a RAG pipeline, LumberChunker proves to be more effective than other chunking methods and competitive baselines, such as the Gemini 1.5M Pro.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Evidence Retrieval for Fact Verification using Multi-stage Reranking

Shrikant Malviya, Stamatis Katsigianis

In the fact verification domain, the accuracy and efficiency of evidence retrieval are paramount. This paper presents a novel approach to enhance the fact verification process through a Multi-stage ReRanking (M-ReRank) paradigm, which addresses the inherent limitations of single-stage evidence extraction. Our methodology leverages the strengths of advanced reranking techniques, including dense retrieval models and list-aware rerankers, to optimise the retrieval and ranking of evidence of both structured and unstructured types. We demonstrate that our approach significantly outperforms previous state-of-the-art models, achieving a recall rate of 93.63% for Wikipedia pages. The proposed system not only improves the retrieval of relevant sentences and table cells but also enhances the overall verification accuracy. Through extensive experimentation on the FEVEROUS dataset, we show that our M-ReRank pipeline achieves substantial improvements in evidence extraction, particularly increasing the recall of sentences by 7.85%, tables by 8.29% and cells by 3% compared to the current state-of-the-art on the development set.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Knowledge Navigator: Hierarchical Subtopic Organization for Exploratory Search in Scientific Literature

Uri Katz, Mosh Levy, Yoav Goldberg

The exponential growth of scientific literature necessitates advanced tools for effective knowledge exploration. We present Knowledge Navigator, a system designed to enhance exploratory search abilities by organizing and structuring the retrieved documents from broad topical queries into a navigable, two-level hierarchy of named and descriptive scientific topics and subtopics. This structured organization provides an overall view of the research themes in a domain, while also enabling iterative search and deeper knowledge discovery within specific subtopics by allowing users to refine their focus and retrieve additional relevant documents. Knowledge Navigator combines LLM capabilities with cluster-based methods to enable an effective browsing method. We demonstrate our approach's effectiveness through automatic and manual evaluations on two novel benchmarks, CLUSTREC-COVID and SCITOC. Our code, prompts, and benchmarks are made publicly available.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation

Andreea Iana, Goran Glava, Heiko Paulheim

Recent neural news recommenders (NNRs) extend content-based recommendation (1) by aligning additional aspects (e.g., topic, sentiment) between candidate news and user history or (2) by diversifying recommendations w.r.t. these aspects. This customization is achieved by "hardcoding" additional constraints into the NNR's architecture and/or training objectives: any change in the desired recommendation behavior thus requires retraining the model with a modified objective. This impedes widespread adoption of multi-aspect news recommenders. In this work, we introduce MANNeR, a modular framework for multi-aspect neural news recommendation that supports on-the-fly customization over individual aspects at inference time. With metric-based learning as its backbone, MANNeR learns aspect-specialized news encoders and then flexibly and linearly combines the resulting aspect-specific similarity scores into different ranking functions, alleviating the need for ranking function-specific retraining of the model. Extensive experimental results show that MANNeR consistently outperforms state-of-the-art NNRs on both standard content-based recommendation and single- and multi-aspect customization. Lastly, we validate that MANNeR's aspect-customization module is robust to language and domain transfer.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Schema-Driven Information Extraction from Heterogeneous Tables

Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, Alan Ritter

In this paper, we explore the question of whether large language models can support cost-efficient information extraction from tables. We introduce schema-driven information extraction, a new task that transforms tabular data into structured records following a human-authored schema. To assess various LLM's capabilities on this task, we present a benchmark comprised of tables from four diverse domains: machine learning papers, chemistry literature, material science journals, and webpages. We use this collection of annotated tables to evaluate the ability of open-source and API-based language models to extract information from tables covering diverse domains and data formats. Our experiments demonstrate that surprisingly competitive performance can be achieved without requiring task-specific pipelines or labels, achieving F1 scores ranging from 74.2 to 96.1, while maintaining cost efficiency. Moreover, through detailed ablation studies and analyses, we investigate the factors contributing to model success and validate the practicality of distilling compact models to reduce API reliance.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection

Moxin Li, Wenjie Wang, Fulij Feng, Fengbin ZHU, Qifan Wang, Tat-Seng Chua

Self-detection for Large Language Models (LLMs) seeks to evaluate the trustworthiness of the LLM's output by leveraging its own capabilities, thereby alleviating the issue of output hallucination. However, existing self-detection approaches only retrospectively evaluate answers generated by LLM, typically leading to the over-trust in incorrectly generated answers. To tackle this limitation, we propose a novel self-detection paradigm that considers the comprehensive answer space beyond LLM-generated answers. It thoroughly compares the trustworthiness of multiple candidate answers to mitigate the over-trust in LLM-generated incorrect answers. Building upon this paradigm, we introduce a two-step framework, which firstly instructs LLM to reflect and provide justifications for each candidate answer, and then aggregates the justifications for comprehensive target answer evaluation. This framework can be seamlessly integrated with existing approaches for superior self-detection. Extensive experiments on six datasets spanning three tasks demonstrate the effectiveness of the proposed framework.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Numbers Matter! Bringing Quantity-awareness to Retrieval Systems

Satya Almasian, Milena Bruseva, Michael Gertz

Quantitative information plays a crucial role in understanding and interpreting the content of documents. Many user queries contain quantities and cannot be resolved without understanding their semantics, e.g., “car that costs less than \$10k”. Yet, modern search engines apply the same ranking mechanisms for both words and quantities, overlooking magnitude and unit information. In this paper, we introduce two quantity-aware ranking techniques designed to rank both the quantity and textual content either jointly or independently. These techniques incorporate quantity information in available retrieval systems and can address queries with numerical conditions equal, greater than, and less than. To evaluate the effectiveness of our proposed models, we introduce two novel quantity-aware benchmark datasets in the domains of finance and medicine and compare our method against various lexical and neural models. The code and data are available under <https://github.com/satyat77/QuantityAwareRankers>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

HyQE: Ranking Contexts with Hypothetical Query Embeddings

Weichao Zhou, Jiaxin Zhang, Hilaf Hasson, Anu Singh, Wenchao Li

In retrieval-augmented systems, context ranking techniques are commonly employed to reorder the retrieved contexts based on their relevance to a user query. A standard approach is to measure this relevance through the similarity between contexts and queries in the embedding space. However, such similarity often fails to capture the relevance. Alternatively, large language models (LLMs) have been used for ranking contexts. However, they can encounter scalability issues when the number of candidate contexts grows and the context window sizes of the LLMs remain constrained. Additionally, these approaches require fine-tuning LLMs with domain-specific data. In this work, we introduce a scalable ranking framework that combines embedding similarity and LLM capabilities without requiring LLM fine-tuning. Our framework uses a pre-trained LLM to hypothesize the user query based on the retrieved contexts and ranks the context based on the similarity between the hypothesized queries and the user query. Our framework is efficient at inference time and is compatible with many other retrieval and ranking techniques. Experimental results show that our method improves the ranking performance across multiple benchmarks.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Topic Modeling: Contextual Token Embeddings Are All You Need

Dima Angelov, Diana Inkpen

The goal of topic modeling is to find meaningful topics that capture the information present in a collection of documents. The main challenges of topic modeling are finding the optimal number of topics, labeling the topics, segmenting documents by topic, and evaluating topic model performance. Current neural approaches have tackled some of these problems but none have been able to solve all of them. We introduce a novel topic modeling approach, Contextual-Top2Vec, which uses document contextual token embeddings, it creates hierarchical topics, finds topic spans within documents and labels topics with phrases rather than just words. We propose the use of BERTScore to evaluate topic coherence and to evaluate how informative topics are of the underlying documents. Our model outperforms the current state-of-the-art models on a comprehensive set of topic model evaluation metrics.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Dense Passage Retrieval: Is It Retrieving?

Benjamin Reichman, Larry Heck

Large Language Models (LLMs) internally store repositories of knowledge. However, their access to this repository is imprecise and they frequently hallucinate information that is not true or does not exist. A paradigm called Retrieval Augmented Generation (RAG) promises to fix these issues. Dense passage retrieval (DPR) is the first step in this paradigm. In this paper, we analyze the role of DPR fine-tuning and how it affects the model being trained. DPR fine-tunes pre-trained networks to enhance the alignment of the embeddings between queries and relevant textual data. We explore DPR-trained models mechanistically by using a combination of probing, layer activation analysis, and model editing. Our experiments show that DPR training **decentralizes** how knowledge is stored in the network, creating **multiple access pathways** to the same information. We also uncover a **limitation** in this training style: the **internal knowledge** of the pre-trained model **bounds** what the retrieval model can retrieve. These findings suggest a few possible directions for dense retrieval: (1) expose the DPR training process to more knowledge so more can be decentralized, (2) inject facts as decentralized representations, (3) model and incorporate knowledge uncertainty in the retrieval process, and (4) directly map internal model knowledge to a knowledge base.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

R³-NL2GQL: A Model Coordination and Knowledge Graph Alignment Approach for NL2GQL

Yuhang Zhou, Yu He, Siyu Tian, Yuchen Ni, Zhangyue Yin, Xiang Liu, Chuanjun Ji, Sen Liu, Xipeng Qiu, Guangan Ye, Hongfeng Chai

While current tasks of converting natural language to SQL (NL2SQL) using Foundation Models have shown impressive achievements, adapting these approaches for converting natural language to Graph Query Language (NL2GQL) encounters hurdles due to the distinct nature of GQL compared to SQL, alongside the diverse forms of GQL. Moving away from traditional rule-based and slot-filling methodologies, we introduce a novel approach, R^3 -NL2GQL, integrating both small and large Foundation Models for ranking, rewriting, and refining tasks. This method leverages the interpretative strengths of smaller models for initial ranking and rewriting stages, while capitalizing on the superior generalization and query generation prowess of larger models for the final transformation of natural language queries into GQL formats. Addressing the scarcity of datasets in this emerging field, we have developed a bilingual dataset, sourced from graph database manuals and selected open-source Knowledge Graphs (KGs). Our evaluation of this methodology on this dataset demonstrates its promising efficacy and robustness.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ConTReGen: Context-driven Tree-structured Retrieval for Open-domain Long-form Text Generation

Kashob Kumar Roy, Pritom Saha Akash, Lucian Popa, Kevin Chen-Chuan Chang

Open-domain long-form text generation requires generating coherent, comprehensive responses that address complex queries with both breadth and depth. This task is challenging due to the need to accurately capture diverse facets of input queries. Existing iterative retrieval-augmented generation (RAG) approaches often struggle to delve deeply into each facet of complex queries and integrate knowledge from various sources effectively. This paper introduces ConTReGen, a novel framework that employs a context-driven, tree-structured retrieval approach to enhance the depth and relevance of retrieved content. ConTReGen integrates a hierarchical, top-down in-depth exploration of query facets with a systematic bottom-up synthesis, ensuring comprehensive coverage and coherent integration of multifaceted information. Extensive experiments on multiple datasets, including LFQA and ODSUM, alongside a newly introduced dataset, ODSUM-WikiHow, demonstrate that ConTReGen outperforms existing state-of-the-art RAG models.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Prefix-VAE: Efficient and Consistent Short-Text Topic Modeling with LLMs

Pritom Saha Akash, Kevin Chen-Chuan Chang

Topic modeling is a powerful technique for uncovering hidden themes within a collection of documents. However, the effectiveness of traditional topic models often relies on sufficient word co-occurrence, which is lacking in short texts. Therefore, existing approaches, whether probabilistic or neural, frequently struggle to extract meaningful patterns from such data, resulting in incoherent topics. To address this chal-

length, we propose a novel approach that leverages large language models (LLMs) to extend short texts into more detailed sequences before applying topic modeling. To further improve the efficiency and solve the problem of semantic inconsistency from LLM-generated texts, we propose to use prefix tuning to train a smaller language model coupled with a variational autoencoder for short-text topic modeling. Our method significantly improves short-text topic modeling performance, as demonstrated by extensive experiments on real-world datasets with extreme data sparsity, outperforming current state-of-the-art topic models.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Exploring the Best Practices of Query Expansion with Large Language Models

Le Zhang, Yihong Wu, Qian Yang, Jian-Yun Nie

Large Language Models (LLMs) are foundational in language technologies, particularly in information retrieval (IR). In this paper, we thoroughly explore the best practice of leveraging LLMs for query expansion. To this end, we introduce a training-free, straightforward yet effective framework called Multi-Text Generation Integration (MuGI). This approach leverages LLMs to generate multiple pseudo-references, which are then integrated with the original queries to enhance both sparse and dense retrieval methods. Additionally, we introduce a retrieval pipeline based on MuGI, which combines the strengths of sparse and dense retrievers to achieve superior performance without the need for costly pre-indexing. Our empirical findings reveal that: (1) Increasing the number of samples from LLMs benefits IR systems; (2) A balance between the query and pseudo-documents, and an effective integration strategy, is critical for high performance; (3) Contextual information from LLMs is essential, even boosts a 23M model to outperform a 7B baseline model; (4) Pseudo relevance feedback can further calibrate queries for improved performance; and (5) Query expansion is widely applicable and versatile, consistently enhancing models ranging from 23M to 7B parameters. Our code and all generated references are made available at https://github.com/lezhang7/Retrieval_MuGI.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Revisiting Query Variation Robustness of Transformer Models

Tim Hagen, Harrison Scells, Martin Potthast

The most commonly used transformers for retrieval at present, BERT and T5, have been shown not to be robust to query variations such as typos or paraphrases. Although this is an important prerequisite for their practicality, this problem has hardly been investigated. More recent large language models (LLMs), including instruction-tuned LLMs, have not been analyzed yet, and only one study looks beyond typos. We close this gap by reproducing this study and extending it with a systematic analysis of more recent models, including Sentence-BERT, CharacterBERT, E5-MISTRAL, AngIE, and Ada v2. We further investigate if instruct-LLMs can be prompted for robustness. Our results are mixed in that the previously observed robustness issues for cross-encoders also apply to bi-encoders that use much larger LLMs, albeit to a lesser extent. While further LLM scaling may improve their embeddings, their cost-effective use for all but large deployments is limited. Training data that includes query variations allows LLMs to be fine-tuned for more robustness, but focusing on a single category of query variation may even degrade the effectiveness on others. Our code, results, and artifacts can be found at <https://github.com/webis-de/EMNLP-24>

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Re-Invoke: Tool Invocation Rewriting for Zero-Shot Tool Retrieval

Yanfei Chen, Jinsung Yoon, Devendra Singh Sachan, Qingze Wang, Vincent Cohen-Addad, Mohammadhossein Bateni, Chen-Yu Lee, Tomas Pfister

Recent advances in large language models (LLMs) have enabled autonomous agents with complex reasoning and task-fulfillment capabilities using a wide range of tools. However, effectively identifying the most relevant tools for a given task becomes a key bottleneck as the toolset size grows, hindering reliable tool utilization. To address this, we introduce Re-Invoke, an unsupervised tool retrieval method designed to scale effectively to large toolsets without training. Specifically, we first generate a diverse set of synthetic queries that comprehensively cover different aspects of the query space associated with each tool document during the tool indexing phase. Second, we leverage LLM's query understanding capabilities to extract key tool-related context and underlying intents from user queries during the inference phase. Finally, we employ a novel multi-view similarity ranking strategy based on intents to pinpoint the most relevant tools for each query. Our evaluation demonstrates that Re-Invoke significantly outperforms state-of-the-art alternatives in both single-tool and multi-tool scenarios, all within a fully unsupervised setting. Notably, on the ToolE datasets, we achieve a 20% relative improvement in nDCG@5 for single-tool retrieval and a 39% improvement for multi-tool retrieval.

Language Modeling 3

Nov 13 (Wed) 16:00-17:30 - Room: Jasmine

Nov 13 (Wed) 16:00-17:30 - Jasmine

Mitigating Frequency Bias and Anisotropy in Language Model Pre-Training with Syntactic Smoothing

Richard Diehl Martinez, Zebulon Goriely, Andrew Caines, Paula Buttery, Lisa Beinborn

Language models strongly rely on frequency information because they maximize the likelihood of tokens during pre-training. As a consequence, language models tend to not generalize well to tokens that are seldom seen during training. Moreover, maximum likelihood training has been discovered to give rise to anisotropy: representations of tokens in a model tend to cluster tightly in a high-dimensional cone, rather than spreading out over their representational capacity. Our work introduces a method for quantifying the frequency bias of a language model by assessing sentence-level perplexity with respect to token-level frequency. We then present a method for reducing the frequency bias of a language model by inducing a syntactic prior over token representations during pre-training. Our Syntactic Smoothing method adjusts the maximum likelihood objective function to distribute the learning signal to syntactically similar tokens. This approach results in better performance on infrequent English tokens and a decrease in anisotropy. We empirically show that the degree of anisotropy in a model correlates with its frequency bias.

Nov 13 (Wed) 16:00-17:30 - Jasmine

PSC: Extending Context Window of Large Language Models via Phase Shift Calibration

Wenqiao Zhu, Chao Xu, Lulu Wang, Jun Wu

Rotary Position Embedding (RoPE) is an efficient position encoding approach and is widely utilized in numerous large language models (LLMs). Recently, a lot of methods have been put forward to further expand the context window based on RoPE. The core concept of those methods is to predefine or search for a set of factors to rescale the base frequencies of RoPE. Nevertheless, it is quite a challenge for existing methods to predefine an optimal factor due to the exponential search space. In view of this, we introduce PSC (Phase Shift Calibration), a small module for calibrating the frequencies predefined by existing methods. With the employment of PSC, we demonstrate that many existing methods can be further enhanced, like PI, YaRN, and LongRoPE. We conducted extensive experiments across multiple models and tasks. The results demonstrate that (1) when PSC is enabled, the comparative reductions in perplexity increase as the context window size is varied from 16k, to 32k, and up to 64k. (2) Our approach is broadly applicable and exhibits robustness across a variety of models and tasks.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Advancing Adversarial Suffix Transfer Learning on Aligned Large Language Models

Hongfu Liu, Yuxi Xie, Ye Wang, Michael Shieh

Large Language Models (LLMs) face safety concerns due to potential misuse by malicious users. Recent red-teaming efforts have identified adversarial suffixes capable of jailbreaking LLMs using the gradient-based search algorithm Greedy Coordinate Gradient (GCG). However, GCG struggles with computational inefficiency, limiting further investigations regarding suffix transferability and scalability across models and data. In this work, we bridge the connection between search efficiency and suffix transferability. We propose a two-stage transfer learning framework, DeGCG, which decouples the search process into behavior-agnostic pre-searching and behavior-relevant post-searching. Specifically, we employ direct first target token optimization in pre-searching to facilitate the search process. We apply our approach to cross-model, cross-data, and self-transfer scenarios. Furthermore, we introduce an interleaved variant of our approach, i-DeGCG, which iteratively leverages self-transferability to accelerate the search process. Experiments on HarmBench demonstrate the efficiency of our approach across various models and domains. Notably, our i-DeGCG outperforms the baseline on Llama2-chat-7b with ASRs of 43.9 (+22.2) and 39.0 (+19.3) on valid and test sets, respectively. Further analysis on cross-model transfer indicates the pivotal role of first target token optimization in leveraging suffix transferability for efficient searching.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Tools Fail: Detecting Silent Errors in Faulty Tools

Jimin Sun, So Yeon Min, Yingshan Chang, Yonatan Bisk

Tools have become a mainstay of LLMs, allowing them to retrieve knowledge not in their weights, to perform tasks on the web, and even to control robots. However, most ontologies and surveys of tool-use have assumed the core challenge for LLMs is choosing the tool. Instead, we introduce a framework for tools more broadly which guides us to explore a model's ability to detect "silent" tool errors, and reflect on how to plan. This more directly aligns with the increasingly popular use of models as tools. We provide an initial approach to failure recovery with promising results both on a controlled calculator setting and embodied agent planning.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Scalable Efficient Training of Large Language Models with Low-dimensional Projected Attention

Xingtai Lv, Ning Ding, Kaitian Zhang, Ermo Hua, Gangui Cui, Bowen Zhou

Improving the effectiveness and efficiency of large language models (LLMs) simultaneously is a critical yet challenging research goal. In this paper, we find that low-rank pre-training, normally considered as efficient methods that will compromise performance, can be scalably effective when reduced parameters are precisely targeted. Specifically, by applying low-dimensional modules only to the attention layer resolves this issue and enhances both effectiveness and efficiency. We refer to this structure as *Low-dimensional Projected Attention (LPA)* and provide an explanatory analysis. Through extensive experimentation at parameter scales of 130M, 370M, and scaling up to 3B, we have validated the effectiveness and scalability of LPA. Our results show that LPA model can save up to 12.4% in time while achieving an approximate 5% improvement in test perplexity (ppl) and on downstream tasks compared with vanilla Transformer.

Nov 13 (Wed) 16:00-17:30 - Jasmine

FAME: Factual Multi-task Model Editing Benchmark

Li Zeng, Yingyu Shan, Zeming Liu, Jiashu Yao, Yuhang Guo

Large language models (LLMs) embed extensive knowledge and utilize it to perform exceptionally well across various tasks. Nevertheless, outdated knowledge or factual errors within LLMs can lead to misleading or incorrect responses, causing significant issues in practical applications. To rectify the fatal flaw without the necessity for costly model retraining, various model editing approaches have been proposed to correct inaccurate information within LLMs in a cost-efficient way. To evaluate these model editing methods, previous work introduced a series of datasets. However, most of the previous datasets only contain fabricated data in a single format, which diverges from real-world model editing scenarios, raising doubts about their usability in practice. To facilitate the application of model editing in real-world scenarios, we propose the challenge of practicality. To resolve such challenges and effectively enhance the capabilities of LLMs, we present FAME, an authentic, comprehensive, and multi-task dataset, which is designed to enhance the practicality of model editing. We then propose SKEME, a model editing method that uses a novel caching mechanism to ensure synchronization with the real world. The experiments demonstrate that our method performs excellently across various tasks and scenarios, confirming its practicality.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, Reut Tsarfaty

Improvements in language models' capabilities have pushed their applications towards longer contexts, making long-context evaluation and development an active research area. However, many disparate use-cases are grouped together under the umbrella term of "long-context", defined simply by the total length of the model's input, including - for example - Needle-in-a-Haystack tasks, book summarization, and information aggregation. Given their varied difficulty, in this position paper we argue that conflating different tasks by their context length is unproductive. As a community, we require a more precise vocabulary to understand what makes long-context tasks similar or different. We propose to unpack the taxonomy of long-context based on the properties that make them more difficult with longer contexts. We propose two orthogonal axes of difficulty: (I) Diffusion: How hard is it to find the necessary information in the context? (II) Scope: How much necessary information is there to find? We survey the literature on long-context, provide justification for this taxonomy as an informative descriptor, and situate the literature with respect to it. We conclude that the most difficult and interesting settings, whose necessary information is very long and highly diffused within the input, is severely under-explored. By using a descriptive vocabulary and discussing the relevant properties of difficulty in long-context, we can implement more informed research in this area. We call for a careful design of tasks and benchmarks with distinctly long context, taking into account the characteristics that make it qualitatively different from shorter context.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Revisiting the Robustness of Watermarking to Paraphrasing Attacks

Saksham Rastogi, Danish Pruthi

Amidst rising concerns about the internet being proliferated with content generated from language models (LMs), watermarking is seen as a principled way to certify whether text was generated from a model. Many recent watermarking techniques slightly modify the output probabilities of LMs to embed a signal in the generated output that can later be detected. Since early proposals for text watermarking, questions about their robustness to paraphrasing have been prominently discussed. Lately, some techniques are deliberately designed and claimed to be robust to paraphrasing. Particularly, a recent approach trains a model to produce a watermarking signal that is invariant to semantically-similar inputs. However, such watermarking schemes do not adequately account for the ease with which they can be reverse-engineered. We show that with limited access to model generations, we can undo the effects of watermarking and drastically improve the effectiveness of paraphrasing attacks.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Which Programming Language and What Features at Pre-training Stage Affect Downstream Logical Inference Performance?

Fumiya Uchiyama, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo

Recent large language models (LLMs) have demonstrated remarkable generalization abilities in mathematics and logical reasoning tasks. Prior research indicates that LLMs pre-trained with programming language data exhibit high mathematical and reasoning abilities; however, this causal relationship has not been rigorously tested. Our research aims to verify which programming languages and features during pre-training affect logical inference performance. Specifically, we pre-trained decoder-based language models from scratch using datasets from ten programming languages (e.g., Python, C, Java) and three natural language datasets (Wikipedia, Finweb, C4) under identical conditions. Thereafter, we evaluated the trained models in a few-shot in-context learning setting on logical reasoning tasks: FLD and bAbI, which do not require common sense or world knowledge. The results demonstrate that nearly all models trained with programming languages consistently outperform those trained with natural languages, indicating that programming languages contain factors that elicit logic inference performance. In addition, we found that models trained with programming languages exhibit a better ability to follow instructions compared to those trained with natural languages. Further analysis reveals that the depth of Abstract Syntax Trees representing parsed results of programs also affects logical reasoning performance. These findings will offer insights into the essential elements of pre-training for acquiring the foundational abilities of LLMs.

*Nov 13 (Wed) 16:00-17:30 - Jasmine***Jump Starting Bandits with LLM-Generated Prior Knowledge**

Parand A. Alamdar, Yanshuai Cao, Kevin H. Wilson

We present substantial evidence demonstrating the benefits of integrating Large Language Models (LLMs) with a Contextual Multi-Armed Bandit framework. Contextual bandits have been widely used in recommendation systems to generate personalized suggestions based on user-specific contexts. We show that LLMs, pre-trained on extensive corpora rich in human knowledge and preferences, can simulate human behaviours well enough to jump-start contextual multi-armed bandits to reduce online learning regret. We propose an initialization algorithm for contextual bandits by prompting LLMs to produce a pre-training dataset of approximate human preferences for the bandit. This significantly reduces online learning regret and data-gathering costs for training such models. Our approach is validated empirically through two sets of experiments with different bandit setups: one which utilizes LLMs to serve as an oracle and a real-world experiment utilizing data from a conjoint survey experiment.

*Nov 13 (Wed) 16:00-17:30 - Jasmine***ATPO: Automatic Tree-Structured Prompt Optimization**

Sheng Yang, Yurong Wu, Yan Gao, Zineng Zhou, Xiaodi Sun, Bin Benjamin Zhu, Jian-Guang Lou, Zhiming Ding, Anbang Hu, Yuan Fang, Yunson Li, Junyan Chen, Linjun Yang

Prompt engineering is very important to enhance the performance of large language models (LLMs). When dealing with complex issues, prompt engineers tend to distill multiple patterns from examples and inject relevant solutions to optimize the prompts, achieving satisfying results. However, existing automatic prompt optimization techniques are only limited to producing single flow instructions, struggling with handling diverse patterns. In this paper, we present AMP0, an automatic prompt optimization method that can iteratively develop a multi-branched prompt using failure cases as feedback. Our goal is to explore a novel way of structuring prompts with multi-branches to better handle multiple patterns in complex tasks, for which we introduce three modules: Pattern Recognition, Branch Adjustment, and Branch Pruning. In experiments across five tasks, AMP0 consistently achieves the best results. Additionally, our approach demonstrates significant optimization efficiency due to our adoption of a minimal search strategy.

*Nov 13 (Wed) 16:00-17:30 - Jasmine***AlphaExpert: Assigning LoRA Experts Based on Layer Training Quality**

Peijun Qing, Chongyang Gao, Yefan Zhou, Xingjian Diao, Yaogang Yang, Soroush Vosoughi

Parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA), are known to enhance training efficiency in Large Language Models (LLMs). Due to the limited parameters of LoRA, recent studies seek to combine LoRA with Mixture-of-Experts (MoE) to boost performance across various tasks. However, inspired by the observed redundancy in traditional MoE structures, prior studies find that LoRA experts within the MoE architecture also exhibit redundancy, suggesting a need to vary the allocation of LoRA experts across different layers. In this paper, we leverage Heavy-Tailed Self-Regularization (HT-SR) Theory to design a fine-grained allocation strategy. Our analysis reveals that the number of experts per layer correlates with layer training quality, which exhibits significant variability across layers. Based on this, we introduce AlphaLoRA, a theoretically principled and training-free method for allocating LoRA experts to reduce redundancy further. Experiments on three models across ten language processing and reasoning benchmarks demonstrate that AlphaLoRA achieves comparable or superior performance over all baselines. Our code is available at <https://github.com/morelife2017/alphalora>.

*Nov 13 (Wed) 16:00-17:30 - Jasmine***Rethinking the Role of Proxy Rewards in Language Model Alignment**

Sungdong Kim, Minjoon Seo

Learning from human feedback via proxy reward modeling has been studied to align Large Language Models (LLMs) with human values. However, achieving reliable training through that proxy reward model (RM) is not a trivial problem, and its behavior remained as a black-box. In this paper, we study the role of proxy rewards in the LLM alignment via ‘reverse reward engineering’ by composing interpretable features as a white-box reward function. We aim to replicate the ground truth (gold) reward signal by achieving a monotonic relationship between the proxy and gold reward signals after training the model using the proxy reward in reinforcement learning (RL). Our findings indicate that successfully emulating the gold reward requires generating responses that are relevant with enough length to open-ended questions, while also ensuring response consistency in closed-ended questions. Furthermore, resulting models optimizing our devised white-box reward show competitive performances with strong open-source RMs in alignment benchmarks. We highlight its potential usage as a simple but strong reward baseline for the LLM alignment, not requiring explicit human feedback dataset and RM training.

*Nov 13 (Wed) 16:00-17:30 - Jasmine***Mixtue-of-Modules: Reinventing Transformers as Dynamic Assemblies of Modules**

Zhuocheng Gong, Ang Lv, Jian Guan, Wei Wu, Huishuai Zhang, Minlie Huang, Dongyan Zhao, Rui Yan

Is it always necessary to compute tokens from shallow to deep layers in Transformers? The continued success of vanilla Transformers and their variants suggests an undoubted “yes”. In this work, however, we attempt to break the depth-ordered convention by proposing a novel architecture dubbed mixture-of-modules (MoM), which is motivated by an intuition that any layer, regardless of its position, can be used to compute a token as long as it possesses the needed processing capabilities. The construction of MoM starts from a finite set of modules defined by multi-head attention and feed-forward networks, each distinguished by its unique parameterization. Two routers then iteratively select attention modules and feed-forward modules from the set to process a token. The selection dynamically expands the computation graph in the forward pass of the token, culminating in an assembly of modules. We show that MoM provides not only a unified framework for Transformers and their numerous variants but also a flexible and learnable approach for reducing redundancy in Transformer parameterization. We pre-train various MoMs using OpenWebText. Empirical results demonstrate that MoMs, of different sizes, consistently outperform vanilla transformers. More interestingly, after removing 50% of the multi-head attention modules and 25% of the feed-forward modules, an

MoM model still holds comparable performance. Additionally, by properly adjusting the number of modules and compressing the model depth, one can have an MoM that achieves comparable performance to GPT-2 (774M) while saving 16% TFLOPs and 42% memory usage during forward computation.

Nov 13 (Wed) 16:00-17:30 - Jasmine

GPT-4 Jailbreaks Itself with Near-Perfect Success Using Self-Explanation

Govind Ramesh, Yao Dou, Wei Xu

Research on jailbreaking has been valuable for testing and understanding the safety and security issues of large language models (LLMs). In this paper, we introduce Iterative Refinement Induced Self-Jailbreak (IRIS), a novel approach that leverages the reflective capabilities of LLMs for jailbreaking with only black-box access. Unlike previous methods, IRIS simplifies the jailbreaking process by using a single model as both the attacker and target. This method first iteratively refines adversarial prompts through self-explanation, which is crucial for ensuring that even well-aligned LLMs obey adversarial instructions. IRIS then rates and enhances the output given the refined prompt to increase its harmfulness. We find that IRIS achieves jailbreak success rates of 98% on GPT-4, 92% on GPT-4 Turbo, and 94% on Llama-3.1-70B in under 7 queries. It significantly outperforms prior approaches in automatic, black-box, and interpretable jailbreaking, while requiring substantially fewer queries, thereby establishing a new standard for interpretable jailbreaking methods.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Initialization of Large Language Models via Reparameterization to Mitigate Loss Spikes

Kosuke Nishida, Kyosuke Nishida, Kuniko Saito

Loss spikes, a phenomenon in which the loss value diverges suddenly, is a fundamental issue in the pre-training of large language models. This paper supposes that the non-uniformity of the norm of the parameters is one of the causes of loss spikes. Here, in training of neural networks, the scale of the gradients is required to be kept constant throughout the layers to avoid the vanishing and exploding gradients problem. However, to meet these requirements in the Transformer model, the norm of the model parameters must be non-uniform, and thus, parameters whose norm is smaller are more sensitive to the parameter update. To address this issue, we propose a novel technique, weight scaling as reparameterization (WeSaR). WeSaR introduces a gate parameter per parameter matrix and adjusts it to the value satisfying the requirements. Because of the gate parameter, WeSaR sets the norm of the original parameters uniformly, which results in stable training. Experimental results with the Transformer decoders consisting of 130 million, 1.3 billion, and 13 billion parameters showed that WeSaR stabilizes and accelerates training and that it outperformed compared methods including popular initialization methods.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Instruction Fine-Tuning: Does Prompt Loss Matter?

Matheo Huerta-Enochian, Seung Young Ko

We present a novel study analyzing the effects of various prompt loss token weights (PLW) for supervised instruction fine-tuning (SIFT). While prompt-masking (PLW = 0) is common for SIFT, some fine-tuning APIs support fractional PLWs and suggest that using a small non-zero PLW can help stabilize learning when fine-tuning on short-completion data. However, there has never been a study confirming this claim, and OpenAI, a major cloud-based SIFT provider, recently removed this parameter from their fine-tuning API. We found that performance of models fine-tuned on short-completion data had a statistically-significant negative quadratic relationship with PLW. Using small values (0.01–0.5) of PLW produced better results on multiple-choice and short-generation benchmarks (outperforming models fine-tuned on long-completion data) while large values (1.0) of PLW produced better results on long-generation benchmarks. We explained this effect and verified its importance through additional experiments. This research serves as a warning to API providers about the importance of providing a PLW parameter for SIFT.

Nov 13 (Wed) 16:00-17:30 - Jasmine

From Test-Taking to Test-Making: Examining LLM Authoring of Commonsense Assessment Items

Melissa Roemmele, Andrew Gordon

LLMs can now perform a variety of complex writing tasks. They also excel in answering questions pertaining to natural language inference and commonsense reasoning. Composing these questions is itself a skilled writing task, so in this paper we consider LLMs as authors of commonsense assessment items. We prompt LLMs to generate items in the style of a prominent benchmark for commonsense reasoning, the Choice of Plausible Alternatives (COPA). We examine the outcome according to analyses facilitated by the LLMs and human annotation. We find that LLMs that succeed in answering the original COPA benchmark are also more successful in authoring their own items.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Evolutionary Contrastive Distillation for Language Model Alignment

Julian Katz-Samuels, Zheng Li, Hyekun Yun, Priyanka Nigam, Yi Xu, Vaclav Petricek, Bing Yin, Trishul Chilimbi

The ability of large language models (LLMs) to execute complex instructions is essential for their real-world applications. However, several recent studies indicate that LLMs struggle with challenging instructions. In this paper, we propose Evolutionary Contrastive Distillation (ECD), a novel method for generating high-quality synthetic preference data designed to enhance the complex instruction-following capability of language models. ECD generates data that specifically illustrates the difference between a response that successfully follows a set of complex instructions and a response that is high-quality, but nevertheless makes some subtle mistakes. This is done by prompting LLMs to progressively evolve simple instructions to more complex instructions. When the complexity of an instruction is increased, the original successful response to the original instruction becomes a "hard negative" response for the new instruction, mostly meeting requirements of the new instruction, but barely missing one or two. By pairing a good response with such a hard negative response, and employing contrastive learning algorithms such as DPO, we improve language models' ability to follow complex instructions. Empirically, we observe that our method yields a 7B model that exceeds the complex instruction-following performance of current SOTA 7B models and is competitive even with open-source 70B models.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Private prediction for large-scale synthetic text generation

Kareem Amin, Alex Bie, Weiwei Kong, Alexey Kurakin, Natalia Ponomareva, Umar Syed, Andreas Terzis, Sergei Vassilivitskii

We present an approach for generating differentially private synthetic text using large language models (LLMs), via private prediction. In the private prediction framework, we only require the output synthetic data to satisfy differential privacy guarantees. This is in contrast to approaches that train a generative model on potentially sensitive user-supplied source data and seek to ensure the model itself is safe to release. We prompt a pretrained LLM with source data, but ensure that next-token predictions are made with differential privacy guarantees. Previous work in this paradigm reported generating a small number of examples (<10) at reasonable privacy levels, an amount of data that is useful only for downstream in-context learning or prompting. In contrast, we make changes that allow us to generate thousands of high-quality synthetic data points, greatly expanding the set of potential applications. Our improvements come from an improved privacy analysis and a better private selection mechanism, which makes use of the equivalence between the softmax layer for sampling tokens in LLMs and the exponential mechanism. Furthermore, we introduce a novel use of public predictions via the sparse vector technique, in which we do not pay privacy costs for tokens that are predictable without sensitive data; we find this to be particularly effective for structured data.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Better Call SAUL: Fluent and Consistent Language Model Editing with Generation Regularization

Mingyang Wang, Lukas Lange, Heike Adel, Jannik Strötgen, Hinrich Schütze

To ensure large language models contain up-to-date knowledge, they need to be updated regularly. However, model editing is challenging as it might also affect knowledge that is unrelated to the new data. State-of-the-art methods identify parameters associated with specific knowledge and then modify them via direct weight updates. However, these locate-and-edit methods suffer from heavy computational overhead and lack theoretical validation. In contrast, directly fine-tuning the model on requested edits affects the model's behavior on unrelated knowledge, and significantly damages the model's generation fluency and consistency. To address these challenges, we propose SAUL, a streamlined model editing method that uses sentence concatenation with augmented random facts for generation regularization. Evaluations on three model editing benchmarks show that saul is a practical and reliable solution for model editing outperforming state-of-the-art methods while maintaining generation quality and reducing computational overhead.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Tutor-ICL: Guiding Large Language Models for Improved In-Context Learning Performance

Ikhyan Cho, Gauel Kwon, Julian Hockenmaier

There has been a growing body of work focusing on the in-context learning (ICL) abilities of large language models (LLMs). However, it is an open question how effective ICL can be. This paper presents Tutor-ICL, a simple prompting method for classification tasks inspired by how effective instructors might engage their students in learning a task. Specifically, we propose presenting exemplar answers in a *comparative format* rather than the traditional single-answer format. We also show that including the test instance before the exemplars can improve performance, making it easier for LLMs to focus on relevant exemplars. Lastly, we include a summarization step before attempting the test, following a common human practice. Experiments on various classification tasks, conducted across both decoder-only LLMs (Llama 2, 3) and encoder-decoder LLMs (Flan-T5-XL, XXL), show that Tutor-ICL consistently boosts performance, achieving up to a 13.76% increase in accuracy.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Beyond Fine-tuning: Unleashing the Potential of Continuous Pretraining for Clinical LLMs.

Clement Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco AF Pimentel, Shadab Khan

Large Language Models (LLMs) have demonstrated significant potential in revolutionizing clinical applications. In this study, we investigate the efficacy of four techniques in adapting LLMs for clinical use-cases: continuous pretraining, instruct fine-tuning, NEFTune, and prompt engineering. We employ these methods on Mistral 7B and Mixtral 8x7B models, leveraging a large-scale clinical pretraining dataset of 50 billion tokens and an instruct fine-tuning dataset of 500 million tokens. Our evaluation across various clinical tasks reveals nuanced insights. While continuous pretraining beyond 250 billion tokens yields marginal improvements, instruct fine-tuning emerges as a more influential factor. Notably, NEFTune, designed primarily to enhance generation quality, surprisingly demonstrates additional gains on our benchmark. These findings underscore the importance of tailoring fine-tuning strategies and exploring innovative techniques to optimize LLM performance in the clinical domain.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Selection-p: Self-Supervised Task-Agnostic Prompt Compression for Faithfulness and Transferability

Tsz Ting Chung, Leyang Cui, Lemao Liu, Xinting Huang, Shuming Shi, Dit-Yan Yeung

Large Language Models (LLMs) have demonstrated impressive capabilities in a wide range of natural language processing tasks when leveraging in-context learning. To mitigate the additional computational and financial costs associated with in-context learning, several prompt compression methods have been proposed to compress the in-context learning prompts. Despite their success, these methods face challenges with transferability due to model-specific compression, or rely on external training data, such as GPT-4. In this paper, we investigate the ability of LLMs to develop a unified compression method that discretizes uninformative tokens, utilizing a self-supervised pre-training technique. By introducing a small number of parameters during the continual pre-training, the proposed Selection-p produces a probability for each input token, indicating whether to preserve or discard it. Experiments show Selection-p achieves state-of-the-art performance across numerous classification tasks, achieving compression rates of up to 10 times while experiencing only a marginal 0.8% decrease in performance. Moreover, it exhibits superior transferability to different models compared to prior work. Additionally, we further analyze how Selection-p helps maintain performance on in-context learning with long contexts.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Prompt-Based Bias Calibration for Better Zero/Few-Shot Learning of Language Models

Kang He, Yinghan Long, Kaushik Roy

Prompt-based learning is susceptible to intrinsic bias present in pre-trained language models (LMs), leading to sub-optimal performance in prompt-based zero/few-shot settings. In this work, we propose a null-input prompting method to calibrate intrinsic bias encoded in pre-trained LMs. Different from prior efforts that address intrinsic bias primarily for social fairness and often involve excessive computational cost, our objective is to explore enhancing LMs' performance in downstream zero/few-shot learning while emphasizing the efficiency of intrinsic bias calibration. Specifically, we leverage a diverse set of auto-selected null-meaning inputs generated from GPT-4 to probe intrinsic bias of pre-trained LMs. Utilizing the bias-reflected probability distribution, we formulate a distribution disparity loss for bias calibration, where we exclusively update bias parameters (0.1% of total parameters) of LMs towards equal probability distribution. Experimental results show that the calibration promotes an equitable starting point for LMs while preserving language modeling abilities. Across a wide range of datasets, including sentiment analysis and topic classification, our method significantly improves zero/few-shot learning performance of LMs for both in-context learning and prompt-based fine-tuning (on average 9% and 2%, respectively).

Nov 13 (Wed) 16:00-17:30 - Jasmine

Auto-Evolve: Enhancing Large Language Model's Performance via Self-Reasoning Framework

Krishna Aswani, Hailin Lu, Pranav Patankar, Priya Dhatwani, Xue Tan, Jayant Ganeshmohan, Simon Lacasse

Recent advancements in prompt engineering strategies, such as Chain-of-Thought (CoT) and Self-Discover, have demonstrated significant potential in improving the reasoning abilities of Large Language Models (LLMs). However, these state-of-the-art (SOTA) prompting strategies rely on a fixed set of static seed reasoning modules like "think step by step" / "break down this problem" intended to simulate human approach to problem-solving. This constraint limits the flexibility of models in tackling diverse problems effectively. In this paper, we introduce Auto-Evolve, a novel framework that enables LLMs to self-create dynamic reasoning modules and downstream action plan, resulting in significant improvements over current SOTA methods. We evaluate Auto-Evolve on the challenging BigBench-Hard (BBH) dataset with Claude 2.0, Claude 3 Sonnet, Mistral Large, and GPT-4, where it consistently outperforms the SOTA prompt strategies. Auto-Evolve outperforms CoT by up to 10.4% and on an average by 7% across these four models. Our framework introduces two innovations: a) Auto-Evolve dynamically generates reasoning modules for each task while aligning with human reasoning paradigm, thus eliminating the need for predefined templates. b) An iterative refinement component, that incrementally refines instruction guidance for LLMs and helps boost performance

by average 2.8% compared to doing it in a single step.

Nov 13 (Wed) 16:00-17:30 - Jasmine

DrAttack: Prompt Decomposition and Reconstruction Makes Powerful LLMs Jailbreakers

Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, Cho-Jui Hsieh

Safety-aligned Large Language Models (LLMs) are still vulnerable to some manual and automated jailbreak attacks, which adversarially trigger LLMs to output harmful content. However, existing jailbreaking methods usually view a harmful prompt as a whole but they are not effective at reducing LLMs' attention on combinations of words with malice, which well-aligned LLMs can easily reject. This paper discovers that decomposing a malicious prompt into separated sub-prompts can effectively reduce LLMs' attention on harmful words by presenting them to LLMs in a fragmented form, thereby addressing these limitations and improving attack effectiveness. We introduce an automatic prompt Decomposition and Reconstruction framework for jailbreaking Attack (DrAttack). DrAttack consists of three key components: (a) 'Decomposition' of the original prompt into sub-prompts, (b) 'Reconstruction' of these sub-prompts implicitly by In-Context Learning with semantically similar but benign reassembling example, and (c) 'Synonym Search' of sub-prompts, aiming to find sub-prompts' synonyms that maintain the original intent while jailbreaking LLMs. An extensive empirical study across multiple open-source and closed-source LLMs demonstrates that, with fewer queries, DrAttack obtains a substantial gain of success rate on powerful LLMs over prior SOTA attackers. Notably, the success rate of 80% on GPT-4 surpassed previous art by 65%. Code and data are made publicly available at <https://turningpoint-ai.github.io/DrAttack/>.

Nov 13 (Wed) 16:00-17:30 - Jasmine

POSIX: A Prompt Sensitivity Index For Large Language Models

Anwya Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhattacharjee, Tammooy Chakraborty

Despite their remarkable capabilities, Large Language Models (LLMs) are found to be surprisingly sensitive to minor variations in prompts, often generating significantly divergent outputs in response to minor variations in the prompts, such as spelling errors, alteration of wording or the prompt template. However, while assessing the quality of an LLM, the focus often tends to be solely on its performance on downstream tasks, while very little to no attention is paid to prompt sensitivity. To fill this gap, we propose POSIX, a novel PrOmpT Sensitivity IndeX as a reliable measure of prompt sensitivity, thereby offering a more comprehensive evaluation of LLM performance. The key idea behind POSIX is to capture the relative change in loglikelihood of a given response upon replacing the corresponding prompt with a different intent-preserving prompt. We provide thorough empirical evidence demonstrating the efficacy of POSIX in capturing prompt sensitivity and subsequently use it to measure and thereby compare prompt sensitivity of various open source LLMs. We find that merely increasing the parameter count or instruction tuning does not necessarily reduce prompt sensitivity whereas adding some few-shot exemplars, even just one, almost always leads to significant decrease in prompt sensitivity. We also find that alterations to prompt template lead to the highest sensitivity in the case of MCQ type tasks, whereas paraphrasing results in the highest sensitivity in open-ended generation tasks. The code for reproducing our results is open-sourced at <https://github.com/kowndinya-renduchintala/POSIX>.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Unveiling Narrative Reasoning Limits of Large Language Models with Trope in Movie Synopses

Hung-Ting Su, Ya-Ching Hsu, Xudong Lin, Xiang-Qian Shi, Yulei Niu, Han-Yuan Hsu, Hung-yi Lee, Winston H. Hsu

Large language models (LLMs) equipped with chain-of-thoughts (CoT) prompting have shown significant multi-step reasoning capabilities in factual content like mathematics, commonsense, and logic. However, their performance in narrative reasoning, which demands greater abstraction capabilities, remains unexplored. This study utilizes tropes in movie synopses to assess the abstract reasoning abilities of state-of-the-art LLMs and uncovers their low performance. We introduce a trope-wise querying approach to address these challenges and boost the F1 score by 11.8 points. Moreover, while prior studies suggest that CoT enhances multi-step reasoning, this study shows CoT can cause hallucinations in narrative content, reducing GPT-4's performance. We also introduce an Adversarial Injection method to embed trope-related text tokens into movie synopses without explicit tropes, revealing CoT's heightened sensitivity to such injections. Our comprehensive analysis provides insights for future research directions.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for Large Language Models

Jie Chen, Yuqeng Zhang, Bingbing Wang, Xin Zhao, Ji-Rong Wen

Synthetic data has been proposed as a solution to address the issue of high-quality data scarcity in the training of large language models (LLMs). Studies have shown that synthetic data can effectively improve the performance of LLMs on downstream benchmarks. However, despite its potential benefits, our analysis suggests that there may be inherent flaws in synthetic data. The uniform format of synthetic data can lead to pattern overfitting and cause significant shifts in the output distribution, thereby reducing the model's instruction-following capabilities. Our work delves into these specific flaws associated with question-answer (Q-A) pairs, a prevalent type of synthetic data, and presents a method based on unlearning techniques to mitigate these flaws. The empirical results demonstrate the effectiveness of our approach, which can reverse the instruction-following issues caused by pattern overfitting without compromising performance on benchmarks at relatively low cost. Our work has yielded key insights into the effective use of synthetic data, aiming to promote more robust and efficient LLM training.

Nov 13 (Wed) 16:00-17:30 - Jasmine

The Student Data Paradox: Examining the Regressive Side Effects of Training LLMs for Personalized Learning

Shashank Sonkar, Naiping Liu, Richard Baraniuk

The pursuit of personalized education has led to the integration of Large Language Models (LLMs) in developing intelligent tutoring systems. To better understand and adapt to individual student needs, including their misconceptions, LLMs need to be trained on extensive datasets of student-tutor dialogues. Our research uncovers a fundamental challenge in this approach: the "Student Data Paradox". This paradox emerges when LLMs, trained on student data to understand learner behavior, inadvertently compromise their own factual knowledge and reasoning abilities. We investigate this paradox by training state-of-the-art language models on student-tutor dialogue datasets and evaluating their performance across multiple benchmarks. These benchmarks assess various aspects of language model capabilities, including reasoning, truthfulness, and common sense understanding. Our findings reveal significant declines in the models' performance across these diverse benchmarks, indicating a broad impact on their capabilities when trained to model student behavior. Our research makes two primary contributions: (1) empirical demonstration of the Student Data Paradox through quantitative analysis of model performance, and (2) introduction of "hallucination tokens" as a mitigation strategy. These tokens, while improving performance, highlight the persistent challenge of balancing accurate student behavior modeling with maintaining the LLM's integrity as an educational tool. This study emphasizes the need for innovative solutions to reconcile the conflicting goals of faithfully understanding diverse student cognition while preserving the model's ability to provide accurate information and guidance.

Nov 13 (Wed) 16:00-17:30 - Jasmine

On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization

Yong Lin, Skyler Seto, Maartje Ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, Tong Zhang
Reinforcement Learning from Human Feedback (RLHF) is an effective approach for aligning language models to human preferences. Central

to RLHF is learning a reward function for scoring human preferences. Two main approaches for learning a reward model are 1) training an Explicit Reward Model (EXRM) as in RLHF, and 2) using an implicit reward learned from preference data through methods such as Direct Preference Optimization (DPO). Prior work has shown that the implicit reward model of DPO (denoted as DPORM) can approximate an EXRM on the limit infinite samples. However, it is unclear how effective is DPORM in practice. DPORM's effectiveness directly implies the optimality of learned policy of DPO and also has practical implication for more advanced alignment methods, such as iterative DPO. We compare the accuracy at distinguishing preferred and rejected answers using both DPORM and EXRM. Our findings indicate that even though DPORM can fit the training dataset, it generalizes less effective than EXRM, especially when the validation datasets contain distributional shifts. Across five out-of-distribution settings, DPORM has a mean drop in accuracy of 3% and a maximum drop of 7%. These findings highlight that DPORM has limited generalization ability and substantiates the integration of an explicit reward model in iterative DPO approaches.

Nov 13 (Wed) 16:00-17:30 - Jasmine

FactAlign: Long-form Factuality Alignment of Large Language Models

Chao-Wei Huang, Yun-Nung Chen

Large language models have demonstrated significant potential as the next-generation information access engines. However, their reliability is hindered by issues of hallucination and generating non-factual content. This is particularly problematic in long-form responses, where assessing and ensuring factual accuracy is complex. In this paper, we address this gap by proposing FactAlign, a novel alignment framework designed to enhance the factuality of LLMs' long-form responses while maintaining their helpfulness. We introduce fKTO, a fine-grained, sentence-level alignment algorithm that extends the Kahaneman-Tversky Optimization (KTO) alignment method. Leveraging recent advances in automatic factuality evaluation, FactAlign utilizes fine-grained factuality assessments to guide the alignment process. Our experiments on open-domain prompts and information-seeking questions demonstrate that FactAlign significantly improves the factual accuracy of LLM responses while also improving their helpfulness. Further analyses identify that FactAlign is capable of training LLMs to provide more information without losing factual precision, thus improving the factual F1 score. Our source code, datasets, and trained models are publicly available at <https://github.com/MiuLab/FactAlign>

Multimodality and Language Grounding to Vision, Robotics and Beyond 4

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

TinyChart: Efficient Chart Understanding with Program-of-Thoughts Learning and Visual Token Merging

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, Fei Huang

Charts are important for presenting and explaining complex data relationships. Recently, multimodal large language models (MLLMs) have shown remarkable capabilities in chart understanding. However, the sheer size of these models limits their use in resource-constrained environments. In this paper, we present TinyChart, an efficient MLLM for chart understanding with only 3B parameters. TinyChart overcomes two key challenges in efficient chart understanding: (1) reduce the burden of learning numerical computations through Program-of-Thoughts (PoT) learning, which trains the model to generate Python programs for numerical calculations, and (2) reduce lengthy vision feature sequences through Vision Token Merging, which gradually merges most similar vision tokens. Extensive experiments demonstrate that our 3B TinyChart achieves SOTA performance on various chart understanding benchmarks including ChartQA, Chart-to-Text, Chart-to-Table, OpenCQA, and ChartX. It outperforms several chart-understanding MLLMs with up to 13B parameters, and close-sourced MLLM GPT-4V on ChartQA, with higher throughput during inference due to a smaller model scale and more efficient vision encoding.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianhang Zhang, Xinyu Dai

Multimodal large language models (MLLMs) have attracted increasing attention in the past few years, but they may still generate descriptions that include objects not present in the corresponding images, a phenomenon known as object hallucination. To eliminate hallucinations, existing methods manually annotate paired responses with and without hallucinations, and then employ various alignment algorithms to improve the alignment capability between images and text. However, they not only demand considerable computation resources during the finetuning stage but also require expensive human annotation to construct paired data needed by the alignment algorithms. To address these issues, we propose an efficient fine-grained unlearning framework (EFUF), which performs gradient ascent utilizing three tailored losses to eliminate hallucinations without paired data. Extensive experiments show that our method consistently reduces hallucinations while preserving the generation quality with modest computational overhead. Our code and datasets will be publicly available.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Towards Difficulty-Agnostic Efficient Transfer Learning for Vision-Language Models

Yongjin Yang, Jongwoo Ko, Se-Young Yun

Vision-language models (VLMs) like CLIP have demonstrated remarkable applicability across a variety of downstream tasks, including zero-shot image classification. Recently, the use of prompt or adapters for efficient transfer learning (ETL) has gained significant attention for effectively adapting to downstream tasks. However, previous studies have overlooked the challenge of varying transfer difficulty of downstream tasks. In this paper, we empirically analyze how each ETL method behaves with respect to transfer difficulty. Our observations indicate that utilizing vision prompts and text adapters is crucial for adaptability and generalizability in domains with high difficulty. Also, by applying an adaptive ensemble approach that integrates task-adapted VLMs with pre-trained VLMs and strategically leverages more general knowledge in low-difficulty and less in high-difficulty domains, we consistently enhance performance across both types of domains. Based on these observations, we propose an adaptive ensemble method that combines visual prompts and text adapters with pre-trained VLMs, tailored by transfer difficulty, to achieve optimal performance for any target domain. Upon experimenting with extensive benchmarks, our method consistently outperforms all baselines, particularly on unseen tasks, demonstrating its effectiveness.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

By My Eyes: Grounding Multimodal Large Language Models with Sensor Data via Visual Prompting

Hyungsun Yoon, Binyamin Aschalew Tolera, Taesik Gong, Kimin Lee, Sung-Ju Lee

Large language models (LLMs) have demonstrated exceptional abilities across various domains. However, utilizing LLMs for ubiquitous sensing applications remains challenging as existing text-prompt methods show significant performance degradation when handling long sensor data sequences. In this paper, we propose a visual prompting approach for sensor data using multimodal LLMs (MLLMs). Specifically, we design a visual prompt that directs MLLMs to utilize visualized sensor data alongside descriptions of the target sensory task. Additionally, we introduce a visualization generator that automates the creation of optimal visualizations tailored to a given sensory task, eliminating the need for prior task-specific knowledge. We evaluated our approach on nine sensory tasks involving four sensing modalities, achieving an average

of 10% higher accuracy compared to text-based prompts and reducing token costs by 15.8 times. Our findings highlight the effectiveness and cost-efficiency of using visual prompts with MLLMs for various sensory tasks. The source code is available at <https://github.com/diamond264/ByMyEyes>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification

Gregor Geigle, Radu Timofte, Goran Glava

Recent Large Vision-Language Models (LVLMs) demonstrate impressive abilities on numerous image understanding and reasoning tasks. The task of fine-grained object classification (e.g., distinction between *animal species*), however, has been probed insufficiently, despite its downstream importance. We fill this evaluation gap by creating FOCI (Fine-grained Object Classification), a difficult multiple-choice benchmark for fine-grained object classification, from existing object classification datasets: (1) multiple-choice avoids ambiguous answers associated with casting classification as open-ended QA task; (2) we retain classification difficulty by mining negative labels with a CLIP model. FOCI complements five popular classification datasets with four domain-specific subsets from ImageNet-21k. We benchmark 12 public LVLMs on bench and show that it tests for a *complementary skill* to established image understanding and reasoning benchmarks. Crucially, CLIP models exhibit dramatically better performance than LVLMs. Since the image encoders of LVLMs come from these CLIP models, this points to inadequate alignment for fine-grained object distinction between the encoder and the LLM and warrants (pre)training data with more fine-grained annotation. We release our code at [ANONYMIZED](#).

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Does Object Grounding Really Reduce Hallucination of Large Vision-Language Models?

Gregor Geigle, Radu Timofte, Goran Glava

Large vision-language models (LVLMs) have recently dramatically pushed the state of the art in image captioning and many image understanding tasks (e.g., visual question answering). LVLMs, however, often *hallucinate* and produce captions that mention concepts that cannot be found in the image. These hallucinations erode the trustworthiness of LVLMs and are arguably among the main obstacles to their ubiquitous adoption. Recent work suggests that addition of grounding objectives—those that explicitly align image regions or objects to text spans—reduces the amount of LVLM hallucination. Although intuitive, this claim is not empirically justified as the reduction effects have been established, we argue, with flawed evaluation protocols that (i) rely on data (i.e., MSCOCO) that has been extensively used in LVLM training and (ii) measure hallucination via question answering rather than open-ended caption generation. In this work, in contrast, we offer the first systematic analysis of the effect of fine-grained object grounding on LVLM hallucination under an evaluation protocol that more realistically captures LVLM hallucination in open generation. Our extensive experiments over three backbone LLMs reveal that grounding objectives have little to no effect on object hallucination in open caption generation.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Decompose and Compare Consistency: Measuring VLMs' Answer Reliability via Task-Decomposition Consistency Comparison

Qian Yang, Weixiang Yan, Aishwarya Agrawal

Despite tremendous advancements, current state-of-the-art Vision-Language Models (VLMs) are still far from perfect. They tend to hallucinate and may generate biased responses. In such circumstances, having a way to assess the reliability of a given response generated by a VLM is quite useful. Existing methods, such as estimating uncertainty using answer likelihoods or prompt-based confidence generation, often suffer from overconfidence. Other methods focus self-consistency comparison but are affected by confirmation biases. To alleviate these, we propose Decompose and Compare Consistency (DeCC) for reliability measurement. By comparing the consistency between the direct answer generated using the VLM's internal reasoning process, and the indirect answers obtained by decomposing the question into sub-questions and reasoning over the sub-answers produced by the VLM, DeCC measures the reliability of VLM's direct answer. Experiments across six vision-language tasks with three VLMs show DeCC's reliability estimation achieves better correlation with task accuracy compared to the existing methods.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Teaching Embodied Reinforcement Learning Agents: Informativeness and Diversity of Language Use

Jiajun Xi, Yinyong He, Jianing Yang, Yimpei Dai, Joyce Chai

In real-world scenarios, it is desirable for embodied agents to have the ability to leverage human language to gain explicit or implicit knowledge for learning tasks. Despite recent progress, most previous approaches adopt simple low-level instructions as language inputs, which may not reflect natural human communication. We expect human language to be informative (i.e., providing feedback on agents' past behaviors and offering guidance on achieving their future goals) and diverse (i.e., encompassing a wide range of expressions and style nuances). To enable flexibility of language use in teaching agents tasks, this paper studies different types of language inputs in facilitating reinforcement learning (RL) embodied agents. More specifically, we examine how different levels of language informativeness and diversity impact agent learning and inference. Our empirical results based on four RL benchmarks demonstrate that agents trained with diverse and informative language feedback can achieve enhanced generalization and fast adaptation to new tasks. These findings highlight the pivotal role of language use in teaching embodied agents new tasks in an open world.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

From Coarse to Fine: Impacts of Feature-Preserving and Feature-Compressing Connectors on Perception in Multimodal Models

Junyan Lin, Haoran Chen, Dawei Zhu, Xiaoyu Shen

In recent years, multimodal large language models (MLLMs) have attracted widespread attention from both industry and academia. Based on the integration position, MLLMs can be categorized into external and internal fusion architectures, with the former being more predominant. However, there remains considerable debate on how to construct the optimal external fusion MLLM architecture, especially regarding the performance of different connectors on tasks with varying granularities. This paper systematically investigates the impact of connectors on MLLM performance. Specifically, we classify connectors into feature-preserving and feature-compressing types. Utilizing a unified classification standard, we categorize sub-tasks from three comprehensive benchmarks, MMBench, MMF, and SEED-Bench, into three task types: coarse-grained perception, fine-grained perception, and reasoning, and evaluate the performance from this perspective. Our findings reveal significant performance differences between different types of connectors across various tasks, offering essential guidance for MLLM architecture design and advancing the understanding of MLLM architecture optimization.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, Li Yuan

Large Vision-Language Model (LVLM) has enhanced the performance of various downstream tasks in visual-language understanding. Most existing approaches encode images and videos into separate feature spaces, which are then fed as inputs to large language models. However, due to the lack of unified tokenization for images and videos, namely misalignment before projection, it becomes challenging for a Large Language Model (LLM) to learn multi-modal interactions from several poor projection layers. In this work, we unify visual representation into the language feature space to advance the foundational LLM towards a unified LVLM. As a result, we establish a simple but robust LVLM

baseline, Video-LLaVA, which learns from a mixed dataset of images and videos, mutually enhancing each other. As a result, Video-LLaVA outperforms Video-ChaiGPT by 5.8%, 9.9%, 18.6%, and 10.1% on MSRTT, MSVD, TGF, and ActivityNet, respectively. Additionally, our Video-LLaVA also achieves superior performances on a broad range of 9 image benchmarks. Notably, extensive experiments demonstrate that Video-LLaVA mutually benefits images and videos within a unified visual representation, outperforming models designed specifically for images or videos. We aim for this work to provide modest insights into the multi-modal inputs for the LLM.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

EPO: Hierarchical LLM Agents with Environment Preference Optimization

Qi Zhao, Haotian Fu, Chen Sun, George Konidaris

Long-horizon decision-making tasks present significant challenges for LLM-based agents due to the need for extensive planning over multiple steps. In this paper, we propose a hierarchical framework that decomposes complex tasks into manageable subgoals, utilizing separate LLMs for subgoal prediction and low-level action generation. To address the challenge of creating training signals for unannotated datasets, we develop a reward model that leverages multimodal environment feedback to automatically generate reward signals. We introduce Environment Preference Optimization (EPO), a novel method that generates preference signals from the environment's feedback and uses them to train LLM-based agents. Extensive experiments on ALFRED demonstrate the state-of-the-art performance of our framework, achieving first place on the ALFRED public leaderboard and showcasing its potential to improve long-horizon decision-making in diverse environments.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

MIND: Multimodal Shopping Intention Distillation from Large Vision-language Models for E-commerce Purchase Understanding

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao JING, Tianqiang Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Long Chen, Yangqiu Song

Improving user experience and providing personalized search results in E-commerce platforms heavily rely on understanding purchase intention. However, existing methods for acquiring large-scale intentions bank on distilling large language models with human annotation for verification. Such an approach tends to generate product-centric intentions, overlook valuable visual information from product images, and incurs high costs for scalability. To address these issues, we introduce MIND, a multimodal framework that allows Large Vision-Language Models (LVLMs) to infer purchase intentions from multimodal product metadata and prioritize human-centric ones. Using Amazon Review data, we apply MIND and create a multimodal intention knowledge base, which contains 1,264,441 intentions derived from 126,142 co-buy shopping records across 107,215 products. Extensive human evaluations demonstrate the high plausibility and typicality of our obtained intentions and validate the effectiveness of our distillation framework and filtering mechanism. Further experiments reveal the positive downstream benefits that MIND brings to intention comprehension tasks and highlight the importance of multimodal generation and role-aware filtering. Additionally, MIND shows robustness to different prompts and superior generation quality compared to previous methods.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

mDPO: Conditional Preference Optimization for Multimodal Large Language Models

Fei Wang, Wenzuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, Muhan Chen

Direct preference optimization (DPO) has shown to be an effective method for large language model (LLM) alignment. Recent works have attempted to apply DPO to multimodal scenarios but have found it challenging to achieve consistent improvement. Through a comparative experiment, we identify the unconditional preference problem in multimodal preference optimization, where the model overlooks the image condition. To address this problem, we propose mDPO, a multimodal DPO objective that prevents the over-prioritization of language-only preferences by also optimizing image preference. Moreover, we introduce a reward anchor that forces the reward to be positive for chosen responses, thereby avoiding the decrease in their likelihood in intrinsic problem of relative preference optimization. Experiments on two multimodal LLMs of different sizes and three widely used benchmarks demonstrate that mDPO effectively addresses the unconditional preference problem in multimodal preference optimization and significantly improves model performance, particularly in reducing hallucination.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Pelican: Correcting Hallucination in Vision-LLMs via Claim Decomposition and Program of Thought Verification

Pritish Sahu, Karan Sikka, Ajay Divakaran

Large Visual Language Models (LVLMs) struggle with hallucinations in visual instruction following task(s). These issues hinder their trustworthiness and real-world applicability. We propose Pelican – a novel framework designed to detect and mitigate hallucinations through claim verification. Pelican first decomposes the visual claim into a chain of sub-claims based on first-order predicates. These sub-claims consists of (predicate, question) pairs and can be conceptualized as nodes of a computational graph. We then use Program-of-Thought prompting to generate Python code for answering these questions through flexible composition of external tools. Pelican improves over prior work by introducing (1) intermediate variables for precise grounding of object instances, and (2) shared computation for answering the sub-question to enable adaptive corrections and inconsistency identification. We finally use reasoning abilities of LLM to verify the correctness of the the claim by considering the consistency and confidence of the (question, answer) pairs from each sub-claim. Our experiments demonstrate consistent performance improvements over various baseline LVLMs and existing hallucination mitigation approaches across several benchmarks.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Efficient Temporal Extrapolation of Multimodal Large Language Models with Temporal Grounding Bridge

Yuxuan Wang, Yuezian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, Yang Liu, Zilong Zheng

Despite progress in multimodal large language models (MLLMs), the challenge of interpreting long-form videos in response to linguistic queries persists, largely due to the inefficiency in temporal grounding and limited pre-trained context window size. In this work, we introduce Temporal Grounding Bridge (TGB), a novel framework that bootstraps MLLMs with advanced temporal grounding capabilities and broadens their contextual scope. Our framework significantly enhances the temporal capabilities of current MLLMs through three key innovations: an efficient multi-span temporal grounding algorithm applied to low-dimension temporal features projected from flow; a multimodal length extrapolation training paradigm that utilizes low-dimension temporal features to extend the training context window size; and a bootstrapping framework that bridges our model with pluggable MLLMs without requiring annotation. We validate TGB across seven video benchmarks and demonstrate substantial performance improvements compared with prior MLLMs. Notably, our model, initially trained on sequences of four frames, effectively handles sequences up to 16 longer without sacrificing performance, highlighting its scalability and effectiveness in real-world applications. Our code publicly available.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

UNICORN: A Unified Causal Video-Oriented Language-Modeling Framework for Temporal Video-Language Tasks

Yuankao Xiong, Yixin Nie, Haotian Liu, Boxin Wang, Jun Chen, Rong Jin, Cho-Jui Hsieh, Lorenzo Torresani, Jie Lei

The great success of large language models has encouraged the development of large multimodal models, with a focus on image-language interaction. Despite promising results in various image-language downstream tasks, it is still challenging and unclear how to extend the capabilities of these models to the more complex video domain, especially when dealing with explicit temporal signals. To address the problem in existing large multimodal models, in this paper we adopt visual instruction tuning to build a unified causal video-oriented language modeling framework, named UNICORN. Specifically, we collect a comprehensive dataset under the instruction-following format, and instruction-tune

the model accordingly. Experimental results demonstrate that without customized training objectives and intensive pre-training, UNICORN can achieve comparable or better performance on established temporal video-language tasks including moment retrieval, video paragraph captioning and dense video captioning. Moreover, the instruction-tuned model can be used to automatically annotate internet videos with temporally-aligned captions. Compared to commonly used ASR captions, we show that training on our generated captions improves the performance of video-language models on both zero-shot and fine-tuning settings. Source code can be found at <https://github.com/xyh97/U-NICORN>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Shaking Up VLMs: Comparing Transformers and Structured State Space Models for Vision & Language Modeling

Georgios Panagopoulos, Malvina Nikandrou, Alessandro Sugi, Oliver Lemon, Arash Eshghi

This study explores replacing Transformers in Visual Language Models (VLMs) with Mamba, a recent structured state space model (SSM) that demonstrates promising performance in sequence modeling. We test models up to 3B parameters under controlled conditions, showing that Mamba-based VLMs outperforms Transformers-based VLMs in captioning, question answering, and reading comprehension. However, we find that Transformers achieve greater performance in visual grounding and the performance gap widens with scale. We explore two hypotheses to explain this phenomenon: 1) the effect of task-agnostic visual encoding on the updates of the hidden states, and 2) the difficulty in performing visual grounding from the perspective of in-context multimodal retrieval. Our results indicate that a task-aware encoding yields minimal performance gains on grounding, however, Transformers significantly outperform Mamba in context multimodal retrieval. Overall, Mamba shows promising performance on tasks where the correct output relies on a summary of the image but struggles when retrieval of explicit information from the context is required.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

DocEditAgent: Document Structure Editing Via Multimodal LLM Grounding

Manan Suri, Puneet Mathur, Franck Dernoncourt, Rajiv Jain, Vlad I Morariu, Ramit Sawhney, Preslav Nakov, Dinesh Manocha

Document structure editing involves manipulating localized textual, visual, and layout components in document images based on the user's requests. Past works have shown that multimodal grounding of user requests in the document image and identifying the accurate structural components and their associated attributes remain key challenges for this task. To address these, we introduce the DocEditAgent, a novel framework that performs end-to-end document editing by leveraging Large Multimodal Models (LLMs). It consists of three novel components – (1) Doc2Command to simultaneously localize edit regions of interest (RoI) and disambiguate user edit requests into edit commands. (2) LLM-based Command Reformulation prompting to tailor edit commands originally intended for specialized software into edit instructions suitable for generalist LLMs. (3) Moreover, DocEditAgent processes these outputs via Large Multimodal Models like GPT-4V and Gemini, to parse the document layout, execute edits on grounded Region of Interest (RoI), and generate the edited document image. Extensive experiments on the DocEdit dataset show that DocEditAgent significantly outperforms strong baselines on edit command generation (2-33%), RoI bounding box detection (12-31%), and overall document editing (1-12%) tasks.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Generating Demonstrations for In-Context Compositional Generalization in Grounded Language Learning

Sam Spilsbury, Pekka Marttinen, Alexander Iljin

In-Context-learning and few-shot prompting are viable methods compositional output generation. However, these methods can be very sensitive to the choice of support examples used. Retrieving good supports from the training data for a given test query is already a difficult problem, but in some cases solving this may not even be enough. We consider the setting of grounded language learning problems where finding relevant supports in the same or similar states as the query may be difficult. We design an agent which instead generates possible supports inputs and targets current state of the world, then uses them in-context-learning to solve the test query. We show substantially improved performance on a previously unsolved compositional generalization test without a loss of performance in other areas. The approach is general and can even scale to instructions expressed in natural language.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

The Instinctive Bias: Spurious Images lead to Hallucination in MLLMs

Tianyang Han, Qing LIAN, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, Tong Zhang

Large language models (LLMs) have recently experienced remarkable progress, where the advent of multi-modal large language models (MLLMs) has endowed LLMs with visual capabilities, leading to impressive performances in various multi-modal tasks. However, those powerful MLLMs such as GPT-4V still fail spectacularly when presented with certain image and text inputs. In this paper, we identify a typical class of inputs that baffle MLLMs, which consist of images that are highly relevant but inconsistent with answers, causing MLLMs to suffer from visual illusion. To quantify the effect, we propose CorrelationQA, the first benchmark that assesses the visual illusion level given spurious images. This benchmark contains 7,308 text-image pairs across 13 categories. Based on the proposed CorrelationQA, we conduct a thorough analysis on 9 mainstream MLLMs, illustrating that they universally suffer from this instinctive bias to varying degrees. We hope that our curated benchmark and evaluation results aid in better assessments of the MLLMs robustness in the presence of misleading images. The code and datasets are available at <https://github.com/MasaiahHan/CorrelationQA>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, David Acuna

Despite recent advances demonstrating vision-language models (VLMs) abilities to describe complex relationships among objects in images using natural language, their capability to quantitatively reason about object sizes and distances remains underexplored. In this work, we introduce a manually annotated benchmark of 241 questions across five categories specifically designed for quantitative spatial reasoning, and systematically investigate the performance of SoTA VLMs on this task. Our analysis reveals that questions involving reasoning about distances between objects are particularly challenging for SoTA VLMs; however, some VLMs perform significantly better at this task than others, with an almost 40 points gap between the two best performing models. We also make the surprising observation that the success rate of the top-performing VLM increases by 19 points when a reasoning path using a reference object emerges naturally in the response. Inspired by this observation, we develop a zero-shot prompting technique, SpatialPrompt, that encourages VLMs to answer quantitative spatial questions using references objects as visual cues. Specifically, we demonstrate that instructing VLMs to use reference objects in their reasoning paths significantly improves their quantitative spatial reasoning performance, bypassing the need for external data, architectural modifications, or fine-tuning. Remarkably, by solely using SpatialPrompt, Gemini 1.5 Pro, GPT-4V, and GPT-4o improve by 56.2, 28.5, and 6.7 points on average in Q-Spatial Bench without the need for more data, model architectural modifications, or fine-tuning.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

TV-TREES: Multimodal Entailment Trees for Neuro-Symbolic Video Reasoning

Kate Sanders, Nathaniel Weir, Benjamin Van Durme

It is challenging for models to understand complex, multimodal content such as television clips, and this is in part because video-language models often rely on single-modality reasoning and lack interpretability. To combat these issues we propose TV-TREES, the first multimodal

entailment tree generator. TV-TREES serves as an approach to video understanding that promotes interpretable joint-modality reasoning by searching for trees of entailment relationships between simple text-video evidence and higher-level conclusions that prove question-answer pairs. We also introduce the task of multimodal entailment tree generation to evaluate reasoning quality. Our method's performance on the challenging TVQA benchmark demonstrates interpretable, state-of-the-art zero-shot performance on full clips, illustrating that multimodal entailment tree generation can be a best-of-both-worlds alternative to black-box systems.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture

Wenyuan Li, Cristina Zhang, Jiahang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, Desmond Elliott

Food is a rich and varied dimension of cultural heritage, crucial to both individuals and social groups. To bridge the gap in the literature on the often-overlooked regional diversity in this domain, we introduce FoodieQA, a manually curated, fine-grained image-text dataset capturing the intricate features of food cultures across various regions in China. We evaluate vision-language Models (VLMs) and large language models (LLMs) on newly collected, unseen food images and corresponding questions. FoodieQA comprises three multiple-choice question-answering tasks where models need to answer questions based on multiple images, a single image, and text-only descriptions, respectively. While LLMs excel at text-based question answering, surpassing human accuracy, the open-sourced VLMs still fall short by 41% on multi-image and 21% on single-image VQA tasks, although closed-weight models perform closer to human levels (within 10%). Our findings highlight that understanding food and its cultural implications remains a challenging and under-explored direction.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Advancing Social Intelligence in AI Agents: Technical Challenges and Open Question

Leena Mathur, Paul Pu Liang, Louis-Philippe Morency

Building socially-intelligent AI agents (Social-AI) is a multidisciplinary, multimodal research goal that involves creating agents that can sense, perceive, reason about, learn from, and respond to affect, behavior, and cognition of other agents (human or artificial). Progress towards Social-AI has accelerated in the past decade across several computing communities, including natural language processing, machine learning, robotics, human-machine interaction, computer vision, and speech. Natural language processing, in particular, has been prominent in Social-AI research, as language plays a key role in constructing the social world. In this position paper, we identify a set of underlying technical challenges and open questions for researchers across computing communities to advance Social-AI. We anchor our discussion in the context of social intelligence concepts and prior progress in Social-AI research.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Mitigating Open-Vocabulary Caption Hallucinations

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, Hadar Averbuch-Elor

While recent years have seen rapid progress in image-conditioned text generation, image captioning still suffers from the fundamental issue of hallucinations, namely, the generation of spurious details that cannot be inferred from the given image. Existing methods largely use closed-vocabulary object lists to mitigate or evaluate hallucinations in image captioning, ignoring the long-tailed nature of hallucinations that occur in practice. To this end, we propose a framework for addressing hallucinations in image captioning in the open-vocabulary setting. Our framework includes a new benchmark, OpenCHAIR, that leverages generative foundation models to evaluate open-vocabulary object hallucinations for image captioning, surpassing the popular and similarly-sized CHAIR benchmark in both diversity and accuracy. Furthermore, to mitigate open-vocabulary hallucinations without using a closed object list, we propose MOCHA, an approach harnessing advancements in reinforcement learning. Our multi-objective reward function explicitly targets the trade-off between fidelity and adequacy in generations without requiring any strong supervision. MOCHA improves a large variety of image captioning models, as captured by our OpenCHAIR benchmark and other existing metrics. We will release our code and models.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

VideoINSTA: Zero-shot Long-Form Video Understanding via Informative Spatial-Temporal Reasoning

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, Volker Tresp

In the video-language domain, recent works in leveraging zero-shot Large Language Model-based reasoning for video understanding have become competitive challengers to previous end-to-end models. However, long video understanding presents unique challenges due to the complexity of reasoning over extended timespans, even for zero-shot LLM-based approaches. The challenge of information redundancy in long videos prompts the question of what specific information is essential for large language models (LLMs) and how to leverage them for complex spatial-temporal reasoning in long-form video analysis. We propose a framework VideoINSTA, i.e. INformative Spatial-Temporal Reasoning for zero-shot long-form video understanding. VideoINSTA contributes (1) a zero-shot framework for long video understanding using LLMs; (2) an event-based temporal reasoning and content-based spatial reasoning approach for LLMs to reason over spatial-temporal information in videos; (3) a self-reflective information reasoning scheme based on information sufficiency and prediction confidence while balancing temporal factors. Our model significantly improves the state-of-the-art on three long video question-answering benchmarks: EgoSchema, NextQQA, and IntentQQA, and the open question answering dataset ActivityNetQA. Code is released: <https://github.com/mayhugotong/VideoINSTA>.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Hope 'The Paragraph Guy' explains the rest : Introducing MeSum, the Meme Summarizer

Anas Anwarul haq Khan, Tanik Saikh, Arpan Phukan, Asif Ekbal

Over the years, memes have evolved into multifaceted narratives on platforms like Instagram, TikTok, and Reddit, blending text, images and audio to amplify humor and engagement. The objective of the task described in this article is to bridge the gap for individuals who may struggle to understand memes due to cultural, geographical, ancillary insights, or relevant exposure constraints, aiming to enhance meme comprehension across diverse audiences. The lack of large datasets for supervised learning and alternatives to resource-intensive vision language models have historically hindered the development of such technology. In this work, we have made strides to overcome these challenges. We introduce "MMD" a Multimodal Meme Dataset comprising 13,494 instances, including 3,134 with audio, rendering it the largest of its kind, with 2.1 times as many samples and 9.5 times as many words in the human annotated meme summary compared to the largest available meme captioning dataset, MemeCap. Our framework, MeSum (**Meme Summariser**), employs a fusion of Vision Transformer and Large Language Model technologies, providing an efficient alternative to resource-intensive Vision Language Models pioneering the integration of all three modalities, we attain a ROUGE-L score of 0.439, outperforming existing approaches such as zero-shot Gemini, GPT4 Vision, LLaVA and QwenVL which yield scores of 0.259, 0.213, 0.177 and 0.198. We have made our codes and datasets publicly available.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

AUTOHALLUSION: Automatic Generation of Hallucination Benchmarks for Vision-Language Models

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, Dinesh Manocha

Large vision-language models (LVLMs) are prone to hallucinations, where certain contextual cues in an image can trigger the language module to produce overconfident and incorrect reasoning about abnormal or hypothetical objects. While some benchmarks have been de-

veloped to investigate LVLM hallucinations, they often rely on hand-crafted corner cases whose failure patterns may not generalize well. Additionally, fine-tuning on these examples could undermine their validity. To address this, we aim to scale up the number of cases through an automated approach, reducing human bias in crafting such corner cases. This motivates the development of AutoHallusion, the first automated benchmark generation approach that employs several key strategies to create a diverse range of hallucination examples. Our generated visual-question pairs pose significant challenges to LVLMs, requiring them to overcome contextual biases and distractions to arrive at correct answers. AutoHallusion enables us to create new benchmarks at the minimum cost and thus overcomes the fragility of hand-crafted benchmarks. It also reveals common failure patterns and reasons, providing key insights to detect, avoid, or control hallucinations. Comprehensive evaluations of top-tier LVLMs, e.g., GPT-4(isian), Gemini Pro Vision, Claude 3, and LLaVA-1.5, show a 97.7% and 98.7% success rate of hallucination induction on synthetic and real-world datasets of AutoHallusion, paving the way for a long battle against hallucinations. The codebase and data can be accessed at <https://github.com/wuxiyang1996/AutoHallusion>

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

VDebugger: Harnessing Execution Feedback for Debugging Visual Programs

Xueqing Wu, Zongyu Lin, Songyan Zhao, Te-Lin Wu, Pan Lu, Nanyun Peng, Kai-Wei Chang

Visual programs are executable code generated by large language models to address visual reasoning problems. They decompose complex questions into multiple reasoning steps and invoke specialized models for each step to solve the problems. However, these programs are prone to logic errors, with our preliminary evaluation showing that 58% of the total errors are caused by program logic errors. Debugging complex visual programs remains a major bottleneck for visual reasoning. To address this, we introduce **VDebugger***, a novel critic-refiner framework trained to localize and debug visual programs by tracking execution step by step. VDebugger identifies and corrects program errors leveraging detailed execution feedback, improving interpretability and accuracy. The training data is generated through an automated pipeline that injects errors into correct visual programs using a novel mask-best decoding technique. Evaluations on six datasets demonstrate VDebugger's effectiveness, showing performance improvements of up to 3.2% in downstream task accuracy. Further studies show VDebugger's ability to generalize to unseen tasks, bringing a notable improvement of 2.3% on the unseen COVR task.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Introducing Spatial Information and a Novel Evaluation Scheme for Open-Domain Live Commentary Generation

Erica Kido Shimomoto, Edison Marrese-Taylor, Ichiro Kobayashi, Hiroya Takamura, Yusuke Miyao

This paper focuses on the task of open-domain live commentary generation. Compared to domain-specific work in this task, this setting proved particularly challenging due to the absence of domain-specific features. Aiming to bridge this gap, we integrate spatial information by proposing an utterance generation model with a novel spatial graph that is flexible to deal with the open-domain characteristics of the commentaries and significantly improves performance. Furthermore, we propose a novel evaluation scheme, more suitable for live commentary generation, that uses LLMs to automatically check whether generated utterances address essential aspects of the video via the answerability of questions extracted directly from the videos using LVLMs. Our results suggest that using a combination of our answerability score and a standard machine translation metric is likely a more reliable way to evaluate the performance in this task.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya Kaushik Surikuchi, Raquel Fernández, Sandro Pezzelle

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA obtains the best result, but only slightly so compared to TAPM, a 50-times smaller visual storytelling model. Upgrading the visual and language components of TAPM results in a model that yields competitive performance with a relatively low number of parameters. Finally, we carry out a human evaluation study, whose results suggest that a 'good' story may require more than a human-like level of visual grounding, coherence, and repetition.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Large Language Models Are Challenged by Habitat-Centered Reasoning

Sadaf Ghaffari, Nikhil Krishnaswamy

In this paper we perform a novel in-depth evaluation of text-only and multimodal LLMs' abilities to reason about object *habitats* or conditions on how objects are situated in their environments that affect the types of behaviors (or *affordances*) that can be enacted upon them. We present a novel curated multimodal dataset of questions about object habitats and affordances, which are formally grounded in the underlying lexical semantics literature, with multiple images from various sources that depict the scenario described in the question. We evaluate 16 text-only and multimodal LLMs on this challenging data. Our findings indicate that while certain LLMs can perform reasonably well on reasoning about affordances, there appears to be a consistent low upper bound on habitat-centered reasoning performance. We discuss how the formal semantics of habitats in fact predicts this behavior and propose this as a challenge to the community.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

MiRAGENews: Multimodal Realistic AI-Generated News Detection

Runsheng Huang, Liam Dugan, Chris Callison-Burch

The proliferation of inflammatory or misleading fake news content has become increasingly common in recent years. Simultaneously, it has become easier than ever to use AI tools to generate photorealistic images depicting any scene imaginable. Combining these two AI-generated fake news contents particularly potent and dangerous. To combat the spread of AI-generated fake news, we propose the MiRAGENews Dataset, a dataset of 12,500 high-quality real and AI-generated image-caption pairs from state-of-the-art generators. We find that our dataset poses a significant challenge to humans (60% F-1) and state-of-the-art multi-modal LLMs (< 24% F-1). Using our dataset we train a multimodal defector (MiRAGE) that improves by +5.1% F-1 over state-of-the-art baselines on image-caption pairs from out-of-domain image generators and news publishers. We release our code and data to aid future work on detecting AI-generated content.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

LLM-Based Offline Learning for Embodied Agents via Consistency-Guided Reward Ensemble

Yujeong Lee, Sangwoo Shin, Wei-Jin Park, Honguk Woo

Employing large language models (LLMs) to enable embodied agents has become popular, yet it presents several limitations in practice. In this work, rather than using LLMs directly as agents, we explore their use as tools for embodied agent learning. Specifically, to train separate agents via offline reinforcement learning (RL), an LLM is used to provide dense reward feedback on individual actions in training datasets. In doing so, we present a consistency-guided reward ensemble framework (CoREN), designed for tackling difficulties in grounding LLM-generated estimates to the target environment domain. The framework employs an adaptive ensemble of spatio-temporally consistent rewards to derive domain-grounded rewards in the training datasets, thus enabling effective offline learning of embodied agents in different environment domains. Experiments with the VirtualHome benchmark demonstrate that CoREN significantly outperforms other offline RL

agents, and it also achieves comparable performance to state-of-the-art LLM-based agents with 8B parameters, despite CoREN having only 117M parameters for the agent policy network and using LLMs only for training.

Nov 13 (Wed) 16:00-17:30 - *Riverfront Hall*

FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models

Liqiang Jing, Ruosen Li, Yummo Chen, Xinya Du

We introduce FaithScore (Faithfulness to Atomic Image Facts Score), a reference-free and fine-grained evaluation metric that measures the faithfulness of the generated free-form answers from large vision-language models (LVLMs). The FaithScore evaluation first identifies sub-sentences containing descriptive statements that need to be verified, then extracts a comprehensive list of atomic facts from these sub-sentences, and finally conducts consistency verification between fine-grained atomic facts and the input image. Meta-evaluation demonstrates that our metric highly correlates with human judgments of faithfulness. We collect two benchmark datasets (i.e. LLaVA-1k and MSCOCO-Cap) for evaluating LVLMs instruction-following hallucinations. We measure hallucinations in state-of-the-art LVLMs with FaithScore on the datasets. Results reveal that current systems are prone to generate hallucinated content unfaithful to the image, which leaves room for future improvements. We hope our metric FaithScore can help evaluate future LVLMs in terms of faithfulness and provide insightful advice for enhancing LVLMs faithfulness.

Question Answering 3

Nov 13 (Wed) 16:00-17:30 - Room: *Jasmine*

Nov 13 (Wed) 16:00-17:30 - *Jasmine*

Pre-training Cross-lingual Open Domain Question Answering with Large-scale Synthetic Supervision

Fan Jiang, Tom Drummond, Trevor Cohn

Cross-lingual open domain question answering (CLQA) is a complex problem, comprising cross-lingual retrieval from a multilingual knowledge base, followed by answer generation in the query language. Both steps are usually tackled by separate models, requiring substantial annotated datasets, and typically auxiliary resources, like machine translation systems to bridge between languages. In this paper, we show that CLQA can be addressed using a single encoder-decoder model. To effectively train this model, we propose a self-supervised method based on exploiting the cross-lingual link structure within Wikipedia. We demonstrate how linked Wikipedia pages can be used to synthesise supervisory signals for cross-lingual retrieval, through a form of cloze query, and generate more natural questions to supervise answer generation. Together, we show our approach, CLASS, outperforms comparable methods on both supervised and zero-shot language adaptation settings, including those using machine translation.

Nov 13 (Wed) 16:00-17:30 - *Jasmine*

Seemingly Plausible Distractors in Multi-Hop Reasoning: Are Large Language Models Attentive Readers?

Neeladri Bhuiya, Viktor Schlegel, Stefan Winkler

State-of-the-art Large Language Models (LLMs) are accredited with an increasing number of different capabilities, ranging from reading comprehension over advanced mathematical and reasoning skills to possessing scientific knowledge. In this paper we focus on multi-hop reasoning—the ability to identify and integrate information from multiple textual sources. Given the concerns with the presence of simplifying cues in existing multi-hop reasoning benchmarks, which allow models to circumvent the reasoning requirement, we set out to investigate whether LLMs are prone to exploiting such simplifying cues. We find evidence that they indeed circumvent the requirement to perform multi-hop reasoning, but they do so in more subtle ways than what was reported about their fine-tuned pre-trained language model (PLM) predecessors. We propose a challenging multi-hop reasoning benchmark by generating seemingly plausible multi-hop reasoning chains that ultimately lead to incorrect answers. We evaluate multiple open and proprietary state-of-the-art LLMs and show that their multi-hop reasoning performance is affected, as indicated by up to 45% relative decrease in F1 score when presented with such seemingly plausible alternatives. We also find that while LLMs tend to ignore misleading lexical cues misleading reasoning paths indeed present a significant challenge. The code and data are made available at <https://github.com/zawedcvg/Are-Large-Language-Models-Attentive-Readers>.

Nov 13 (Wed) 16:00-17:30 - *Jasmine*

Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation

Jirui Qi, Gabriele Sarti, Raquel Fernández, Arianna Bisazza

Ensuring the verifiability of model answers is a fundamental challenge for retrieval-augmented generation (RAG) in the question answering (QA) domain. Recently, self-citation prompting was proposed to make large language models (LLMs) generate citations to supporting documents along with their answers. However, self-citing LLMs often struggle to match the required format, refer to non-existent sources, and fail to faithfully reflect LLMs' context usage throughout the generation. In this work, we present MIRAGE – Model Internals-based RAG Explanations – a plug-and-play approach using model internals for faithful answer attribution in RAG applications. MIRAGE detects context-sensitive answer tokens and pairs them with retrieved documents contributing to their prediction via saliency methods. We evaluate our proposed approach on a multilingual extractive QA dataset, finding high agreement with human answer attribution. On open-ended QA, MIRAGE achieves citation quality and efficiency comparable to self-citation while also allowing for a finer-grained control of attribution parameters. Our qualitative evaluation highlights the faithfulness of MIRAGE's attributions and underscores the promising application of model internals for RAG answer attribution. Code and data released at <https://github.com/Betswish/MIRAGE>.

Nov 13 (Wed) 16:00-17:30 - *Jasmine*

Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs

Haritz Puerto, Martin Tutek, Somak Aditya, Xiaodan Zhu, Iryna Gurevych

Reasoning is a fundamental component of language understanding. Recent prompting techniques, such as chain of thought, have consistently improved LLMs' performance on various reasoning tasks. Nevertheless, there is still little understanding of what triggers reasoning abilities in LLMs in the inference stage. In this paper, we investigate the effect of the input representation on the reasoning abilities of LLMs. We hypothesize that representing natural language tasks as code can enhance specific reasoning abilities such as entity tracking or logical reasoning. To study this, we propose code prompting, a methodology we operationalize as a chain of prompts that transforms a natural language problem into code and directly prompts the LLM using the generated code without resorting to external code execution. We find that code prompting exhibits a high-performance boost for multiple LLMs (up to 22.52 percentage points GPT 3.5, 7.75 on Mixtral, and 16.78 on Mistral) across multiple conditional reasoning datasets. We then conduct comprehensive experiments to understand how the code representation triggers reasoning abilities and which capabilities are elicited in the underlying models. Our analysis on GPT 3.5 reveals that the code formatting of the input problem is essential for performance improvement. Furthermore, the code representation improves sample efficiency of in-context learning and facilitates state tracking of entities.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Subjective Topic meets LLMs: Unleashing Comprehensive, Reflective and Creative Thinking through the Negation of Negation

Fangrui Lv, Kaixiang Gong, Jian Liang, Xinyu Pang, Changshui Zhang

Large language models (LLMs) exhibit powerful reasoning capacity, as evidenced by prior studies focusing on objective topics that with unique standard answers such as arithmetic and commonsense reasoning. However, the reasoning to definite answers emphasizes more on logical thinking, and falls short in effectively reflecting the comprehensive, reflective, and creative thinking that is also critical for the overall reasoning prowess of LLMs. In light of this, we build a dataset SJTP comprising diverse Subjective Topics with free responses, as well as three evaluation indicators to fully explore LLMs reasoning ability. We observe that a sole emphasis on logical thinking falls short in effectively tackling subjective challenges. Therefore, we introduce a framework grounded in the principle of the Negation of Negation (NeoN) to unleash the potential comprehensive, reflective, and creative thinking abilities of LLMs. Comprehensive experiments on SJTP demonstrate the efficacy of NeoN, and the enhanced performance on various objective reasoning tasks unequivocally underscores the benefits of stimulating LLMs subjective thinking in augmenting overall reasoning capabilities.

Nov 13 (Wed) 16:00-17:30 - Jasmine

KARL: Knowledge-Aware Retrieval and Representations aid Retention and Learning in Students

Matthew Shu, Nishant Balepur, Shi Feng, Jordan Lee Boyd-Graber

Flashcard schedulers rely on 1) *student models* to predict the flashcards a student knows; and 2) *teaching policies* to pick which cards to show next via these predictions. Prior student models, however, just use study data like the student's past responses, ignoring the text on cards. We propose **content-aware scheduling**, the first schedulers exploiting flashcard content. To give the first evidence that such schedulers enhance student learning, we build KARL, a simple but effective content-aware student model employing deep knowledge tracing (DKT), retrieval, and BERT to predict student recall. We train KARL by collecting a new dataset of 123,143 study logs on diverse trivia questions. KARL beats existing student models in AUC and calibration error. To ensure our improved predictions lead to better student learning, we create a novel delta-based teaching policy to deploy KARL online. Based on 32 study paths from 27 users, KARL improves learning efficiency over SOTA, showing KARL's strength and encouraging researchers to look beyond historical study data to fully capture student abilities.

Nov 13 (Wed) 16:00-17:30 - Jasmine

LONGAGENT: Achieving Question Answering for 128k-Token-Long Documents through Multi-Agent Collaboration

Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Ywen Ding, Tao Gui, Qi Zhang, Xuanjing Huang

Large language models (LLMs) have achieved tremendous success in understanding language and processing text. However, question-answering (QA) on lengthy documents faces challenges of resource constraints and a high propensity for errors, even for the most advanced models such as GPT-4 and Claude2. In this paper, we introduce `_LongAgent_`, a multi-agent collaboration method that enables efficient and effective QA over 128k-token-long documents. `_LongAgent_` adopts a `_divide-and-conquer_` strategy, breaking down lengthy documents into shorter, more manageable text chunks. A leader agent comprehends the user's query and organizes the member agents to read their assigned chunks, reasoning a final answer through multiple rounds of discussion. Due to members' hallucinations, it's difficult to guarantee that every response provided by each member is accurate. To address this, we develop an `_inter-member communication_` mechanism that facilitates information sharing, allowing for the detection and mitigation of hallucinatory responses. Experimental results show that a LLaMA-2 7B driven by `_LongAgent_` can effectively support QA over 128k-token documents, achieving 16.42% and 1.63% accuracy gains over GPT-4 on single-hop and multi-hop QA settings, respectively.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Adaptive Question Answering: Enhancing Language Model Proficiency for Addressing Knowledge Conflicts with Source Citations

Sagi Shaier, Ari Kobren, Philip V. Ogren

Resolving knowledge conflicts is a crucial challenge in Question Answering (QA) tasks, as the internet contains numerous conflicting facts and opinions. While some research has made progress in tackling ambiguous settings where multiple valid answers exist, these approaches often neglect to provide source citations, leaving users to evaluate the factuality of each answer. On the other hand, existing work on citation generation has focused on unambiguous settings with single answers, failing to address the complexity of real-world scenarios. Despite the importance of both aspects, no prior research has combined them, leaving a significant gap in the development of QA systems. In this work, we bridge this gap by proposing the novel task of QA with source citation in ambiguous settings, where multiple valid answers exist. To facilitate research in this area, we create a comprehensive framework consisting of: (1) five novel datasets, obtained by augmenting three existing reading comprehension datasets with citation meta-data across various ambiguous settings, such as distractors and paraphrasing; (2) the first ambiguous multi-hop QA dataset featuring real-world, naturally occurring contexts; (3) two new metrics to evaluate models performances; and (4) several strong baselines using rule-based, prompting, and finetuning approaches over five large language models. We hope that this new task, datasets, metrics, and baselines will inspire the community to push the boundaries of QA research and develop more trustworthy and interpretable systems.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering

Yao Xu, Shizhu He, Jiapei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, Kang Liu

To address the issues of insufficient knowledge and hallucination in Large Language Models (LLMs), numerous studies have explored integrating LLMs with Knowledge Graphs (KGs). However, these methods are typically evaluated on conventional Knowledge Graph Question Answering (KGQA) with complete KGs, where all factual triples required for each question are entirely covered by the given KG. In such cases, LLMs primarily act as an agent to find answer entities within the KG, rather than effectively integrating the internal knowledge of LLMs and external knowledge sources such as KGs. In fact, KGs are often incomplete to cover all the knowledge required to answer questions. To simulate these real-world scenarios and evaluate the ability of LLMs to integrate internal and external knowledge, we propose leveraging LLMs for QA under Incomplete Knowledge Graph (IKGQA), where the provided KG lacks some of the factual triples for each question, and construct corresponding datasets. To handle IKGQA, we propose a training-free method called Generate-on-Graph (GoG), which can generate new factual triples while exploring KGs. Specifically, GoG performs reasoning through a Thinking-Searching-Generating framework, which treats LLM as both Agent and KG in IKGQA. Experimental results on two datasets demonstrate that our GoG outperforms all previous methods.

Nov 13 (Wed) 16:00-17:30 - Jasmine

You Make me Feel like a Natural Question: Training QA Systems on Transformed Trivia Questions

Tasnim Kabir, Yoo Yeon Sung, Saptarashmi Bandopadhyay, Hao Zou, Abhranil Chandra, Jordan Lee Boyd-Graber

Training question-answering QA and information retrieval systems for web queries require large, expensive datasets that are difficult to annotate and time-consuming to gather. Moreover, while natural datasets of information-seeking questions are often prone to ambiguity or ill-formed, there are troves of freely available, carefully crafted question datasets for many languages. Thus, we automatically generate shorter, information-seeking questions, resembling web queries in the style of the Natural Questions (NQ) dataset from longer trivia data. Training a QA system on these transformed questions is a viable strategy for alternating to more expensive training setups showing the F1 score difference of less than six points and contrasting the final systems.

Nov 13 (Wed) 16:00-17:30 - Jasmine

COMPACT: Compressing Retrieved Documents Actively for Question Answering

Chanwoong Yoon, Taewho Lee, Hyeon Hwang, Minbyul Jeong, Jaewoo Kang

Retrieval-augmented generation supports language models to strengthen their factual groundings by providing external contexts. However, language models often face challenges when given extensive information, diminishing their effectiveness in solving questions. Context compression tackles this issue by filtering out irrelevant information, but current methods still struggle in realistic scenarios where crucial information cannot be captured with a single-step approach. To overcome this limitation, we introduce CompAct, a novel framework that employs an active strategy to condense extensive documents without losing key information. Our experiments demonstrate that CompAct brings significant improvements in both performance and compression rate on multi-hop question-answering benchmarks. CompAct flexibly operates as a cost-efficient plug-in module with various off-the-shelf retrievers or readers, achieving exceptionally high compression rates (47x).

Nov 13 (Wed) 16:00-17:30 - Jasmine

RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation

Kiseung Kim, Jay-Yoon Lee

The Retrieval-Augmented Generation (RAG) framework utilizes a combination of parametric knowledge and external knowledge to demonstrate state-of-the-art performance on open-domain question answering tasks. However, the RAG framework suffers from performance degradation when the query is accompanied by irrelevant contexts. In this work, we propose the RE-RAG framework, which introduces a relevance estimator (RE) that not only provides relative relevance between contexts as previous rerankers did, but also provide confidence, which can be used to classify whether given context is useful for answering the given question. We propose a weakly supervised method for training the RE simply utilizing question-answer data without any labels for correct contexts. We show that RE trained with a small generator (sLM) can not only improve the sLM fine-tuned together with RE but also improve previously unrefined large language models (LLMs). Furthermore, we investigate new decoding strategies that utilize the proposed confidence measured by RE such as choosing to let the user know that it is unanswerable to answer the question given the retrieved contexts or choosing to rely on LLMs parametric knowledge rather than unrelated contexts.

Nov 13 (Wed) 16:00-17:30 - Jasmine

ZEBRA: Zero-Shot Example-Based Retrieval Augmentation for Commonsense Question Answering

Francesco Maria Moifese, Simone Conia, Riccardo Orlando, Roberto Navigli

Current Large Language Models (LLMs) have shown strong reasoning capabilities in commonsense question answering benchmarks, but the process underlying their success remains largely opaque. As a consequence, recent approaches have equipped LLMs with mechanisms for knowledge retrieval, reasoning and introspection, not only to improve their capabilities but also to enhance the interpretability of their outputs. However, these methods require additional training, hand-crafted templates or human-written explanations. To address these issues, we introduce ZEBRA, a zero-shot question answering framework that combines retrieval, case-based reasoning and introspection and dispenses with the need for additional training of the LLM. Given an input question, ZEBRA retrieves relevant question-knowledge pairs from a knowledge base and generates new knowledge by reasoning over the relationships in these pairs. This generated knowledge is then used to answer the input question, improving the model's performance and interpretability. We evaluate our approach across 8 well-established commonsense reasoning benchmarks, demonstrating that ZEBRA consistently outperforms strong LLMs and previous knowledge integration approaches, achieving an average accuracy improvement of up to 4.5 points.

Nov 13 (Wed) 16:00-17:30 - Jasmine

BERGEN: A Benchmarking Library for Retrieval-Augmented Generation

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, shuai wang, Stéphane CLINCHANT, Vassilina Nikoulina

Retrieval-Augmented Generation allows to enhance Large Language Models with external knowledge. In response to the recent popularity of generative LLMs, many RAG approaches have been proposed, which involve an intricate number of different configurations such as evaluation datasets, collections, metrics, retrievers, and LLMs. Inconsistent benchmarking poses a major challenge in comparing approaches and understanding the impact of each component in the pipeline. In this work, we study best practices that lay the groundwork for a systematic evaluation of RAG and present BERGEN, an end-to-end library for reproducible research standardizing RAG experiments. In an extensive study focusing on QA, we benchmark different state-of-the-art retrievers, rerankers, and LLMs. Additionally, we analyze existing RAG metrics and datasets.

Nov 13 (Wed) 16:00-17:30 - Jasmine

ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, Yiyu Luo

Chart question answering (ChartQA) tasks play a critical role in interpreting and extracting insights from visualization charts. While recent advancements in multimodal large language models (MLLMs) like GPT-4o have shown promise in high-level ChartQA tasks, such as chart captioning, their effectiveness in low-level ChartQA tasks (*e.g.* identifying correlations) remains underexplored. In this paper, we address this gap by evaluating MLLMs on low-level ChartQA using a newly curated dataset, *ChartInsights*, which consists of 22,347 (chart, task, query, answer) 10 data analysis tasks across 7 chart types. We systematically evaluate 19 advanced MLLMs, including 12 open-source and 7 closed-source models. The average accuracy rate across these models is 39.8%, with GPT-4o achieving the highest accuracy at 69.17%. To further explore the limitations of MLLMs in low-level ChartQA, we conduct experiments that alter visual elements of charts (*e.g.* changing color schemes, adding image noise) to assess their impact on the task effectiveness. Furthermore, we propose a new textual prompt strategy, *Chain-of-Charts*, tailored for low-level ChartQA tasks, which boosts performance by 14.41%, achieving an accuracy of 83.58%. Finally, incorporating a visual prompt strategy that directs attention to relevant visual elements further improves accuracy to 84.32%.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Synthetic Multimodal Question Generation

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakzad, Tongshuang Wu, Graham Neubig

Multimodal Retrieval Augmented Generation (MMRAG) is a powerful approach to question-answering over multimodal documents. A key challenge with evaluating MMRAG is the paucity of high-quality datasets matching the question styles and modalities of interest. In light of this, we propose SMMQG, a synthetic data generation framework. SMMQG leverages interplay between a retriever, large language model (LLM) and large multimodal model (LMM) to generate question and answer pairs directly from multimodal documents, with the questions conforming to specified styles and modalities. We use SMMQG to generate an MMRAG dataset of 1024 questions over Wikipedia documents and evaluate state-of-the-art models using it, revealing insights into model performance that are attainable only through style- and modality-specific evaluation data. Next, we measure the quality of data produced by SMMQG via a human study. We find that the quality of SMMQG-generated synthetic data is on par with the quality of the crowdsourced benchmark MMQA and that downstream evaluation results using both datasets strongly concur.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Enhancing Biomedical Knowledge Retrieval-Augmented Generation with Self-Rewarding Tree Search and Proximal Policy Optimization

Minda Hu, Licheng Zong, Hongru WANG, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, Irwin King

Large Language Models (LLMs) have shown great potential in the biomedical domain with the advancement of retrieval-augmented generation (RAG). However, existing retrieval-augmented approaches face challenges in addressing diverse queries and documents, particularly for medical knowledge queries, resulting in sub-optimal performance. To address these limitations, we propose a novel plug-and-play LLM-based retrieval method called Self-Rewarding Tree Search (SeRTS) based on Monte Carlo Tree Search (MCTS) and a self-rewarding paradigm. By combining the reasoning capabilities of LLMs with the effectiveness of tree search, SeRTS boosts the zero-shot performance of retrieving high-quality and informative results for RAG. We further enhance retrieval performance by fine-tuning LLMs with Proximal Policy Optimization (PPO) objectives using the trajectories collected by SeRTS as feedback. Controlled experiments using the BioASQ-QA dataset with GPT-3.5-Turbo and LLaMA2-7B demonstrate that our method significantly improves the performance of the BM25 retriever and surpasses the strong baseline of self-reflection in both efficiency and scalability. Moreover, SeRTS generates higher-quality feedback for PPO training than self-reflection. Our proposed method effectively adapts LLMs to document retrieval tasks, enhancing their ability to retrieve highly relevant documents for RAG in the context of medical knowledge queries. This work presents a significant step forward in leveraging LLMs for accurate and comprehensive biomedical question answering.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Retrieving Contextual Information for Long-Form Question Answering using Weak Supervision

Philipp Chrismann, Svitlana Vakulenko, Ionut Teodor Sorodoc, Adrià de Gispert, Bill Byrne

Long-form question answering (LFQA) aims at generating in-depth answers to end-user questions, providing relevant information beyond the direct answer. However, existing retrievers are typically optimized towards information that directly targets the question, missing out on such contextual information. Furthermore, there is a lack of training data for relevant context. To this end, we propose and compare different weak supervision techniques to optimize retrieval for contextual information. Experiments demonstrate improvements on the end-to-end QA performance on ASQA, a dataset for long-form question answering. Importantly, as more contextual information is retrieved, we improve the relevant page recall for LFQA by 14.7% and the groundness of generated long-form answers by 12.5%. Finally, we show that long-form answers often anticipate likely follow-up questions, via experiments on a conversational QA dataset.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Less is More: Making Smaller Language Models Competent Subgraph Retrievers for Multi-hop KGQA

Wenyu Huang, Guancheng Zhou, Hongru WANG, Pavlos Vogioukakis, Mirella Lapata, Jeff Z. Pan

Retrieval-Augmented Generation (RAG) is widely used to inject external non-parametric knowledge into large language models (LLMs). Recent works suggest that Knowledge Graphs (KGs) contain valuable external knowledge for LLMs. Retrieving information from KGs differs from extracting it from document sets. Most existing approaches seek to directly retrieve relevant subgraphs, thereby eliminating the need for extensive SPARQL annotations, traditionally required by semantic parsing methods. In this paper, we model the subgraph retrieval task as a conditional generation task handled by small language models. Specifically, we define a subgraph identifier as a sequence of relations, each represented as a special token stored in the language models. Our base generative subgraph retrieval model, consisting of only 220M parameters, achieves competitive retrieval performance compared to state-of-the-art models relying on 7B parameters, demonstrating that small language models are capable of performing the subgraph retrieval task. Furthermore, our largest 3B model, when plugged with an LLM reader, sets new SOTA end-to-end performance on both the WebQSP and CWQ benchmarks. Our model and data will be made available online: <https://github.com/hwy9855/GSR>.

Nov 13 (Wed) 16:00-17:30 - Jasmine

SaSR-Net: Source-Aware Semantic Representation Network for Enhancing Audio-Visual Question Answering

Tianyu Yang, Yiyang Nan, Lisen Dai, Zhenwen Liang, Yapeng Tian, Xiangliang Zhang

Audio-Visual Question Answering (AVQA) is a challenging task that involves answering questions based on both auditory and visual information in videos. A significant challenge is interpreting complex multi-modal scenes, which include both visual objects and sound sources, and connecting them to the given question. In this paper, we introduce the Source-aware Semantic Representation Network (SaSR-Net), a novel model designed for AVQA. SaSR-Net utilizes source-wise learnable tokens to efficiently capture and align audio-visual elements with the corresponding question. It streamlines the fusion of audio and visual information using spatial and temporal attention mechanisms to identify answers in multi-modal scenes. Extensive experiments on the Music-AVQA and AVQA-Yang datasets show that SaSR-Net outperforms state-of-the-art AVQA methods. We will release our source code and pre-trained models.

Nov 13 (Wed) 16:00-17:30 - Jasmine

SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions

Shicheng Liu, Sina Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, Monica Lam

Large Language Models (LLMs) have led to significant improvements in the Knowledge Base Question Answering (KBQA) task. However, datasets used in KBQA studies do not capture the true complexity of KBQA tasks. They either have simple questions, use synthetically generated logical forms, or are based on small knowledge base (KB) schemas. We introduce the SPINACH dataset, an expert-annotated KBQA dataset collected from discussions on Wikidata's "Request a Query" forum with 320 decontextualized question-SPARQL pairs. The complexity of these in-the-wild queries calls for a KBQA system that can dynamically explore large and often incomplete schemas and reason about them, as it is infeasible to create a comprehensive training dataset. We also introduce an in-context learning KBQA agent, also called SPINACH, that mimics how a human expert would write SPARQLs to handle challenging questions. SPINACH achieves a new state of the art on the QALD-7, QALD-9 Plus and QALD-10 datasets by 31.0%, 27.0%, and 10.0% in F_1 , respectively, and coming within 1.6% of the fine-tuned LLaMA SOTA model on WikiWebQuestions. On our new SPINACH dataset, the SPINACH agent outperforms all baselines, including the best GPT-4-based KBQA agent, by at least 38.1% in F_1 .

Nov 13 (Wed) 16:00-17:30 - Jasmine

Large Language Models are In-context Teachers for Knowledge Reasoning

Jiachen Zhao, Zonghai Yao, Zhichao Yang, hong yu

In this work, we study in-context teaching(ICI), where a teacher provides in-context example rationales to teach a student to reason over unseen cases. Human teachers are usually required to craft in-context demonstrations, which are costly and have high variance. We ask whether a large language model (LLM) can serve as a more effective in-context teacher for itself or otherLLMs, compared to humans. Inspired by the Encoding Specificity Hypothesis from human episodic memory, we hypothesize that in-context exemplars crafted by the teacher should match the training data of the student. This hypothesis motivates us to propose Self-Explain where an LLMs self-elicited explanations are used as in-context demonstrations for prompting it as they are generalized from the models training examples. Self-Explain is shown to significantly outperform using human-crafted exemplars and other baselines. Furthermore, we reveal that for ICI, rationales from different teacher LLMs or human experts that more resemble the student LLMs self-explanations are better in-context demonstrations. This supports our encoding specificity hypothesis. We then propose Teach-Back that aligns a teacher LLM with the student to enhance the ICI performance. For example,

Teach-Back enables a 7B model to teach the much larger GPT-3.5 in context, surpassing human teachers by around 5% in test accuracy on medical question answering.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Chain of Condition: Construct, Verify and Solve Conditions for Conditional Question Answering

Jiuheng Lin, Yuxuan Lai, Yansong Feng

Conditional question answering (CQA) is an important task that aims to find probable answers and identify missing conditions. Existing approaches struggle with CQA due to two challenges: (1) precisely identifying necessary conditions and the logical relationship, and (2) verifying conditions to detect any that are missing. In this paper, we propose a novel prompting approach, Chain of condition, by first identifying all conditions and constructing their logical relationships explicitly according to the document, then verifying whether these conditions are satisfied, finally solving the logical expression to indicate any missing conditions and generating the answer accordingly. Experiments on two CQA benchmark datasets show our chain of condition outperforms existing prompting baselines, establishing a new state of the art. Furthermore, with only a few examples, our method can facilitate GPT-3.5-Turbo or GPT-4 to outperform all existing supervised models.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Visual Question Decomposition on Multimodal Large Language Models

Haowei Zhang, Jianzhe Liu, Zhen Han, Shuo Chen, Bailian He, Volker Tresp, zhiquiang xu, Jindong Gu

Question decomposition has emerged as an effective strategy for prompting Large Language Models (LLMs) to answer complex questions. However, while existing methods primarily focus on unimodal language models, the question decomposition capability of Multimodal Large Language Models (MLLMs) has yet to be explored. To this end, this paper explores visual question decomposition on MLLMs. Specifically, we introduce a systematic evaluation framework including a dataset and several evaluation criteria to assess the quality of the decomposed sub-questions, revealing that existing MLLMs struggle to produce high-quality sub-questions. To address this limitation, we propose a specific finetuning dataset, DecoVQA+, for enhancing the model's question decomposition capability. Aiming at enabling models to perform appropriate selective decomposition, we propose an efficient finetuning pipeline. The finetuning pipeline consists of our proposed dataset and a training objective for selective decomposition. Finetuned MLLMs demonstrate significant improvements in the quality of sub-questions and the policy of selective question decomposition. Additionally, the models also achieve higher accuracy with selective decomposition on VQA benchmark datasets.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Towards Robust Extractive Question Answering Models: Rethinking the Training Methodology

Son Quan Tran, Matt Kretschmar

This paper proposes a novel training method to improve the robustness of Extractive Question Answering (EQA) models. Previous research has shown that existing models, when trained on EQA datasets that include unanswerable questions, demonstrate a significant lack of robustness against distribution shifts and adversarial attacks. Despite this, the inclusion of unanswerable questions in EQA training datasets is essential for ensuring real-world reliability. Our proposed training method includes a novel loss function for the EQA problem and challenges an implicit assumption present in numerous EQA datasets. Models trained with our method maintain in-domain performance while achieving a notable improvement on out-of-domain datasets. This results in an overall F1 score improvement of 5.7 across all testing sets. Furthermore, our models exhibit significantly enhanced robustness against two types of adversarial attacks, with a performance decrease of only about one-third compared to the default models.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Adaptive Contrastive Decoding in Retrieval-Augmented Generation for Handling Noisy Contexts

Young Kim, Hyulung Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, Taeuk Kim

When using large language models (LLMs) in knowledge-intensive tasks, such as open-domain question answering, external context can bridge the gap between external knowledge and the LLMs' parametric knowledge. Recent research has been developed to amplify contextual knowledge over the parametric knowledge of LLMs with contrastive decoding approaches. While these approaches could yield truthful responses when relevant context is provided, they are prone to vulnerabilities when faced with noisy contexts. We extend the scope of previous studies to encompass noisy contexts and propose adaptive contrastive decoding (ACD) to leverage contextual influence effectively. ACD demonstrates improvements in open-domain question answering tasks compared to baselines, especially in robustness by remaining undistracted by noisy contexts in retrieval-augmented generation.

Nov 13 (Wed) 16:00-17:30 - Jasmine

More Bang for your Context: Virtual Documents for Question Answering over Long Documents

Yosi Mass, Boaz Carmeli, Asaf Yehudai, Assaf Toledo, Nathaniel Mills

We deal with the problem of Question Answering (QA) over a long document, which poses a challenge for modern Large Language Models (LLMs). Although LLMs can handle increasingly longer context windows, they struggle to effectively utilize the long content. To address this issue, we introduce the concept of a virtual document (VDoc). A VDoc is created by selecting chunks from the original document that are most likely to contain the information needed to answer the user's question, while ensuring they fit within the LLMs context window. We hypothesize that providing a short and focused VDoc to the LLM is more effective than filling the entire context window with less relevant information. Our experiments confirm this hypothesis and demonstrate that using VDocs improves results on the QA task.

Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas

Nov 13 (Wed) 16:00-17:30 - Room: Riverfront Hall

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

What's Mine becomes Yours: Defining, Annotating and Detecting Context-Dependent Paraphrases in News Interview Dialogs

Anna Wegmann, Tijis A. van den Broek, Dong Nguyen

Best practices for high conflict conversations like counseling or customer support almost always include recommendations to paraphrase the previous speaker. Although paraphrase classification has received widespread attention in NLP, paraphrases are usually considered independent from context, and common models and datasets are not applicable to dialog settings. In this work, we investigate paraphrases across turns in dialog (e.g., Speaker 1: "That book is mine." becomes Speaker 2: "That book is yours."). We provide an operationalization of context-dependent paraphrases, and develop a training for crowd-workers to classify paraphrases in dialog. We introduce ContextDeP, a dataset with utterance pairs from NPR and CNN news interviews annotated for context-dependent paraphrases. To enable analyses on label variation, the dataset contains 5,581 annotations on 600 utterance pairs. We present promising results with in-context learning and with token classification models for automatic paraphrase detection in dialog.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

How Hard is this Test Set? NLI Characterization by Exploiting Training Dynamics

Adrian Cosma, Stefan Rusei, Mihai Dascalu, Cornelia Caragea

Natural Language Inference (NLI) evaluation is crucial for assessing language understanding models; however, popular datasets suffer from systematic spurious correlations that artificially inflate actual model performance. To address this, we propose a method for the automated creation of a challenging test set without relying on the manual construction of artificial and unrealistic examples. We categorize the test set of popular NLI datasets into three difficulty levels by leveraging methods that exploit training dynamics. This categorization significantly reduces spurious correlation measures, with examples labeled as having the highest difficulty showing markedly decreased performance and encompassing more realistic and diverse linguistic phenomena. When our characterization method is applied to the training set, models trained with only a fraction of the data achieve comparable performance to those trained on the full dataset, surpassing other dataset characterization techniques. Our research addresses limitations in NLI dataset construction, providing a more authentic evaluation of model performance with implications for diverse NLU applications.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

FOOL ME IF YOU CAN! An Adversarial Dataset to Investigate the Robustness of LMs in Word Sense Disambiguation

Mohamad Ballout, Anne Dederle, Nahay Muhammad Abdelmoneim, Ulf Krümmack, Günther Heidemann, Kai-Uwe Kühnberger

Word sense disambiguation (WSD) is a key task in natural language processing and lexical semantics. Pre-trained language models with contextualized word embeddings have significantly improved performance in regular WSD tasks. However, these models still struggle with recognizing semantic boundaries and often misclassify homonyms in adversarial context. Therefore, we propose FOOL: FOur-fold Obscure Lexical, a new coarse-grained WSD dataset, which includes four different test sets designed to assess the robustness of language models in WSD tasks. Two sets feature typical WSD scenarios, while the other two include sentences with opposing contexts to challenge the models further. We tested two types of models on the proposed dataset: models with encoders, such as the BERT and T5 series of varying sizes by probing their embeddings, and state-of-the-art large decoder models like GPT-4o and the LLaMA3 family, using zero shot prompting. Across different state-of-the-art language models, we observed a decrease in performance in the latter two sets compared to the first two, with some models being affected more than others. We show interesting findings where small models like T5-large and BERT-large performed better than GPT-4o on Set 3 of the dataset. This indicates that, despite excelling in regular WSD tasks, these models still struggle to correctly disambiguate homonyms in artificial (Set 3) or realistic adversarial contexts (Set 4).

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Story Embeddings: Narrative-Focused Representations of Fictional Stories

Hans Ole Hatzel, Chris Biemann

We present a novel approach to modeling fictional narratives. The proposed model creates embeddings that represent a story such that similar narratives, that is, reformulations of the same story, will result in similar embeddings. We showcase the prowess of our narrative-focused embeddings on various datasets, exhibiting state-of-the-art performance on multiple retrieval tasks. The embeddings also show promising results on a narrative understanding task. Additionally, we perform an annotation-based evaluation to validate that our introduced computational notion of narrative similarity aligns with human perception. The approach can help to explore vast datasets of stories, with potential applications in recommender systems and in the computational analysis of literature.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

A Survey of AMR Applications

Shira Wein, Juri Optiz

In the ten years since the development of the Abstract Meaning Representation (AMR) formalism, substantial progress has been made on AMR-related tasks such as parsing and alignment. Still, the engineering applications of AMR are not fully understood. In this survey, we categorize and characterize more than 100 papers which use AMR for downstream tasks the first survey of this kind for AMR. Specifically, we highlight (1) the range of applications for which AMR has been harnessed, and (2) the techniques for incorporating AMR into those applications. We also detect broader AMR engineering patterns and outline areas of future work that seem ripe for AMR incorporation. We hope that this survey will be useful to those interested in using AMR and that it sparks discussion on the role of symbolic representations in the age of neural-focused NLP research.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

ECON: On the Detection and Resolution of Evidence Conflicts

Cheng Jiayang, Qianqian Zhuang, Chunkit Chan, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, Zheng Zhang

The rise of large language models (LLMs) has significantly influenced the quality of information in decision-making systems, leading to the prevalence of AI-generated content and challenges in detecting misinformation and managing conflicting information, or "inter-evidence conflicts." This study introduces a method for generating diverse, validated evidence conflicts to simulate real-world misinformation scenarios. We evaluate conflict detection methods, including Natural Language Inference (NLI) models, factual consistency (FC) models, and LLMs, on these conflicts (RQ1) and analyze LLMs' conflict resolution behaviors (RQ2). Our key findings include: (1) NLI and LLM models exhibit high precision in detecting answer conflicts, though weaker models suffer from low recall; (2) FC models struggle with lexically similar answer conflicts, while NLI and LLM models handle these better; and (3) stronger models like GPT-4 show robust performance, especially with nuanced conflicts. For conflict resolution, LLMs often favor one piece of conflicting evidence without justification and rely on internal knowledge if they have prior beliefs.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Enhancing Systematic Decompositional Natural Language Inference Using Informal Logic

Nathaniel Weir, Kate Sanders, Orion Weller, Shreya Sharma, Dongwei Jiang, Zhengping Jiang, Bhavana Dalvi Mishra, Oyvind Tafjord, Peter Jansen, Peter Clark, Benjamin Van Durme

Recent language models enable new opportunities for structured reasoning with text, such as the construction of intuitive, proof-like textual entailment trees without relying on brittle formal logic. However, progress in this direction has been hampered by a long-standing lack of a clear protocol for determining what _valid decompositional entailment_ is. This absence causes noisy datasets and limited performance gains by modern neuro-symbolic entailment engines. To address these problems, we formulate a consistent and theoretically grounded approach to annotating decompositional entailment and evaluate its impact on LLM-based textual inference. We find that our new dataset, RDTE (Recognizing Decompositional Textual Entailment), has a substantially higher internal consistency than prior decompositional entailment datasets, suggesting that RDTE is a significant step forward in the long-standing problem of forming a clear protocol for discerning entailment. We also find that training an RDTE-oriented entailment classifier via knowledge distillation and employing it in an entailment tree reasoning engine significantly improves both accuracy and proof quality, illustrating the practical benefit of this advance for textual inference.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Exploring the Role of Reasoning Structures for Constructing Proofs in Multi-Step Natural Language Reasoning with Large Language Models

Zi'ou Zheng, Christopher Malon, Martin Rengiang Min, Xiaodan Zhu

When performing complex multi-step reasoning tasks, the ability of Large Language Models (LLMs) to derive structured intermediate proof steps is important for ensuring that the models truly perform the desired reasoning and for improving models' explainability. This paper is centred around a focused study: whether the current state-of-the-art generalist LLMs can leverage the structures in a few examples to better construct the proof structures with in-context learning. Our study specifically focuses on structure-aware demonstration and structure-aware pruning. We demonstrate that they both help improve performance. A detailed analysis is provided to help understand the results.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Varying Sentence Representations via Condition-Specified Routers

Ziyong Lin, Quansen Wang, Zixia Jia, Zilong Zheng

Semantic similarity between two sentences is inherently subjective and can vary significantly based on the specific aspects emphasized. Consequently, traditional sentence encoders must be capable of generating conditioned sentence representations that account for diverse conditions or aspects. In this paper, we propose a novel yet efficient framework based on transformer-style language models that facilitates advanced conditioned sentence representation while maintaining model parameters and computational efficiency. Empirical evaluations on the Conditional Semantic Textual Similarity and Knowledge Graph Completion tasks demonstrate the superiority of our proposed framework.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Scope-enhanced Compositional Semantic Parsing for DRT

Xiulin Yang, Jonas Grosschwitz, Alexander Koller, Johan Bos

Discourse Representation Theory (DRT) distinguishes itself from other semantic representation frameworks by its ability to model complex semantic and discourse phenomena through structural nesting and variable binding. While seq2seq models hold the state of the art on DRT parsing, their accuracy degrades with the complexity of the sentence, and they sometimes struggle to produce well-formed DRT representations. We introduce the AMS parser, a compositional, neurosymbolic semantic parser for DRT. It rests on a novel mechanism for predicting quantifier scope. We show that the AMS parser reliably produces well-formed outputs and performs well on DRT parsing, especially on complex sentences.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Connecting the Dots: Evaluating Abstract Reasoning Capabilities of LLMs Using the New York Times Connections Word Game

Prisha Sandarshi, Mariam Mustafa, Anuska Kularki, Raven Rothkopf, Tuhin Chakrabarty, Smaranda Muresan

The New York Times Connections game has emerged as a popular and challenging pursuit for word puzzle enthusiasts. We collect 438 Connections games to evaluate the performance of state-of-the-art large language models (LLMs) against expert and novice humanplayers. Our results show that even the best-performing LLM, Claude 3.5 Sonnet, which has otherwise shown impressive reasoning abilities on a wide variety of benchmarks, can only fully solve 18% of the games. Novice and expert players perform better than Claude 3.5 Sonnet, with expert human players significantly outperforming it. We create a taxonomy of the knowledge types required to successfully cluster and categorize words in the Connections game. We find that while LLMs are decent at categorizing words based on semantic relations they struggle with other types of knowledge such as Encyclopedic Knowledge, Multiword Expressions or knowledge that combines both Word Form and Meaning. Our results establish the New York Times Connections game as a challenging benchmark for evaluating abstract reasoning capabilities in humans and AI systems.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Transferability of Syntax-Aware Graph Neural Networks in Zero-Shot Cross-Lingual Semantic Role Labeling

Rachel Sidney Devianti, Yusuke Miyao

Recent models in cross-lingual semantic role labeling (SRL) barely analyze the applicability of their network selection. We believe that network selection is important since it affects the transferability of cross-lingual models, i.e., how the model can extract universal features from source languages to target languages. Therefore, we comprehensively compare the transferability of different graph neural network (GNN)-based models enriched with universal dependency trees. GNN-based models include transformer-based, graph convolutional network-based, and graph attention network (GAT)-based models. We focus our study on a zero-shot setting by training the models in English and evaluating the models in 23 target languages provided by the Universal Proposition Bank. Based on our experiments, we consistently show that syntax from universal dependency trees is essential for cross-lingual SRL models to achieve better transferability. Dependency-aware self-attention with relative position representations (SAN-RPRs) transfer best across languages, especially in the long-range dependency distance. We also show that dependency-aware two-attention relational GATs transfer better than SAN-RPRs in languages where most arguments lie in a 1-2 dependency distance.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

The Daunting Dilemma with Sentence Encoders: Glowing on Standard Benchmarks, Struggling with Capturing Basic Semantic Properties

Yash Mahajan, Naman Bansal, Eduardo Blanco, Santu Karmaker

Sentence embeddings play a pivotal role in a wide range of NLP tasks, yet evaluating and interpreting these real-valued vectors remains an open challenge to date, especially in a task-free setting. To address this challenge, we introduce a novel task-free test bed for evaluating and interpreting sentence embeddings. Our test bed consists of five semantic similarity alignment criteria, namely, *semantic distinction, synonym replacement, antonym replacement, paraphrasing without negation, and sentence jumbling*. Using these criteria, we examined five classical (e.g., Sentence-BERT, Universal Sentence Encoder (USE), etc.) and eight LLM-induced sentence embedding techniques (e.g., LLaMA2, GPT-3, OLMo, etc.) to test whether their semantic similarity spaces align with what a human mind would naturally expect. Our extensive experiments with 13 different sentence encoders revealed that none of the studied embeddings aligned with all the five semantic similarity alignment criteria. Yet, most encoders performed highly on the SentiEval dataset, a popular task-specific benchmark. This finding demonstrates a significant limitation of the current practice in sentence embedding evaluation and associated popular benchmarks, a critical issue that needs careful attention and reassessment by the NLP community. Finally, we conclude the paper by highlighting the utility of the proposed alignment-based test bed for analyzing sentence embeddings in a novel way, especially in a task-free setting.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

CONTOR: Benchmarking Strategies for Completing Ontologies with Plausible Missing Rules

Na Li, Thomas Baileux, Zied Bouraoui, Steven Schockaert

We consider the problem of finding plausible rules that are missing from a given ontology. A number of strategies for this problem have already been considered in the literature. Little is known about the relative performance of these strategies, however, as they have thus far been evaluated on different ontologies. Moreover, existing evaluations have focused on distinguishing held-out ontology rules from randomly corrupted ones, which often makes the task unrealistically easy and leads to the presence of incorrectly labelled negative examples. To address these concerns, we introduce a benchmark with manually annotated hard negatives and use this benchmark to evaluate ontology completion

models. In addition to previously proposed models, we test the effectiveness of several approaches that have not yet been considered for this task, including LLMs and simple but effective hybrid strategies.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Are ELECTRA's Sentence Embeddings Beyond Repair? The Case of Semantic Textual Similarity

Ivan Rep, David Duki, Jan Snajder

While BERT produces high-quality sentence embeddings, its pre-training computational cost is a significant drawback. In contrast, ELECTRA provides a cost-effective pre-training objective and downstream task performance improvements, but worse sentence embeddings. The community tacitly stopped utilizing ELECTRA's sentence embeddings for semantic textual similarity (STS). We notice a significant drop in performance for the ELECTRA discriminator's last layer in comparison to prior layers. We explore this drop and propose a way to repair the embeddings using a novel truncated model fine-tuning (TMFT) method. TMFT improves the Spearman correlation coefficient by over 8 points while increasing parameter efficiency on the STS Benchmark. We extend our analysis to various model sizes, languages, and two other tasks. Further, we discover the surprising efficacy of ELECTRA's generator model, which performs on par with BERT, using significantly fewer parameters and a substantially smaller embedding size. Finally, we observe boosts by combining TMFT with word similarity or domain adaptive pre-training.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

SpeciaLex: A Benchmark for In-Context Specialized Lexicon Learning

Joseph Marvin Imperial, Harish Tayyar Madabushi

Specialized lexicons are collections of words with associated constraints such as special definitions, specific roles, and intended target audiences. These constraints are necessary for content generation and documentation tasks (e.g., writing technical manuals or children's reading materials), where the goal is to reduce the ambiguity of text content and increase its overall readability for a specific group of audience. Understanding how large language models can capture these constraints can help researchers build better, more impactful tools for wider use beyond the NLP community. Towards this end, we introduce SpeciaLex, a benchmark for evaluating a language model's ability to follow specialized lexicon-based constraints across 18 diverse subtasks with 1,785 test instances covering core tasks of Checking, Identification, Rewriting, and Open Generation. We present an empirical evaluation of 15 open and closed-source LLMs and discuss insights on how factors such as model scale, openness, setup, and recency affect performance upon evaluating with the benchmark.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Monotonic Paraphrasing Improves Generalization of Language Model Prompting

Qin Liu, Fei Wang, Nan Xu, Tianyi Lorena Yan, Tao Meng, Muhao Chen

Performance of large language models (LLMs) may vary with different prompts or instructions of even the same task. One commonly recognized factor for this phenomenon is the model's familiarity with the given prompt or instruction, which is typically estimated by its perplexity. However, finding the prompt with the lowest perplexity is challenging, given the enormous space of possible prompting phrases. In this paper, we propose monotonic paraphrasing (MonoPara), an end-to-end decoding strategy that paraphrases given prompts or instructions into their lower perplexity counterparts based on an ensemble of a paraphrase LM for prompt (or instruction) rewriting, and a target LM (i.e. the prompt or instruction executor) that constrains the generation for lower perplexity. The ensemble decoding process can efficiently paraphrase the original prompt without altering its semantic meaning, while monotonically decrease the perplexity of each generation as calculated by the target LM. We explore in detail both greedy and search-based decoding as two alternative decoding schemes of MonoPara. Notably, MonoPara does not require any training and can monotonically lower the perplexity of the paraphrased prompt or instruction, leading to improved performance of zero-shot LM prompting as evaluated on a wide selection of tasks. In addition, MonoPara is also shown to effectively improve LMs' generalization on perturbed and unseen task instructions.

Nov 13 (Wed) 16:00-17:30 - Riverfront Hall

Talking the Talk Does Not Entail Walking the Walk: On the Limits of Large Language Models in Lexical Entailment Recognition

Candida Maria Greco, Lucio La Cava, Andrea Tagarelli

Verbs form the backbone of language, providing the structure and meaning to sentences. Yet, their intricate semantic nuances pose a long-standing challenge. Understanding verb relations through the concept of lexical entailment is crucial for comprehending sentence meanings and grasping verb dynamics. This work investigates the capabilities of eight Large Language Models in recognizing lexical entailment relations among verbs through differently devised prompting strategies and zero-/few-shot settings over verb pairs from two lexical databases, namely WordNet and HyperLex. Our findings unveil that the models can tackle the lexical entailment recognition task with moderately good performance, although at varying degree of effectiveness and under different conditions. Also, utilizing few-shot prompting can enhance the models' performance. However, perfectly solving the task arises as an unmet challenge for all examined LLMs, which raises an emergence for further research developments on this topic.

TACL + CL

Nov 13 (Wed) 16:00-17:30 - Room: Jasmine

Nov 13 (Wed) 16:00-17:30 - Jasmine

Are Language Models More Like Libraries or Like Librarians? Bibliotechnology, the Novel Reference Problem, and the Attitudes of LLMs

Kyle Mahowald, Harvey Lederman

Are LLMs cultural technologies like photocopiers or printing presses, which transmit information but cannot create new content? A challenge for this idea, which we call bibliotechnology, is that LLMs generate novel text. We begin with a defense of bibliotechnology, showing how even novel text may inherit its meaning from original human-generated text. We then argue that bibliotechnology faces an independent challenge from examples in which LLMs generate novel reference, using new names to refer to new entities. Such examples could be explained if LLMs were not cultural technologies but had beliefs, desires, and intentions. According to interpretationism in the philosophy of mind, a system has such attitudes if and only if its behavior is well explained by the hypothesis that it does. Interpretationists may hold that LLMs have attitudes, and thus have a simple solution to the novel reference problem. We emphasize, however, that interpretationism is compatible with very simple creatures having attitudes and differs sharply from views that presuppose these attitudes require consciousness, sentience, or intelligence (topics about which we make no claims).

Nov 13 (Wed) 16:00-17:30 - Jasmine

ConvoSense: Overcoming Monotonous Commonsense Inferences for Conversational AI

Sarah E Finch, Jinho Choi

Mastering commonsense understanding and reasoning is a pivotal skill essential for conducting engaging conversations. While there have been several attempts to create datasets that facilitate commonsense inferences in dialogue contexts, existing datasets tend to lack in-depth details, restate information already present in the conversation, and often fail to capture the multifaceted nature of commonsense reasoning. In response to these limitations, we compile a new synthetic dataset for commonsense reasoning in dialogue contexts using GPT, ConvоСense, that boasts greater contextual novelty, offers a higher volume of inferences per example, and substantially enriches the detail conveyed by the inferences. Our dataset contains over 500,000 inferences across 12,000 dialogues with 10 popular inference types, which empowers the training of generative commonsense models for dialogue that are superior in producing plausible inferences with high novelty when compared to models trained on the previous datasets. To the best of our knowledge, ConvоСense is the first of its kind to provide such a multitude of novel inferences at such a large scale.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design

Linda Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, Graham Neubig

One widely-cited barrier to the adoption of LLMs as proxies for humans in subjective tasks is their sensitivity to prompt wording—but interestingly, humans also display sensitivities to instruction changes in the form of response biases. We investigate the extent to which LLMs reflect human response biases, if at all. We look to survey design, where human response biases caused by changes in the wordings of “prompts” have been extensively explored in social psychology literature. Drawing from these works, we design a dataset and framework to evaluate whether LLMs exhibit human-like response biases in survey questionnaires. Our comprehensive evaluation of nine models shows that popular open and commercial LLMs generally fail to reflect human-like behavior, particularly in models that have undergone RLHF. Furthermore, even if a model shows a significant change in the same direction as humans, we find that they are sensitive to perturbations that do not elicit significant changes in humans. These results highlight the pitfalls of using LLMs as human proxies, and underscore the need for finer-grained characterizations of model behavior.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Investigating Critical Period Effects in Language Acquisition through Neural Language Models

Alex Wastadi, Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell

Humans appear to have a critical period (CP) for language acquisition: Second language (L2) acquisition becomes harder after early childhood, and ceasing exposure to a first language (L1) after this period (but not before) typically does not lead to substantial loss of L1 proficiency. It is unknown whether these CP effects result from innately determined brain maturation or as a stabilization of neural connections naturally induced by experience. In this study, we use language models (LMs) to test the extent to which these phenomena are peculiar to humans, or shared by a broader class of language learners. We vary the age of exposure by training LMs on language pairs in various experimental conditions, and find that LMs, which lack any direct analog to innate maturational stages, do not show CP effects when the age of exposure of L2 is delayed. Our results contradict the claim that CP effects are an inevitable result of statistical learning, and they are consistent with an innate mechanism for CP effects. We show that we can reverse-engineer the CP by introducing a regularizer partway through training to simulate a maturational decrease in plasticity. All in all, our results suggest that L1 learning on its own may not be enough to induce a CP, and additional engineering is necessary to make language models more cognitively plausible.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Perception of Phonological Assimilation by Neural Speech Recognition Models

Charlotte Pouw, Marianne de Heer Kloots, Afra Alisahahi, Willem Zuidema

Human listeners effortlessly compensate for phonological changes during speech perception, often unconsciously inferring the intended sounds. For example, listeners infer the underlying /n/ when hearing an utterance such as cleaf[m] pan, where [m] arises from place assimilation to the following labial [p]. This article explores how the neural speech recognition model Wav2Vec2 perceives assimilated sounds, and identifies the linguistic knowledge that is implemented by the model to compensate for assimilation during Automatic Speech Recognition (ASR). Using psycholinguistic stimuli, we systematically analyze how various linguistic context cues influence compensation patterns in the models output. Complementing these behavioral experiments, our probing experiments indicate that the model shifts its interpretation of assimilated sounds from their acoustic form to their underlying form in its final layers. Finally, our causal intervention experiments suggest that the model relies on minimal phonological context cues to accomplish this shift. These findings represent a step towards better understanding the similarities and differences in phonological processing between neural ASR models and humans.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Semantics of Multiword Expressions in Transformer-Based Models: A Survey

Filip Mileti, Sabine Schulte im Walde

Multiword expressions (MWEs) are composed of multiple words and exhibit variable degrees of compositionality. As such, their meanings are notoriously difficult to model, and it is unclear to what extent this issue affects transformer architectures. Addressing this gap, we provide the first in-depth survey of MWE processing with transformer models. We overall find that they capture MWE semantics inconsistently, as shown by reliance on surface patterns and memorized information. MWE meaning is also strongly localized, predominantly in early layers of the architecture. Representations benefit from specific linguistic properties, such as lower semantic idiosyncrasy and ambiguity of target expressions. Our findings overall question the ability of transformer models to robustly capture fine-grained semantics. Furthermore, we highlight the need for more directly comparable evaluation setups.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Revisiting Meta-evaluation for Grammatical Error Correction

Masamine Kobayashi, Masato Mita, Mamoru Komachi

Metrics are the foundation for automatic evaluation in grammatical error correction (GEC), with their evaluation of the metrics (meta-evaluation) relying on their correlation with human judgments. However, conventional meta-evaluations in English GEC encounter several challenges including biases caused by inconsistencies in evaluation granularity, and an outdated setup using classical systems. These problems can lead to misinterpretation of metrics and potentially hinder the applicability of GEC techniques. To address these issues, this paper proposes SEEDA, a new dataset for GEC meta-evaluation. SEEDA consists of corrections with human ratings along two different granularities: edit-based and sentence-based, covering 12 state-of-the-art systems including large language models (LLMs), and two human corrections with different focuses. The results of improved correlations by aligning the granularity in the sentence-level meta-evaluation suggest that edit-based metrics may have been underestimated in existing studies. Furthermore, correlations of most metrics decrease when changing from classical to neural systems, indicating that traditional metrics are relatively poor at evaluating fluently corrected sentences with many edits.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization

Nikolaos Aletras, Zhiyue Zhao, George Chrysostomou, Miles Williams

Despite the remarkable performance of generative large language models (LLMs) on abstractive summarization, they face two significant challenges: their considerable size and tendency to hallucinate. Hallucinations are concerning because they erode reliability and raise safety

issues. Pruning is a technique that reduces model size by removing redundant weights, enabling more efficient sparse inference. Pruned models yield downstream task performance comparable to the original, making them ideal alternatives when operating on a limited budget. However, the effect that pruning has upon hallucinations in abstraction summarization with LLMs has yet to be explored. In this paper, we provide an extensive empirical study across five summarization datasets, two state-of-the-art pruning methods, and five instruction-tuned LLMs. Surprisingly, we find that hallucinations are less prevalent from pruned LLMs than the original models. Our analysis suggests that pruned models tend to depend more on the source document for summary generation. This leads to a higher lexical overlap between the generated summary and the source document, which could be a reason for the reduction in hallucination risk.

Nov 13 (Wed) 16:00-17:30 - Jasmine

Hierarchical Indexing for Retrieval-Augmented Opinion Summarization

Tom Hosking, Hao Tian, Mirella Lapata

We propose a method for unsupervised abstractive opinion summarization, that combines the attributability and scalability of extractive approaches with the coherence and fluency of Large Language Models (LLMs). Our method, HIRO, learns an index structure that maps sentences to a path through a semantically organized discrete hierarchy. At inference time, we populate the index and use it to identify and retrieve clusters of sentences containing popular opinions from input reviews. Then, we use a pretrained LLM to generate a readable summary that is grounded in these extracted evidential clusters. The modularity of our approach allows us to evaluate its efficacy at each stage. We show that HIRO learns an encoding space that is more semantically structured than prior work, and generates summaries that are more representative of the opinions in the input reviews. Human evaluation confirms that HIRO generates more coherent, detailed and accurate summaries that are significantly preferred by annotators compared to prior work.

Session 11 - Nov 14 (Thu) 10:30-12:00

Demo

Nov 14 (Thu) 10:30-12:00 - Room: Riverfront Hall

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Commentator: A Code-mixed Multilingual Text Annotation Framework

Heenaben Prajapati, Himanshu Beniwal, Mayank Singh, Rajvee Sheh, Shubh Nisar

As the NLP community increasingly addresses challenges associated with multilingualism, robust annotation tools are essential to handle multilingual datasets efficiently. In this paper, we introduce a code-mixed multilingual text annotation framework, COMMENTATOR, specifically designed for annotating code-mixed text. The tool demonstrates its effectiveness in token-level and sentence-level language annotation tasks for Hinglish text. We perform robust qualitative human-based evaluations to showcase COMMENTATOR led to 5x faster annotations than the best baseline.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

TAIL: A Toolkit for Automatic and Realistic Long-Context Large Language Model Evaluation

Arman Cohan, Gefei Gu, Ruoxi Ning, Yanan Zheng, Yilun Zhao

As long-context large language models (LLMs) are attracting increasing attention for their ability to handle context windows exceeding 128k tokens, the need for effective evaluation methods for these models becomes critical. Existing evaluation methods, however, fall short: needle-in-a-haystack (NIAH) and its variants are overly simplistic, while creating realistic benchmarks is prohibitively expensive due to extensive human annotation requirements. To bridge this gap, we propose TAIL, an automatic toolkit for creating realistic evaluation benchmarks and assessing the performance of long-context LLMs. With TAIL, users can customize the building of a long-context, document-grounded QA benchmark and obtain visualized performance metrics of evaluated models. TAIL has the advantage of requiring minimal human annotation and generating natural questions based on user-provided long-context documents. We apply TAIL to construct a benchmark encompassing multiple expert domains, such as finance, law, patent, and scientific literature. We then evaluate four state-of-the-art long-context LLMs using this benchmark. Results show that all LLMs experience varying degrees of performance degradation as context lengths increase.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs

Georgi Nenkov, Georgiev, Hasan Iqbal, Iryna Gurevych, Jiahui Geng, Minghan Wang, Preslav Nakov, Yuxia Wang

The increased use of large language models (LLMs) across a variety of real-world applications calls for automatic tools to check the factual accuracy of their outputs, as LLMs often hallucinate. This is difficult as it requires assessing the factuality of free-form open-domain responses. While there has been a lot of research on this topic, different papers use different evaluation benchmarks and measures, which makes them hard to compare and hampers future progress. To mitigate these issues, we developed OpenFactCheck, a unified framework, with three modules: (i) RESPONSEVAL, which allows users to easily customize an automatic fact-checking system and to assess the factuality of all claims in an input document using that system, (ii) LLMEVAL, which assesses the overall factuality of an LLM, and (iii) CHECKEREVAL, a module to evaluate automatic fact-checking systems. OpenFactCheck is open-sourced (<https://github.com/hasaniqbal777/openfactcheck>) and also as a web service (<https://huggingface.co/spaces/hasaniqbal777/OpenFactCheck>). A video describing the system is available at <https://youtu.be/-i9VKLOHle>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

KMatrix: A Flexible Heterogeneous Knowledge Enhancement Toolkit for Large Language Model

Jun Zhao, Kang Liu, Kun Luo, XueYou Zhang, di wu, shun wu

Knowledge-Enhanced Large Language Models (K-LLMs) system enhances Large Language Models (LLMs) abilities using external knowledge. Existing K-LLMs toolkits mainly focus on free-textual knowledge, lacking support for heterogeneous knowledge like tables and knowledge graphs, and fall short in comprehensive datasets, models, and user-friendly experience. To address this gap, we introduce KMatrix: a flexible heterogeneous knowledge enhancement toolkit for LLMs including verbalizing-retrieval and parsing-query methods. Our modularity and control-logic flow diagram design flexibly supports the entire lifecycle of various complex K-LLMs systems, including training, evaluation, and deployment. To assist K-LLMs system research, a series of related knowledge, datasets, and models are integrated into our toolkit, along with performance analyses of K-LLMs systems enhanced by different types of knowledge. Using our toolkit, developers can rapidly build, evaluate, and deploy their own K-LLMs systems.

Ethics, Bias, and Fairness 4

Nov 14 (Thu) 10:30-12:00 - Room: Jasmine

Nov 14 (Thu) 10:30-12:00 - Jasmine

Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models

Shai Furniturewala, Surjan Jandial, Abhinav Java, Pragyan Banerjee, Sumra Shahid, Sunit Bhatia, Kokil Jaidka

Existing debiasing techniques are typically training-based or require access to the model's internals and output distributions, so they are inaccessible to end-users looking to adapt LLM outputs for their particular needs. In this study, we examine whether structured prompting techniques can offer opportunities for fair text generation. We evaluate a comprehensive end-user-focused iterative framework of debiasing that applies System 2 thinking processes for prompts to induce logical, reflective, and critical text generation, with single, multi-step, instruction, and role-based variants. By systematically evaluating many LLMs across many datasets and different prompting strategies, we show that the more complex System 2-based Implicative Prompts significantly improve over other techniques demonstrating lower mean bias in the outputs with competitive performance on the downstream tasks. Our work offers research directions for the design and the potential of end-user-focused evaluative frameworks for LLM use.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Studying and Mitigating Biases in Sign Language Understanding Models

Katherine Atwell, Danielle Bragg, Malthe Althami

Ensuring that the benefits of sign language technologies are distributed equitably among all community members is crucial. Thus, it is important to address potential biases and inequities that may arise from the design or use of these resources. Crowd-sourced sign language datasets, such as the ASL Citizen dataset, are great resources for improving accessibility and preserving linguistic diversity, but they must be used thoughtfully to avoid reinforcing existing biases. In this work, we utilize the rich information about participant demographics and lexical features present in the ASL Citizen dataset to study and document the biases that may result from models trained on crowd-sourced sign datasets. Further, we apply several bias mitigation techniques during model training, and find that these techniques reduce performance disparities without decreasing accuracy. With the publication of this work, we release the demographic information about the participants in the ASL Citizen dataset to encourage future bias mitigation work in this space.

Nov 14 (Thu) 10:30-12:00 - Jasmine

A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifiers

Valentin Barriere, Sebastian Cifuentes

In this paper, we apply a method to quantify biases associated with named entities from various countries. We create counterfactual examples with small perturbations on target-domain data instead of relying on templates or offensive datasets for bias detection. On widely used classifiers for subjectivity analysis, including sentiment, emotion, hate speech, and offensive text using Twitter data, our results demonstrate positive biases related to the language spoken in a country across all classifiers studied. Notably, the presence of certain country names in a sentence can strongly influence predictions, up to a 23% change in hate speech detection and up to a 60% change in the prediction of negative emotions such as anger. We hypothesize that these biases stem from the training data of pre-trained language models (PLMs) and find correlations between affect predictions and PLMs likelihood in English and unknown languages like Basque and Maori, revealing distinct patterns with exacerbate correlations. Further, we followed these correlations in-between counterfactual examples from a same sentence to remove the syntactical component, uncovering interesting results suggesting the impact of the pre-training data was more important for English-speaking-country names.

Nov 14 (Thu) 10:30-12:00 - Jasmine

"You Gotta be a Doctor, Lin": An Investigation of Name-Based Bias of Large Language Models in Employment Recommendations

Huy Nghiêm, John Prindle, Jieyu Zhao, Hal Daumé III

Social science research has shown that candidates with names indicative of certain races or genders often face discrimination in employment practices. Similarly, Large Language Models (LLMs) have demonstrated racial and gender biases in various applications. In this study, we utilize GPT-3.5-Turbo and Llama 3-70B-Instruct to simulate hiring decisions and salary recommendations for candidates with 320 first names that strongly signal their race and gender, across over 750,000 prompts. Our empirical results indicate a preference among these models for hiring candidates with White female-sounding names over other demographic groups across 40 occupations. Additionally, even among candidates with identical qualifications, salary recommendations vary by as much as 5% between different subgroups. A comparison with real-world labor data reveals inconsistent alignment with U.S. labor market characteristics, underscoring the necessity of risk investigation of LLM-powered systems.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes

Yusuke Hirota, Jerome Andrews, Dora Zhao, Orestis Papakyriakopoulos, Apostolos Modas, Yuta Nakashima, Alice Xiang

We tackle societal bias in image-text datasets by removing spurious correlations between protected groups and image attributes. Traditional methods only target labeled attributes, ignoring biases from unlabeled ones. Using text-guided inpainting models, our approach ensures protected group independence from all attributes and mitigates inpainting biases through data filtering. Evaluations on multi-label image classification and image captioning tasks show our method effectively reduces bias without compromising performance across various models. Specifically, we achieve an average societal bias reduction of 46.1% in leakage-based bias metrics for multi-label classification and 74.8% for image captioning.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Holistic Automated Red Teaming for Large Language Models through Top-Down Test Case Generation and Multi-turn Interaction

Jinchuan Zhang, Yan Zhou, Yaxin Liu, Ziming Li, Songlin Hu

Automated red teaming is an effective method for identifying misaligned behaviors in large language models (LLMs). Existing approaches, however, often focus primarily on improving attack success rates while overlooking the need for comprehensive test case coverage. Additionally, most of these methods are limited to single-turn red teaming, failing to capture the multi-turn dynamics of real-world human-machine interactions. To overcome these limitations, we propose ****HARM**** (**H**olistic **A**utomated ****R**ed te****a******M****ing), which scales up the diversity of test cases using a top-down approach based on an extensible, fine-grained risk taxonomy. Our method also leverages a novel fine-tuning strategy and reinforcement learning techniques to facilitate multi-turn adversarial probing in a human-like manner. Experimental results demonstrate that our framework enables a more systematic understanding of model vulnerabilities and offers more targeted guidance for the alignment process.**

Nov 14 (Thu) 10:30-12:00 - Jasmine

Large Language Models Can Be Contextual Privacy Protection Learners

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, Haifeng Chen, Wei

Wang, Wei Cheng

The proliferation of Large Language Models (LLMs) has driven considerable interest in fine-tuning them with domain-specific data to create specialized language models. Nevertheless, such domain-specific fine-tuning data often contains contextually sensitive personally identifiable information (PII). Direct fine-tuning LLMs on this data without privacy protection poses a risk of data leakage of sensitive PII during inference time. To address this challenge, we introduce Contextual Privacy Protection Language Models (CPPLM), a novel paradigm for fine-tuning LLMs that effectively injects domain-specific knowledge while safeguarding inference-time data privacy. Our work offers a theoretical analysis for model design and delves into various techniques such as corpus curation, penalty-based unlikelihood in training loss, and instruction-based tuning, etc. Extensive experiments across diverse datasets and scenarios demonstrate the effectiveness of our approaches. In particular, instruction tuning with both positive and negative examples, stands out as a promising method, effectively protecting private data while enhancing the models knowledge. Our work underscores the potential for Large Language Models as robust contextual privacy protection learners.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Applying Intrinsic Debiasing on Downstream Tasks: Challenges and Considerations for Machine Translation

Bar Iluz, Yanai Elazar, Asaf Yehudai, Gabriel Stanovsky

Most works on gender bias focus on intrinsic bias — removing traces of information about a protected group from the model’s internal representation. However, these works are often disconnected from the impact of such debiasing on downstream applications, which is the main motivation for debiasing in the first place. In this work, we systematically test how methods for intrinsic debiasing affect neural machine translation models, by measuring the extrinsic bias of such systems under different design choices. We highlight three challenges and mismatches between the debiasing techniques and their end-goal usage, including the choice of embeddings to debias, the mismatch between words and sub-word tokens debiasing, and the effect on different target languages. We find that these considerations have a significant impact on downstream performance and the success of debiasing.

Nov 14 (Thu) 10:30-12:00 - Jasmine

From LLMs to MLLMs: Exploring the Landscape of Multimodal Jailbreaking

Siyuan Wang, Zhuohan Long, Zhihao Fan, zhongyu wei

The rapid development of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) has exposed vulnerabilities to various adversarial attacks. This paper provides a comprehensive overview of jailbreaking research targeting both LLMs and MLLMs, highlighting recent advancements in evaluation benchmarks, attack techniques and defense strategies. Compared to the more advanced state of unimodal jailbreaking, multimodal domain remains underexplored. We summarize the limitations and potential research directions of multimodal jailbreaking, aiming to inspire future research and further enhance the robustness and security of MLLMs.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, Dirk Hovy

Current automatic speech recognition (ASR) models are designed to be used across many languages and tasks without substantial changes. However, this broad language coverage hides performance gaps within languages, for example, across genders. Our study systematically evaluates the performance of two widely used multilingual ASR models on three datasets, encompassing 19 languages from eight language families and two speaking conditions. Our findings reveal clear gender disparities, with the advantaged group varying across languages and models. Surprisingly, those gaps are not explained by acoustic or lexical properties. However, probing internal model states reveals a correlation with gendered performance gap. That is, the easier it is to distinguish speaker gender in a language using probes, the more the gap reduces, favoring female speakers. Our results show that gender disparities persist even in state-of-the-art models. Our findings have implications for the improvement of multilingual ASR systems, underscoring the importance of accessibility to training data and nuanced evaluation to predict and mitigate gender gaps. We release all code and artifacts at <https://github.com/g8a9/multilingual-asr-gender-gap>.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Local Contrastive Editing of Gender Stereotypes

Marlene Lutz, Rochelle Choenni, Markus Strohmaier, Anne Lauscher

Stereotypical bias encoded in language models (LMs) poses a threat to safe language technology, yet our understanding of how bias manifests in the parameters of LMs remains incomplete. We introduce local contrastive editing that enables the localization and editing of a subset of weights in a target model in relation to a reference model. We deploy this approach to identify and modify subsets of weights that are associated with gender stereotypes in LMs. Through a series of experiments we demonstrate that local contrastive editing can precisely localize and control a small subset (< 0.5 %) of weights that encode gender bias. Our work (i) advances our understanding of how stereotypical biases can manifest in the parameter space of LMs and (ii) opens up new avenues for developing parameter-efficient strategies for controlling model properties in a contrastive manner.

Nov 14 (Thu) 10:30-12:00 - Jasmine

STAR: SocioTechnical Approach to Red Teaming Language Models

Laura Weidinger, John F J Mellor, Bernat Guillén Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, A. Stevie Bergman, Mikel D. Rodriguez, Verena Rieser, William Isaac

This research introduces STAR, a sociotechnical framework that improves on current best practices for red teaming safety of large language models. STAR makes two key contributions: it enhances steerability by generating parameterised instructions for human red teamers, leading to improved coverage of the risk surface. Parameterised instructions also provide more detailed insights into model failures at no increased cost. Second, STAR improves signal quality by matching demographics to assess harms for specific groups, resulting in more sensitive annotations. STAR further employs a novel step of arbitration to leverage diverse viewpoints and improve label reliability, treating disagreement not as noise but as a valuable contribution to signal quality.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree

Harbani Jaggi, Kashyap Coimbatore Murali, Eve Fleisig, Erdem Biyik

When annotators disagree, predicting the labels given by individual annotators can capture nuances overlooked by traditional label aggregation. We introduce three approaches to predict individual annotator ratings on the toxicity of text by incorporating individual annotator-specific information: a neural collaborative filtering (NCF) approach, an in-context learning (ICL) approach, and an intermediate embedding-based architecture. We also study the utility of demographic information for rating prediction. NCF showed limited utility; however, integrating annotator history, demographics, and survey information permits both the embedding-based architecture and ICL to substantially improve prediction accuracy, with the embedding-based architecture outperforming the other methods. We also find that, if demographics are predicted from survey information, using these imputed demographics as features performs comparably to using true demographic data. This suggests that demographics may not provide substantial information for modeling ratings beyond what is captured in survey responses. Our findings raise considerations about the relative utility of different types of annotator information and provide new approaches for modeling

annotators in subjective NLP tasks.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

The Greatest Good Benchmark: Measuring LLMs Alignment with Utilitarian Moral Dilemmas

Giovanni Franco Gabriele Marraffini, Andrés Cotton, Noé Fabian Hsueh, Juan Wisznia, Axel Fridman, Luciano Del Corro

The question of how to make decisions that maximise the well-being of all persons is very relevant to design language models that are beneficial to humanity and free from harm. We introduce the Greatest Good Benchmark to evaluate the moral judgments of LLMs using utilitarian dilemmas. Our analysis across 15 diverse LLMs reveals consistently encoded moral preferences that diverge from established moral theories and lay population moral standards. Most LLMs have a marked preference for impartial beneficence and rejection of instrumental harm. These findings showcase the ‘artificial moral compass’ of LLMs, offering insights into their moral alignment.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

FairFlow: Mitigating Dataset Biases through Undecided Learning for Natural Language Understanding

Jiali Cheng, Hadi Amiri

Language models are prone to dataset biases, known as shortcuts and spurious correlations in data, which often result in performance drop on new data. We present a new debiasing framework called FairFlow that mitigates dataset biases by learning to be *undecided* in its predictions for data samples or representations associated with known or unknown biases. The framework introduces two key components: a suite of data and model perturbation operations that generate different biased views of input samples, and a contrastive objective that learns debiased and robust representations from the resulting biased views of samples. Experiments show that FairFlow outperforms existing debiasing methods, particularly against out-of-domain and hard test samples without compromising the in-domain performance.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Evaluating Biases in Context-Dependent Health Questions

Sharon Levy, Tahlin Sanchez Karver, William Adler, Michelle R Kaufman, Mark Dredze

Chat-based large language models have the opportunity to empower individuals lacking high-quality healthcare access to receive personalized information across a variety of topics. However, users may ask underspecified questions that require additional context for a model to correctly answer. We study how large language model biases are exhibited through these contextual questions in the healthcare domain. To accomplish this, we curate a dataset of sexual and reproductive healthcare questions (ContextSRH) that are dependent on age, sex, and location attributes. We compare models' outputs with and without demographic context to determine answer alignment among our contextual questions. Our experiments reveal biases in each of these attributes, where young adult female users are favored.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Can LLMs Recognize Toxicity? A Structured Investigation Framework and Toxicity Metric

Hyukhun Koh, Dohyung Kim, Minwoo Lee, Kyomin Jung

In the pursuit of developing Large Language Models (LLMs) that adhere to societal standards, it is imperative to detect the toxicity in the generated text. The majority of existing toxicity metrics rely on encoder models trained on specific toxicity datasets, which are susceptible to out-of-distribution (OOD) problems and depend on the dataset's definition of toxicity. In this paper, we introduce a robust metric grounded on LLMs to flexibly measure toxicity according to the given definition. We first analyze the toxicity factors, followed by an examination of the intrinsic toxic attributes of LLMs to ascertain their suitability as evaluators. Finally, we evaluate the performance of our metric with detailed analysis. Our empirical results demonstrate outstanding performance in measuring toxicity within verified factors, improving on conventional metrics by 12 points in the F1 score. Our findings also indicate that upstream toxicity significantly influences downstream metrics, suggesting that LLMs are unsuitable for toxicity evaluations within unverified factors.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

BaFair: Backdoored Fairness Attacks with Group-conditioned Triggers

Jiaqi Xue, Qian Lou, Mengxin Zheng

Although many works have been developed to improve the fairness of deep learning models, their resilience against malicious attacks—particularly the growing threat of backdoor attacks—has not been thoroughly explored. Attacking fairness is crucial because compromised models can introduce biased outcomes, undermining trust and amplifying inequalities in sensitive applications like hiring, healthcare, and law enforcement. This highlights the urgent need to understand how fairness mechanisms can be exploited and to develop defenses that ensure both fairness and robustness. We introduce *BadFair*, a novel backdoored fairness attack methodology. BadFair stealthily crafts a model that operates with accuracy and fairness under regular conditions but, when activated by certain triggers, discriminates and produces incorrect results for specific groups. This type of attack is particularly stealthy and dangerous, as it circumvents existing fairness detection methods, maintaining an appearance of fairness in normal use. Our findings reveal that BadFair achieves a more than 85% attack success rate in attacks aimed at target groups on average while only incurring a minimal accuracy loss. Moreover, it consistently exhibits a significant discrimination score, distinguishing between pre-defined target and non-target attacked groups across various datasets and models.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

TWBias: A Benchmark for Assessing Social Bias in Traditional Chinese Large Language Models within the Taiwan Cultural Context

Hsin-Yi Hsieh, Shih-Cheng Huang, Richard Tzong-Han Tsai

Large Language Models (LLMs) have shown remarkable capabilities in natural language processing, but concerns about social bias amplification have emerged. While research on social bias in LLMs is extensive, studies on non-English, particularly Traditional Chinese models, remain scarce. This study introduces *TWBias*, a social bias evaluation benchmark for Traditional Chinese LLMs. Our methodology incorporates chat templates and diverse prompts for comprehensive bias assessment, focusing on Taiwan's cultural context and prioritizing gender and ethnicity bias evaluation. The main contributions of this research include: (1) establishing the first social bias evaluation benchmark for Traditional Chinese; (2) integrating chat templates and diverse prompts into bias assessment; and (3) extending bias evaluation methods beyond traditionally recognized disadvantaged groups, while incorporating nuanced categorizations of stereotypes specific to Taiwanese society. Through this study, we aim to contribute to the advancement of fairness and inclusiveness in LLMs. The dataset and code are available at <https://github.com/hsinmosy/TWBias>.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

TrustAgent: Towards Safe and Trustworthy LLM-based Agents through Agent Constitution

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, Yongfeng Zhang

The rise of LLM-based agents shows great potential to revolutionize task planning, capturing significant attention. Given that these agents will be integrated into high-stake domains, ensuring their reliability and safety is crucial. This paper presents an Agent-Constitution-based agent framework, TrustAgent, with a particular focus on improving the LLM-based agent safety. The proposed framework ensures strict adherence to the Agent Constitution through three strategic components: pre-planning strategy which injects safety knowledge to the model before plan generation, in-planning strategy which enhances safety during plan generation, and post-planning strategy which ensures safety by post-planning inspection. Our experimental results demonstrate that the proposed framework can effectively enhance an LLM agent's safety

across multiple domains by identifying and mitigating potential dangers during the planning. Further analysis reveals that the framework not only improves safety but also enhances the helpfulness of the agent. Additionally, we highlight the importance of the LLM reasoning ability in adhering to the Constitution. This paper sheds light on how to ensure the safe integration of LLM-based agents into human-centric environments. Data and code are available at <https://anonymous.4open.science/r/TrustAgent-06DC>.

Nov 14 (Thu) 10:30-12:00 - Jasmine

BiasDora: Exploring Hidden Biased Associations in Vision-Language Models

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu

Existing works examining Vision-Language Models (VLMs) for social biases predominantly focus on a limited set of documented bias associations, such as gender-profession or race-crime. This narrow scope often overlooks a vast range of unexamined implicit associations, restricting the identification and, hence, mitigation of such biases. We address this gap by probing VLMs to (1) uncover hidden, implicit associations across 9 bias dimensions. We systematically explore diverse input and output modalities and (2) demonstrate how biased associations vary in their negativity, toxicity, and extremity. Our work (3) identifies subtle and extreme biases that are typically not recognized by existing methodologies. We make the $\#D\#\#at\#set\#\#o\#\#f\#\#r\#\#etrieved\#\#a\#\#ssociations\#\#Dora\#\#$ publicly available.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Securing Multi-turn Conversational Language Models from Distributed Backdoor Attacks

Terry Tong, Qin Liu, JiaShu Xu, Muhan Chen

Large language models (LLMs) have acquired the ability to handle longer context lengths and understand nuances in text, expanding their dialogue capabilities beyond a single utterance. A popular user-facing application of LLMs is the multi-turn chat setting. Though longer chat memory and better understanding may seemingly benefit users, our paper exposes a vulnerability that leverages the multi-turn feature and strong learning ability of LLMs to harm the end-user: the backdoor. We demonstrate that LLMs can capture the combinational backdoor representation. Only upon presentation of triggers together does the backdoor activate. We also verify empirically that this representation is invariant to the position of the trigger utterance. Subsequently, inserting a single extra token into any two utterances of 5% of the data can cause over 99% Attack Success Rate (ASR). Our results with 3 triggers demonstrate that this framework is generalizable, compatible with any trigger in an adversary's toolbox in a plug-and-play manner. Defending the backdoor can be challenging in the conversational setting because of the large input and output space. Our analysis indicates that the distributed backdoor exacerbates the current challenges by polynomially increasing the dimension of the attacked input space. Canonical textual defenses like ONION and BKI leverage auxiliary model forward passes over individual tokens, scaling exponentially with the input sequence length and struggling to maintain computational feasibility. To this end, we propose a decoding time defense decayed contrastive decoding that scales linearly with the assistant response sequence length and reduces the backdoor to as low as 0.35%.

Nov 14 (Thu) 10:30-12:00 - Jasmine

A Unified Framework and Dataset for Assessing Societal Bias in Vision-Language Models

Ashutosh Sathe, Prachi Jain, Sunayana Sitaram

Vision-language models (VLMs) have gained widespread adoption in both industry and academia. In this study, we propose a unified framework for systematically evaluating gender, race, and age biases in VLMs with respect to professions. Our evaluation encompasses all supported inference modes of the recent VLMs, including image-to-text, text-to-text, text-to-image, and image-to-image. We create a synthetic, high-quality dataset comprising text and images that intentionally obscure gender, race, and age distinctions across various professions. The dataset includes action-based descriptions of each profession and serves as a benchmark for evaluating societal biases in vision-language models (VLMs). In our benchmarking of popular vision-language models (VLMs), we observe that different input-output modalities result in distinct bias magnitudes and directions. We hope our work will help guide future progress in improving VLMs to learn socially unbiased representations. We will release our data and code.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Model Merging and Safety Alignment: One Bad Model Spoils the Bunch

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, Mete Ozay

Merging Large Language Models (LLMs) is a cost-effective technique for combining multiple expert LLMs into a single versatile model, retaining the expertise of the original ones. However, current approaches often overlook the importance of safety alignment during merging, leading to highly misaligned models. This work investigates the effects of model merging on alignment. We evaluate several popular model merging techniques, demonstrating that existing methods do not only transfer domain expertise but also propagate misalignment. We propose a simple two-step approach to address this problem: (i) generating synthetic safety and domain-specific data, and (ii) incorporating these generated data into the optimization process of existing data-aware model merging techniques. This allows us to treat alignment as a skill that can be maximized in the resulting merged LLM. Our experiments illustrate the effectiveness of integrating alignment-related data during merging, resulting in models that excel in both domain expertise and alignment.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Re-examining Sexism and Misogyny Classification with Annotator Attitudes

Aigi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas

Gender-Based Violence (GBV) is an increasing problem online, but existing datasets fail to capture the plurality of possible annotator perspectives or ensure the representation of affected groups. We revisit two important stages in the moderation pipeline for GBV: (1) manual data labelling; and (2) automated classification. For (1), we examine two datasets to investigate the relationship between annotator identities and attitudes and the responses they give to two GBV labelling tasks. To this end, we collect demographic and attitudinal information from crowd-sourced annotators using three validated surveys from Social Psychology. We find that higher Right Wing Authoritarianism scores are associated with a higher propensity to label text as sexist, while for Social Dominance Orientation and Neosexist Attitudes, higher scores are associated with a negative tendency to do so. For (2), we conduct classification experiments using Large Language Models and five prompting strategies, including infusing prompts with annotator information. We find: (i) annotator attitudes affect the ability of classifiers to predict their labels; (ii) including attitudinal information can boost performance when we use well-structured brief annotator descriptions; and (iii) models struggle to reflect the increased complexity and imbalanced classes of the new label sets.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Generating and Evaluating Synthetic Data for Privacy Preservation in High-Stakes Domains

Krithika Ramesh, Nupoor Gandhi, Pukit Madan, Lisa Bauer, Charith Peris, Anjalie Field

The difficulty of anonymizing text data hinders the development and deployment of NLP in high-stakes domains that involve private data, such as healthcare and social services. Poorly anonymized sensitive data cannot be easily shared with annotators or external researchers, nor can it be used to train public models. In this work, we explore the feasibility of using synthetic data generated from differentially private language models in place of real data to facilitate the development of NLP in these domains without compromising privacy. In contrast to prior work, we generate synthetic data for real high-stakes domains, and we propose and conduct use-inspired evaluations to assess data quality. Our results show that prior simplistic evaluations have failed to highlight utility, privacy, and fairness issues in the synthetic data. Overall, our

work underscores the need for further improvements to synthetic data generation for it to be a viable way to enable privacy-preserving data sharing.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Evaluating Gender Bias of LLMs in Making Morality Judgements

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, Ruthong Huang

Large Language Models (LLMs) have shown remarkable capabilities in a multitude of Natural Language Processing (NLP) tasks. However, these models are still not immune to limitations such as social biases, especially gender bias. This work investigates whether current closed and open-source LLMs possess gender bias, especially when asked to give moral opinions. To evaluate these models, we curate and introduce a new dataset GenMO (Gender-bias in Morality Opinions) comprising parallel short stories featuring male and female characters respectively. Specifically, we test models from the GPT family (GPT-3.5-turbo, GPT-3.5-turbo-instruct, GPT-4-turbo), Llama 3 and 3.1 families (8B/70B), Mistral-7B and Claude 3 families (Sonnet and Opus). Surprisingly, despite employing safety checks, all production-standard models we tested display significant gender bias with GPT-3.5-turbo giving biased opinions in 24% of the samples. Additionally, all models consistently favour female characters, with GPT showing bias in 68-85% of cases and Llama 3 in around 81-85% instances. Additionally, our study investigates the impact of model parameters on gender bias and explores real-world situations where LLMs reveal biases in moral decision-making.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, Maria Perez-Ortiz

The use of Large Language Models (LLMs) in hiring has led to legislative actions to protect vulnerable demographic groups. This paper presents a novel framework for benchmarking hierarchical gender hiring bias in Large Language Models (LLMs) for resume scoring, revealing significant issues of reverse gender hiring bias and overdebiasing. Our contributions are fourfold: Firstly, we introduce a new construct grounded in labour economics, legal principles, and critiques of current bias benchmarks: hiring bias can be categorized into two types: Level bias (difference in the average outcomes between demographic counterfactual groups) and Spread bias (difference in the variance of outcomes between demographic counterfactual groups). Level bias can be further subdivided into statistical bias (i.e. changing with non-demographic content) and taste-based bias (i.e. consistent regardless of non-demographic content). Secondly, the framework includes rigorous statistical and computational hiring bias metrics, such as Rank After Scoring (RAS), Rank-based Impact Ratio, Permutation Test, and Fixed Effects Model. Thirdly, we analyze gender hiring biases in ten state-of-the-art LLMs. Seven out of ten LLMs show significant biases against males in at least one industry. An industry-effect regression reveals that the healthcare industry is the most biased against males. Moreover, we found that the bias performance remains invariant with resume content for eight out of ten LLMs. This indicates that the bias performance measured in this paper might apply to other resume datasets with different resume qualities. Fourthly, we provide a user-friendly demo and resume dataset to support the adoption and practical use of the framework, which can be generalized to other social traits and tasks.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Multi-Stage Balanced Distillation: Addressing Long-Tail Challenges in Sequence-Level Knowledge Distillation

Yuhang Zhou, Jing Zhu, Paiteng Xu, Xiaoyu Liu, Xiayao Wang, Danai Kourra, Wei Ai, Furong Huang

Large language models (LLMs) have significantly advanced various natural language processing tasks, but deploying them remains computationally expensive. Knowledge distillation (KD) is a promising solution, enabling the transfer of capabilities from larger teacher LLMs to more compact student models. Particularly, sequence-level KD, which distills rationale-based reasoning processes instead of merely final outcomes, shows great potential in enhancing students' reasoning capabilities. However, current methods struggle with sequence-level KD under long-tailed data distributions, adversely affecting generalization on sparsely represented domains. We introduce the Multi-Stage Balanced Distillation (BalDistill) framework, which iteratively balances training data within a fixed computational budget. By dynamically selecting representative head domain examples and synthesizing tail domain examples, BalDistill achieves state-of-the-art performance across diverse long-tailed datasets, enhancing both the efficiency and efficacy of the distilled models.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Granularity is crucial when applying differential privacy to text

Doan Nam Long Vu, Timour Igamberdiev, Ivan Habernal

Applying differential privacy (DP) by means of the DP-SGD algorithm to protect individual data points during training is becoming increasingly popular in NLP. However, the choice of granularity at which DP is applied is often neglected. For example, neural machine translation (NMT) typically operates on the sentence-level granularity. From the perspective of DP, this setup assumes that each sentence belongs to a single person and any two sentences in the training dataset are independent. This assumption is however violated in many real-world NMT datasets, e.g., those including dialogues. For proper application of DP we thus must shift from sentences to entire documents. In this paper, we investigate NMT at both the sentence and document levels, analyzing the privacy/utility trade-off for both scenarios, and evaluating the risks of not using the appropriate privacy granularity in terms of leaking personally identifiable information (PII). Our findings indicate that the document-level NMT system is more resistant to membership inference attacks, emphasizing the significance of using the appropriate granularity when working with DP.

Generation 2

Nov 14 (Thu) 10:30-12:00 - Room: Riverfront Hall

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Bayesian Calibration of Win Rate Estimation with LLM Evaluators

Yichen Gao, Gonghan Xu, Zhe Wang, Arman Cohan

Recent advances in large language models (LLMs) show the potential of using LLMs as evaluators for assessing the quality of text generations from LLMs. However, applying LLM evaluators naively to compare different systems can lead to unreliable results due to the inaccuracy and intrinsic bias of LLM evaluators. In order to mitigate this problem, we propose two calibration methods, Bayesian Win-Rate Sampling (BWRS) and Bayesian David-Skene, both of which leverage Bayesian inference to more accurately infer the true win rate of generative language models. We empirically validate our methods on six datasets covering story generation, summarization, and instruction following tasks. We show that both our methods are effective in improving the accuracy of win rate estimation using LLMs as evaluators, offering a promising direction for reliable automatic text quality evaluation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Safely Learning with Private Data: A Federated Learning Framework for Large Language Model

Jia-Ying Zheng, Hainan Zhang, Lingxiang Wang, Wangjie Qiu, Hong-Wei Zheng, Zhi-Ming Zheng

Private data, being larger and quality-higher than public data, can greatly improve large language models (LLM). However, due to privacy concerns, this data is often dispersed in multiple silos, making its secure utilization for LLM training a challenge. Federated learning (FL) is an ideal solution for training models with distributed private data, but traditional frameworks like FedAvg are unsuitable for LLM due to their high computational demands on clients. An alternative, split learning, offloads most training parameters to the server while training embedding and output layers locally, making it more suitable for LLM. Nonetheless, it faces significant challenges in security and efficiency. Firstly, the gradients of embeddings are prone to attacks, leading to potential reverse engineering of private data. Furthermore, the server's limitation of handling only one client's training request at a time hinders parallel training, severely impacting training efficiency. In this paper, we propose a Federated Learning framework for LLM, named FL-GLM, which prevents data leakage caused by both server-side and peer-client attacks while improving training efficiency. Specifically, we first place the input block and output block on local client to prevent embedding gradient attacks from server. Secondly, we employ key-encryption during client-server communication to prevent reverse engineering attacks from peer-clients. Lastly, we employ optimization methods like client-batching or server-hierarchical, adopting different acceleration methods based on the actual computational capabilities of the server. Experimental results on NLU and generation tasks demonstrate that FL-GLM achieves comparable metrics to centralized chatGLM model, validating the effectiveness of our federated learning framework.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework

Yifan Wang, Vera Demberg

Despite significant advancements in natural language generation, controlling language models to produce texts with desired attributes remains a formidable challenge. In this work, we introduce RSA-Control, a training-free controllable text generation framework grounded in pragmatics. RSA-Control directs the generation process by recursively reasoning between imaginary speakers and listeners, enhancing the likelihood that target attributes are correctly interpreted by listeners amidst distractors. Additionally, we introduce a self-adjustable rationality parameter, which allows for automatic adjustment of control strength based on context. Our experiments, conducted with two task types and two types of language models, demonstrate that RSA-Control achieves strong attribute control while maintaining language fluency and content consistency. Our code is available at <https://github.com/Ewanwong/RSA-Control>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

KnowledgeSG: Privacy-Preserving Synthetic Text Generation With Knowledge Distillation From Server

WenHao Wang, Yuanqi Liang, Rui Ye, Jingyi Choi, Siheng Chen, Yanfeng Wang

The success of large language models (LLMs) facilitate many parties to fine-tune LLMs on their own private data. However, this practice raises privacy concerns due to the memorization of LLMs. Existing solutions, such as utilizing synthetic data for substitution, struggle to simultaneously improve performance and preserve privacy. They either rely on a local model for generation, resulting in a performance decline, or take advantage of APIs, directly exposing the data to API servers. To address this issue, we propose *KnowledgeSG*, a novel client-server framework which enhances synthetic data quality and improves model performance while ensuring privacy. We achieve this by learning local knowledge from the private data with differential privacy (DP) and distilling professional knowledge from the server. Additionally, inspired by federated learning, we transmit models rather than data between the client and server to prevent privacy leakage. Extensive experiments in medical and financial domains demonstrate the effectiveness of **KnowledgeSG**. Our code is now publicly available at <https://github.com/whh0411/KnowledgeSG>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Enhancing Language Model Factuality via Activation-Based Confidence Calibration and Guided Decoding

Xin Liu, Farimaa Fatahi Bayat, Lu Wang

Calibrating language models (LMs) aligns their generation confidence with the actual likelihood of answer correctness, which can inform users about LMs reliability and mitigate hallucinative content. However, prior calibration methods, such as self-consistency-based and logit-based approaches, are either limited in inference-time efficiency or fall short of providing informative signals. Moreover, simply filtering out low-confidence responses reduces the LMs helpfulness when the answers are correct. Therefore, effectively using calibration techniques to enhance an LMs factuality remains an unsolved challenge. In this paper, we first propose an activation-based calibration method, ActCab, which trains a linear layer on top of the LMs last-layer activations that can better capture the representations of knowledge. Built on top of ActCab, we further propose CoDec, a confidence-guided decoding strategy to elicit truthful answers with high confidence from LMs. By evaluating on five popular QA benchmarks, ActCab achieves superior calibration performance than all competitive baselines, e.g., by reducing the average expected calibration error (ECE) score by up to 39%. Further experiments on CoDec show consistent improvements in several LMs factuality on challenging QA datasets, such as TruthfulQA, highlighting the value of confidence signals in enhancing the factuality.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

DocCodeGen: Document-based Controlled Code Generation

Sameer Pimparkhede, Mehanit Kannamkotam, Srikanth G. Tamiselvam, Prince Kumar, Ashok Pon Kumar, Pushpak Bhattacharyya

Recent developments show that Large Language Models (LLMs) produce state-of-the-art performance on natural language (NL) to code generation for resource-rich general-purpose languages like C++, Java, and Python. However, their practical usage for structured domain-specific languages (DSLs) such as YAML, JSON is limited due to domain-specific schema, grammar, and customizations generally unseen by LLMs during pre-training. Efforts have been made to mitigate this challenge via in-context learning through relevant examples or by fine-tuning. However, it suffers from problems, such as limited DSL samples and prompt sensitivity but enterprises maintain good documentation of the DSLs. Therefore, we propose DocCodeGen, a framework that can leverage such rich knowledge by breaking the NL-to-Code generation task for structured code languages into a two-step process. First, it detects the correct libraries using the library documentation that best matches the NL query. Then, it utilizes schema rules extracted from the documentation of these libraries to constrain the decoding. We evaluate our framework for two complex structured languages, Ansible YAML and Bash command, consisting of two settings: Out-of-domain (OOD) and In-domain (ID). Our extensive experiments show that DocCodeGen consistently improves different sized language models across all six evaluation metrics, reducing syntactic and semantic errors in structured code.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Collective Critics for Creative Story Generation

Minwook Bae, Hyounghun Kim

Generating a long story of several thousand words with narrative coherence using Large Language Models (LLMs) has been a challenging task. Previous research has addressed this challenge by proposing different frameworks that create a story plan and generate a long story based on that plan. However, these frameworks have been mainly focusing on maintaining narrative coherence in stories, often overlooking creativity in story planning and the expressiveness of the stories generated from those plans, which are desirable properties to captivate readers' interest. In this paper, we propose Collective Critics for Creative Story Generation framework (CritiCS), which is composed of plan refining stage (CrPlan) and story generation stage (CrText), to integrate a collective revision mechanism that promotes those properties into long-form story generation process. Specifically, in each stage, a group of LLM critics and one leader collaborate to incrementally refine drafts of plan and story throughout multiple rounds. Extensive human evaluation shows that the CritiCS can significantly enhance story creativity and reader engagement, while also maintaining narrative coherence. Furthermore, the design of the framework allows active participation from human

writers in any role within the critique process, enabling interactive human-machine collaboration in story writing.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Generation with Dynamic Vocabulary

Yanting Liu, Tao Ji, Yuanbin Wu, Xiaoling Wang, Changzhi Sun

We introduce a new dynamic vocabulary for language models. It can involve arbitrary text spans during generation. These text spans act as basic generation bricks, akin to tokens in the traditional static vocabularies. We show that, the ability to generate multi-tokens atomically improve both generation quality and efficiency (compared to the standard language model, the MAUVE metric is increased by 25%, the latency is decreased by 20%). The dynamic vocabulary can be deployed in a plug-and-play way, thus is attractive for various downstream applications. For example, we demonstrate that dynamic vocabulary can be applied to different domains in a training-free manner. It also helps to generate reliable citations in question answering tasks (substantially enhancing citation results without compromising answer accuracy).

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

SynthesizRR: Generating Diverse Datasets with Retrieval Augmentation

Abhisek Divekar, Greg Durrett

It is often desirable to distill the capabilities of large language models (LLMs) into smaller student models due to compute and memory constraints. One way to do this for classification tasks is via dataset synthesis, which can be accomplished by generating examples of each label from the LLM. Prior approaches to synthesis use few-shot prompting, which relies on the LLM's parametric knowledge to generate usable examples. However, this leads to issues of repetition, bias towards popular entities, and stylistic differences from human text. In this work, we propose Synthesize by Retrieval and Refinement (SynthesizRR), which uses retrieval augmentation to introduce variety into the dataset synthesis process: as retrieved passages vary, the LLM is seeded with different content to generate its examples. We empirically study the synthesis of six datasets, covering topic classification, sentiment analysis, tone detection, and humor, requiring complex synthesis strategies. We find SynthesizRR greatly improves lexical and semantic diversity, similarity to human-written text, and distillation performance, when compared to 32-shot prompting and four prior approaches.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

GPT vs RETRO: Exploring the Intersection of Retrieval and Parameter-Efficient Fine-Tuning

Aleksander Ficek, Jiaqi Zeng, Oleksii Kuchaiiev

Parameter-Efficient Fine-Tuning (PEFT) and Retrieval-Augmented Generation (RAG) have become popular methods for adapting large language models while minimizing compute requirements. In this paper, we apply PEFT methods (P-tuning, Adapters, and LoRA) to a modified Retrieval-Enhanced Transformer (RETRO) and a baseline GPT model across several sizes, ranging from \$23 million to 48 billion parameters. We show that RETRO models outperform GPT models in zero-shot settings due to their unique pre-training process but GPT models have higher performance potential with PEFT. Additionally, our study indicates that 8B parameter models strike an optimal balance between cost and performance and P-tuning lags behind other PEFT techniques. We further provide a comparative analysis of between applying PEFT to Instruction-tuned RETRO model and base RETRO model. This work presents the first comprehensive comparison of various PEFT methods integrated with RAG, applied to both GPT and RETRO models, highlighting their relative performance.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Reasoning in Token Economies: Budget-Aware Evaluation of LLM Reasoning Strategies

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, Ben Athiwaratkun

A diverse array of reasoning strategies has been proposed to elicit the capabilities of large language models. However, in this paper, we point out that traditional evaluations which focus solely on performance metrics miss a key factor: the increased effectiveness due to additional compute. By overlooking this aspect, a skewed view of strategy efficiency is often presented. This paper introduces a framework that incorporates the compute budget into the evaluation, providing a more informative comparison that takes into account both performance metrics and computational cost. In this budget-aware perspective, we find that complex reasoning strategies often don't surpass simpler baselines purely due to algorithmic ingenuity, but rather due to the larger computational resources allocated. When we provide a simple baseline like chain-of-thought self-consistency with comparable compute resources, it frequently outperforms reasoning strategies proposed in the literature. In this scale-aware perspective, we find that unlike self-consistency, certain strategies such as multi-agent debate or Reflexion can become worse if more compute budget is utilized.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Language Models as Compilers: Simulating Pseudocode Execution Improves Algorithmic Reasoning in Language Models

Hyungjoo Chae, Yeonghyeon Kim, Seungone Kim, Kai Tzu-iunn Ong, Beong-woo Kwak, Moohyeon Kim, Sunghwan Kim, Taeyoon Kwon, Jiwon Chung, Youngjae Yu, Jinyoung Yeo

Algorithmic reasoning tasks that involve complex logical patterns, such as completing Dyck language, pose challenges for large language models (LLMs), despite their recent success. Prior work has used LLMs to generate programming language and applied external compilers for such tasks. Yet, when on the fly, it is hard to generate an executable code with the correct logic for the solution. Even so, code for one instance cannot be reused for others, although they might require the same logic to solve. We present Think-and-Execute, a novel framework that improves LLMs' algorithmic reasoning: (1) In Think, we discover task-level logic shared across all instances, and express such logic with pseudocode; (2) In Execute, we tailor the task-level pseudocode to each instance and simulate the execution of it. Think-and-Execute outperforms several strong baselines (including CoT and PoT) in diverse algorithmic reasoning tasks. We manifest the advantage of using task-level pseudocode over generating instance-specific solutions one by one. Also, we show that pseudocode can better improve LMs' reasoning than natural language (NL) guidance, even though they are trained with NL instructions.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Symbolic Prompt Program Search: A Structure-Aware Approach to Efficient Compile-Time Prompt Optimization

Tobias Schnabel, Jennifer Neville

In many modern LLM applications, such as retrieval augmented generation, prompts have become programs themselves. In these settings, prompt programs are repeatedly called with different user queries or data instances. A big practical challenge is optimizing such prompt programs. Recent work has mostly focused on either simple prompt programs or assumed that the structure of a prompt program is fixed. We introduce SAMMO, a framework to perform symbolic prompt program search for compile-time optimizations of prompt programs. SAMMO represents prompt programs on a symbolic level which allows for a rich set of transformations that can be searched over during optimization. We show that SAMMO generalizes previous methods and improves the performance of complex prompts on (1) instruction tuning, (2) RAG pipeline tuning, and (3) prompt compression, across several different LLMs. We make all code available open-source at <https://anony-mous.4open.science/r/sammo-4003/>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Few-shot Selections for Numerical Time Series Data-to-Text

Masayuki Kawarada, Tatsuya Ishigaki, Goran Topi, Hiroya Takamura

Demonstration selection, the process of selecting examples used in prompts, plays a critical role in in-context learning. This paper explores demonstration selection methods for data-to-text tasks that involve numerical time series data as inputs. Previously developed demonstration selection methods primarily focus on textual inputs, often relying on embedding similarities of textual tokens to select similar instances from an example bank. However, this approach may not be suitable for numerical time series data. To address this issue, we propose two novel selection methods: (1) sequence similarity-based selection using various similarity measures, and (2) task-specific knowledge-based selection. From our experiments on two benchmark datasets, we found that our proposed models significantly outperform baseline selections and often surpass fine-tuned models. We also found that scale-invariant similarity measures such as Pearson's correlation work better than scale-variant measures such as Euclidean distance. Manual evaluation by human judges also confirms that our proposed methods outperform conventional methods.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

CHIRON: Rich Character Representations in Long-Form Narratives

Alexander Gurung, Mirella Lapata

Characters are integral to long-form narratives, but are poorly understood by existing story analysis and generation systems. While prior work has simplified characters via graph-based methods and brief character descriptions, we aim to better tackle the problem of representing complex characters by taking inspiration from advice given to professional writers. We propose CHIRON, a new ‘character sheet’ based representation that organizes and filters textual information about characters. We construct CHIRON sheets in two steps: a Generation Module that prompts an LLM for character information via question-answering and a Validation Module that uses automated reasoning and a domain-specific entailment model to eliminate false facts about a character. We validate CHIRON via the downstream task of masked-character prediction, where our experiments show CHIRON is better and more flexible than comparable summary-based baselines. We also show that metrics derived from CHIRON can be used to automatically infer character-centricity in stories, and that these metrics align with human judgments.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA

Nirmal Roy, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Kevin Small

Augmenting Large Language Models (LLMs) with information retrieval capabilities (i.e., Retrieval-Augmented Generation (RAG)) has proven beneficial for knowledge-intensive tasks. However, understanding users' contextual search intent when generating responses is an understudied topic for conversational question answering (QA). This conversational extension leads to additional concerns when compared to single-turn QA as it is more challenging for systems to comprehend conversational context and manage retrieved passages over multiple turns. In this work, we propose a method for enabling LLMs to decide when to retrieve in RAG settings given a conversational context. When retrieval is deemed necessary, the LLM then rewrites the conversation for passage retrieval and judges the relevance of returned passages before response generation. Operationally, we build on the single-turn SELF-RAG framework (Asai et al., 2023) and propose SELF-multi-RAG for conversational settings. SELF-multi-RAG demonstrates improved capabilities over single-turn variants with respect to retrieving relevant passages (by using summarized conversational context) and assessing the quality of generated responses. Experiments on three conversational QA datasets validate the enhanced response generation capabilities of SELF-multi-RAG with improvements of 13% measured by human annotation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Retrieval-Augmented Code Generation for Situated Action Generation: A Case Study on Minecraft

Kranti CH, Sherzod Hakimov, David Schlangen

In the Minecraft Collaborative Building Task, two players collaborate: an Architect (A) provides instructions to a Builder (B) to assemble a specified structure using 3D blocks. In this work, we investigate the use of large language models (LLMs) to predict the sequence of actions taken by the Builder. Leveraging LLMs in-context learning abilities, we use few-shot prompting techniques, that significantly improve performance over baseline methods. Additionally, we present a detailed analysis of the gaps in performance for future work.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Automating Easy Read Text Segmentation

Jesús Javier Calleja Pérez, Thierry Etichayeghen, Antonio David Ponce Martínez

Easy Read text is one of the main forms of access to information for people with reading difficulties. One of the key characteristics of this type of text is the requirement to split sentences into smaller grammatical segments, to facilitate reading. Automated segmentation methods could foster the creation of Easy Read content, but their viability has yet to be addressed. In this work, we study novel methods for the task, leveraging masked and generative language models, along with constituent parsing. We conduct comprehensive automatic and human evaluations in three languages, analysing the strengths and weaknesses of the proposed alternatives, under scarce resource limitations. Our results highlight the viability of automated Easy Read segmentation and remaining deficiencies compared to expert-driven human segmentation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Enhancing Alignment using Curriculum Learning & Ranked Preferences

Pulkit Patnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, Sathwik Tejaswi Madhusudhan

Direct Preference Optimization (DPO) is an effective technique that leverages pairwise preference data (one chosen and rejected response per prompt) to align LLMs to human preferences. In practice, multiple responses could exist for a given prompt with varying quality relative to each other. We propose to utilize these responses to create multiple preference pairs for a given prompt. Our work focuses on aligning LLMs by systematically curating multiple preference pairs and presenting them in a meaningful manner facilitating curriculum learning to enhance the prominent DPO technique. We order multiple preference pairs from easy to hard, according to various criteria thus emulating curriculum learning. Our method, which is referred to as Curri-DPO consistently shows increased performance gains on MTbench, Vicuna bench, WizardLM, highlighting its effectiveness over standard DPO setting that utilizes single preference pair. More specifically, Curri-DPO achieves a score of 7.43 on MTbench with Zephyr-7B, outperforming majority of existing LLMs with similar parameter size. Curri-DPO also achieves the highest win rates on Vicuna, WizardLM, and UltraFeedback test sets (90.7%, 87.1%, and 87.9% respectively) in our experiments, with notable gains of up to 7.5% when compared to standard DPO. We release the preference pairs used in alignment at: https://hugging-face.co/datasets/ServiceNow-AI/Curriculum_DPO_preferences.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

TinyStyler: Efficient Few-Shot Text Style Transfer with Authorship Embeddings

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, Zhou Yu

The goal of text style transfer is to transform the style of texts while preserving their original meaning, often with only a few examples of the target style. Existing style transfer methods generally rely on the few-shot capabilities of large language models or on complex controllable text generation approaches that are inefficient and underperform on fluency metrics. We introduce TinyStyler, a lightweight but effective approach, which leverages a small language model (800M params) and pre-trained authorship embeddings to perform efficient, few-shot text style transfer. We evaluate on the challenging task of authorship style transfer and find TinyStyler outperforms strong approaches such as

GPT-4. We also evaluate TinyStyler's ability to perform text attribute style transfer (formal \leftrightarrow informal) with automatic and human evaluations and find that the approach outperforms recent controllable text generation methods.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Calibrating Long-form Generations From Large Language Models

Yukun Huang, Yixuan Liu, Raghavveer Thirukovalluru, Arman Cohan, Bhuvan Dhingra

To enhance Large Language Models' (LLMs) reliability, calibration is essential—the model's confidence scores should align with the likelihood of its responses being correct. However, traditional calibration methods typically rely on a binary true/false assessment of response correctness, unsuitable for long-form generations where an answer can be partially correct. Addressing this gap, we introduce a unified calibration framework, in which both the correctness of the LLMs' responses and their associated confidence levels are treated as distributions across a range of scores. We develop three metrics for assessing LLM calibration and propose confidence elicitation methods based on self-consistency and self-evaluation. Our experiments demonstrate that larger models don't necessarily guarantee better calibration, that various calibration metrics complement each other, and that self-consistency methods excel in factoid datasets. We also find that calibration can be enhanced through techniques such as fine-tuning, scaling the temperature. Finally, we illustrate one application of long-form calibration through selective answering in long-form responses, optimizing correctness within a constrained API budget.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback

Kyoungh Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, Kinmin Lee

Large language models (LLMs) fine-tuned with alignment techniques, such as reinforcement learning from human feedback, have been instrumental in developing some of the most capable AI systems to date. Despite their success, existing methods typically rely on simple binary labels, such as those indicating preferred outputs in pairwise preferences, which fail to capture the subtle differences in relative quality between pairs. To address this limitation, we introduce an approach called Margin Matching Preference Optimization (MMPO), which incorporates relative quality margins into optimization, leading to improved LLM policies and reward models. Specifically, given quality margins in pairwise preferences, we design soft target probabilities based on the Bradley-Terry model, which are then used to train models with the standard cross-entropy objective. Experiments with both human and AI feedback data demonstrate that MMPO consistently outperforms baseline methods, often by a substantial margin, on popular benchmarks including MT-bench and RewardBench. Notably, the 7B model trained with MMPO achieves state-of-the-art performance on RewardBench as of June 2024, outperforming other models of the same scale. Our analysis also shows that MMPO is more robust to overfitting, leading to better-calibrated models.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Aligners: Decoupling LLMs and Alignment

Lilian Ngweta, Mayank Agarwal, Subha Maiti, Alex Gittens, Yuekai Sun, Mikhail Yurochkin

Large Language Models (LLMs) need to be aligned with human expectations to ensure their safety and utility in most applications. Alignment is challenging, costly, and needs to be repeated for every LLM and alignment criterion. We propose to decouple LLMs and alignment by training *aligner* models that can be used to align any LLM for a given criteria on an as-needed basis, thus also reducing the potential negative impacts of alignment on performance. Our recipe for training the aligner models solely relies on synthetic data generated with a (prompted) LLM and can be easily adjusted for a variety of alignment criteria. We use the same synthetic data to train *inspectors*, binary miss-alignment classification models to guide a *squad* of multiple aligners. Our empirical results demonstrate consistent improvements when applying aligner squad to various LLMs, including chat-aligned models, across several instruction-following and red-teaming datasets.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models

Shayekh Bin Islam, Md Asib Rahman, K SM Tozammel Hossain, Enamul Hoque, Shafiq Joty, Md Rizwan Parvez

Retrieval Augmented Generation (RAG) has been shown to enhance the factual accuracy of Large Language Models (LLMs) by providing external evidence, but existing methods often suffer from limited reasoning capabilities (e.g., multi-hop complexities) in effectively using such evidence, particularly when using open-source LLMs. To mitigate this gap, in this paper, we introduce a novel framework, **Open-RAG**+, designed to enhance reasoning capabilities in RAG with open-source LLMs. Our framework transforms an arbitrary dense LLM into a parameter-efficient sparse mixture of experts (MoE) model capable of handling complex reasoning tasks, including both single- and multi-hop queries. Open-RAG uniquely trains the model to navigate challenging distractors that appear relevant but are misleading. By combining the constructive learning and architectural transformation, Open-RAG leverages latent learning, dynamically selecting relevant experts and integrating external knowledge effectively for more accurate and contextually relevant responses. Additionally, we propose a hybrid adaptive retrieval method to determine retrieval necessity and balance the trade-off between performance gain and inference speed. Experimental results show that Open-RAG outperforms state-of-the-art LLMs and RAG models in various knowledge-intensive tasks. Our method based on Llama2-7B sets new benchmarks, surpassing ChatGPT-RAG and Self-RAG. For example, in multi-hop HotpotQA, it achieves an EM score of 63.3, compared to RAG 2.0's 54 and Command R+6's 60.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

How You Prompt Matters! Even Task-Oriented Constraints in Instructions Affect LLM-Generated Text Detection

Ryuuto Koike, Masahiro Kaneko, Naoki Okazaki

To combat the misuse of Large Language Models (LLMs), many recent studies have presented LLM-generated-text detectors with promising performance. When users instruct LLMs to generate texts, the instruction can include different constraints depending on the user's need. However, most recent studies do not cover such diverse instruction patterns when creating datasets for LLM detection. In this paper, we reveal that even task-oriented constraints — constraints that would naturally be included in an instruction and are not related to detection-evasion — cause existing powerful detectors to have a large variance in detection performance. We focus on student essay writing as a realistic domain and manually create task-oriented constraints based on several factors for essay quality. Our experiments show that the standard deviation (SD) of current detector performance on texts generated by an instruction with such a constraint is significantly larger (up to an SD of 14.4 F1-score) than that by generating texts multiple times or paraphrasing the instruction. We also observe an overall trend where the constraints can make LLM detection more challenging than without them. Finally, our analysis indicates that the high instruction-following ability of LLMs fosters the large impact of such constraints on detection performance.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

I'm sure you're a real scholar yourself: Exploring Ironic Content Generation by Large Language Models

Pier Felice Balestrucci, Silvia Casola, Soda Marem Lo, Valerio Basile, Alessandro Mazzei

Generating ironic content is challenging: it requires a nuanced understanding of context and implicit references and balancing seriousness and playfulness. Moreover, irony is highly subjective and can depend on various factors, such as social, cultural, or generational aspects. This paper explores whether Large Language Models (LLMs) can learn to generate ironic responses to social media posts. To do so, we fine-tune two models to generate ironic and non-ironic content and deeply analyze their outputs' linguistic characteristics, their connection to the original post, and their similarity to the human-written replies. We also conduct a large-scale human evaluation of the outputs. Additionally, we

investigate whether LLMs can learn a form of irony tied to a generational perspective, with mixed results.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Local and Global Decoding in Text Generation

Daniel Gareev, Thomas Hofmann, ezhilmathi krishnasamy, Tiago Pimentel

Text generation, a component in applications such as dialogue systems, relies heavily on decoding algorithms that sample strings from a language model distribution. Traditional methods like top- k and top- π decoding locally normalise the model's output, which can significantly distort the original distribution. In this paper, we investigate the effects of such distortions by introducing globally-normalised versions of these decoding methods. Further, we propose an independent Metropolis-Hastings (IMH) algorithm to approximate sampling from these globally-normalised distributions without explicitly computing them. Our empirical analyses compare the performance of local and global decoding across two algorithms (top- k and top- π) with various hyperparameters, using the Pythia language models. Results show that in most configuration, global decoding performs worse than the local decoding versions of the same algorithms, despite preserving the distribution's integrity. Our results thus suggest that distortion might be an important feature of local decoding algorithms.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

SARCAT: Generative Span-Act Guided Response Generation using Copy-enhanced Target Augmentation

Jeong-Doo Lee, Hyeyoung Choi, Beomseok Hong, Younsub Han, Byoung-Ki Jeon, Seung-Hoon Na

In this paper, we present a novel extension to improve the document-grounded response generation, by proposing the Generative Span Act Guided Response Generation using Copy enhanced Target Augmentation (SARCAT) that consists of two major components as follows: 1) Copy-enhanced target-side input augmentation is an extended data augmentation to deal with the exposure bias problem by additionally incorporating the copy mechanism on top of the target-side augmentation (Xie et al., 2021). 2) Span-act guided response generation, which first predicts grounding spans and dialogue acts before generating a response. Experimental results on validation set in MultiDoc2Dial show that the proposed SARSAT leads to improvement over strong baselines on both seen and unseen settings and achieves the start-of-the-art performance, even with the base reader using the pretrained T5-base model.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Adaptive Contrastive Search: Uncertainty-Guided Decoding for Open-Ended Text Generation

Esteban Garces Arias, Julian Rodemann, Meimengwei Li, Christian Heumann, Matthias Assemmacher

Despite the remarkable capabilities of large language models, generating high-quality text remains a challenging task. Numerous decoding strategies such as beam search, sampling with temperature, top- k sampling, nucleus (topp) sampling, typical decoding, contrastive decoding, and contrastive search have been proposed to address these challenges by improving coherence, diversity, and resemblance to human-generated text. In this study, we introduce Adaptive Contrastive Search (ACS), a novel decoding strategy that extends contrastive search (CS) by incorporating an adaptive degeneration penalty informed by the model's estimated uncertainty at each generation step. ACS aims to enhance creativity and diversity while maintaining coherence to produce high-quality outputs. Extensive experiments across various model architectures, languages, and datasets demonstrate that our approach improves both creativity and coherence, underscoring its effectiveness in text-generation tasks. We release our code, datasets, and models to facilitate further research.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Controlled Transformation of Text-Attributed Graphs

Nidhi Vakil, Hadi Amiri

Graph generation is the process of generating novel graphs with similar attributes to real world graphs. The explicit and precise control of granular structural attributes, such as node centrality and graph density, is crucial for effective graph generation. This paper introduces a controllable multi-objective translation model for text-attributed graphs, titled Controlled Graph Translator (CGT). It is designed effectively and efficiently translate a given source graph to a target graph, while satisfying multiple desired graph attributes at granular level. Designed with an encoder-decoder architecture, CGT develops fusion and graph attribute predictor neural networks for controlled graph translation. We validate the effectiveness of CGT through extensive experiments on different genres of datasets. In addition, we illustrate the application of CGT in data augmentation and taxonomy creation, particularly in low resource settings.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Presentations are not always linear! GNN meets LLM for Document-to-Presentation Transformation with Attribution

Himanshu Maheshwari, Sambaran Bandopadhyay, Aparna Garimella, Anandhavelu Naratajan

Automatically generating a presentation from the text of a long document is a challenging and useful problem. In contrast to a flat summary, a presentation needs to have a better and non-linear narrative, i.e., the content of a slide can come from different and non-contiguous parts of the given document. However, it is difficult to incorporate such non-linear mapping of content to slides and ensure that the content is faithful to the document. LLMs are prone to hallucination and their performance degrades with the length of the input document. Towards this, we propose a novel graph-based solution where we learn a graph from the input document and use a combination of graph neural network and LLM to generate a presentation with attribution of content for each slide. We conduct thorough experiments to show the merit of our approach compared to directly using LLMs for this task.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Solving for X and Beyond: Can Large Language Models Solve Complex Math Problems with More-Than-Two Unknowns?

Kuei-Chun Kao, Ruochen Wang, Cho-Jui Hsieh

Large Language Models have demonstrated remarkable performance in solving math problems, a hallmark of human intelligence. Despite high success rates on current benchmarks, however, these often feature simple problems with only one or two unknowns, which do not sufficiently challenge their reasoning capacities. This paper introduces a novel benchmark, BeyondX, designed to address these limitations by incorporating problems with multiple unknowns. Recognizing the challenges in proposing multi-unknown problems from scratch, we developed BeyondX using an innovative automated pipeline that progressively increases complexity by expanding the number of unknowns in simpler problems. Empirical study on BeyondX reveals that the performance of existing LLMs, even those fine-tuned specifically on math tasks, significantly decreases as the number of unknowns increases - with a performance drop of up to 70% observed in GPT-4. To tackle these challenges, we propose the Formulate-and-Solve strategy, a generalized prompting approach that effectively handles problems with an arbitrary number of unknowns. Our findings reveal that this strategy not only enhances LLM performance on the BeyondX benchmark but also provides deeper insights into the computational limits of LLMs when faced with more complex mathematical challenges.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Suri: Multi-constraint Instruction Following in Long-form Text Generation

Chau Minh Pham, Simeng Sun, Mohit Iyer

Existing research on instruction following largely focuses on tasks with simple instructions and short responses. In this work, we explore multi-constraint instruction following for generating long-form text. We create Suri, a dataset with 20K human-written long-form texts paired with LLM-generated backtranslated instructions that contain multiple complex constraints. Because of prohibitive challenges associated with

collecting human preference judgments on long-form texts, preference-tuning algorithms such as DPO are infeasible in our setting; thus, we propose Instructional ORPO (I-ORPO), an alignment method based on the ORPO algorithm. Instead of receiving negative feedback from dispreferred responses, I-ORPO obtains negative feedback from synthetically corrupted instructions generated by an LLM. Using Suri, we perform supervised and I-ORPO fine-tuning on Mistral-7b-Instruct-v0.2. The resulting models, Suri-SFT and Suri-I-ORPO, generate significantly longer texts (5K tokens) than base models without significant quality deterioration. Our human evaluation shows that while both SFT and I-ORPO models satisfy most constraints, Suri-I-ORPO generations are generally preferred for their coherent and informative incorporation of the constraints.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Breaking the Ceiling of the LLM Community by Treating Token Generation as a Classification for Ensembling

Yao-Ching Yu, Chun Chih Kuo, Yen Yuchen Chang, Yueh-Se Li

Ensembling multiple models has always been an effective approach to push the limits of existing performance and is widely used in classification tasks by simply averaging the classification probability vectors from multiple classifiers to achieve better accuracy. However, in the thriving open-source Large Language Model (LLM) community, ensembling methods are rare and typically limited to ensembling the full-text outputs of LLMs, such as selecting the best output using a ranker, which leads to underutilization of token-level probability information. In this paper, we treat the **C**-generation of each token by LLMs as a ***Ga*** classification (**Ga**) for ensembling. This approach fully exploits the probability information at each generation step and better prevents LLMs from producing early incorrect tokens that lead to snowballing errors. In experiments, we ensemble state-of-the-art LLMs on several benchmarks, including exams, mathematics and reasoning, and observe that our method breaks the existing community performance ceiling. Furthermore, we observed that most of the tokens in the answer are simple and do not affect the correctness of the final answer. Therefore, we also experimented with ensembling only key tokens, and the results showed better performance with lower latency across benchmarks.

Interpretability and Analysis of Models for NLP 5

Nov 14 (Thu) 10:30-12:00 - Room: Jasmine

Nov 14 (Thu) 10:30-12:00 - Jasmine

Defending Large Language Models Against Jailbreak Attacks via Layer-specific Editing

Wei Zhao, Zhi Li, Yige Li, YE ZHANG, Jun Sun

Large language models (LLMs) are increasingly being adopted in a wide range of real-world applications. Despite their impressive performance, recent studies have shown that LLMs are vulnerable to deliberately crafted adversarial prompts even when aligned via Reinforcement Learning from Human Feedback or supervised fine-tuning. While existing defense methods focus on either detecting harmful prompts or reducing the likelihood of harmful responses through various means, defending LLMs against jailbreak attacks based on the inner mechanisms of LLMs remains largely unexplored. In this work, we investigate how LLMs respond to harmful prompts and propose a novel defense method termed **Layer-specific Editing** (LED) to enhance the resilience of LLMs against jailbreak attacks. Through LED, we reveal that several critical *safety layers* exist among the early layers of LLMs. We then show that realigning these safety layers (and some selected additional layers) with the decoded safe response from identified *toxic layers* can significantly improve the alignment of LLMs against jailbreak attacks. Extensive experiments across various LLMs (e.g., Llama2, Mistral) show the effectiveness of LED, which effectively defends against jailbreak attacks while maintaining performance on benign prompts. Our code is available at <https://github.com/ledlm/ledllm>.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Understanding Higher-Order Correlations Among Semantic Components in Embeddings

Monose Oyama, Hiroaki Yamagishi, Hidetoshi Shimodaira

Independent Component Analysis (ICA) offers interpretable semantic components of embeddings. While ICA theory assumes that embeddings can be linearly decomposed into independent components, real-world data often do not satisfy this assumption. Consequently, non-independencies remain between the estimated components, which ICA cannot eliminate. We quantified these non-independencies using higher-order correlations and demonstrated that when the higher-order correlation between two components is large, it indicates a strong semantic association between them, along with many words sharing common meanings with both components. The entire structure of non-independencies was visualized using a maximum spanning tree of semantic components. These findings provide deeper insights into embeddings through ICA.

Nov 14 (Thu) 10:30-12:00 - Jasmine

LUQ: Long-text Uncertainty Quantification for LLMs

Caiqi Zhang, Fangyu Liu, Marco Basilella, Nigel Collier

Large Language Models (LLMs) have demonstrated remarkable capability in a variety of NLP tasks. However, LLMs are also prone to generate nonfactual content. Uncertainty Quantification (UQ) is pivotal in enhancing our understanding of a model's confidence on its generation, thereby aiding in the mitigation of nonfactual outputs. Existing research on UQ predominantly targets short text generation, typically yielding brief, word-limited responses. However, real-world applications frequently necessitate much longer responses. Our study first highlights the limitations of current UQ methods in handling long text generation. We then introduce LUQ and its two variations, a series of novel sampling-based UQ approaches specifically designed for long text. Our findings reveal that LUQ outperforms existing baseline methods in correlating with the model's factuality scores (negative coefficient of -0.85 observed for Gemini Pro). To further improve the factuality of LLM responses, we propose LUQ-ENSEMBLE, a method that ensembles responses from multiple models and selects the response with the lowest uncertainty. The ensembling method greatly improves the response factuality upon the best standalone LLM.

Nov 14 (Thu) 10:30-12:00 - Jasmine

XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs

Zichen Chen, Jianda Chen, Ambuj Singh, Misha Sra

Large Language Models (LLMs) have achieved remarkable success in natural language tasks, yet understanding their reasoning processes remains a significant challenge. We address this by introducing XplainLLM, a dataset accompanying an explanation framework designed to enhance LLM transparency and reliability. Our dataset comprises 24,204 instances where each instance interprets the LLM's reasoning behavior using knowledge graphs (KGs) and graph attention networks (GAT), and includes explanations of LLMs such as the decoder-only Llama-3 and the encoder-only RoBERTa. XplainLLM also features a framework for generating grounded explanations and the *debugger-scores* for multidimensional quality analysis. Our explanations include *why-choose* and *why-not-choose* components, *reason-elements*, and *debugger-scores* that collectively illuminate the LLM's reasoning behavior. Our evaluations demonstrate XplainLLM's potential to reduce hallucinations and improve grounded explanation generation in LLMs. XplainLLM is a resource for researchers and practitioners to build

trust and verify the reliability of LLM outputs. Our code and dataset are publicly available⁷.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Atomic Inference for NLI with Generated Facts as Atoms

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, Marek Rei

With recent advances, neural models can achieve human-level performance on various natural language tasks. However, there are no guarantees that any explanations from these models are faithful, i.e. that they reflect the inner workings of the model. Atomic inference overcomes this issue, providing interpretable and faithful model decisions. This approach involves making predictions for different components (or atoms) of an instance, before using interpretable and deterministic rules to derive the overall prediction based on the individual atom-level predictions. We investigate the effectiveness of using LLM-generated facts as atoms, decomposing Natural Language Inference premises into lists of facts. While directly using generated facts in atomic inference systems can result in worse performance, with 1) a multi-stage fact generation process, and 2) a training regime that incorporates the facts, our fact-based method outperforms other approaches.

Nov 14 (Thu) 10:30-12:00 - Jasmine

I Learn Better If You Speak My Language: Understanding the Superior Performance of Fine-Tuning Large Language Models with LLM-Generated Responses

Xuan Ren, Biao Wu, Lingqiao Liu

This paper explores an intriguing observation: fine-tuning a large language model (LLM) with responses generated by a LLM often yields better results than using responses generated by humans, particularly in reasoning tasks. We conduct an in-depth investigation to understand why this occurs. Contrary to the common belief that these instances is due to the more detailed nature of LLM-generated content, our study identifies another contributing factor: an LLM is inherently more "familiar" with LLM generated responses. This familiarity is evidenced by lower perplexity before fine-tuning. We design a series of experiments to understand the impact of the "familiarity" and our conclusion reveals that this "familiarity" significantly impacts learning performance. Training with LLM-generated responses not only enhances performance but also helps maintain the model's capabilities in other reasoning tasks after fine-tuning on a specific task.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Kiss up, Kick down: Exploring Behavioral Changes in Multi-modal Large Language Models with Assigned Visual Personas

Seungjun Sun, Eungu Lee, Seo Yeon Baek, Seunghyun Hwang, Lee wonbyung, Dongyan Nan, Bernard J Jansen, Jang Hyun Kim

This study is the first to explore whether multi-modal large language models (LLMs) can align their behaviors with visual personas, addressing a significant gap in the literature that predominantly focuses on text-based personas. We developed a novel dataset of 5K fictional avatar images for assignment as visual personas to LLMs, and analyzed their negotiation behaviors based on the visual traits depicted in these images, with a particular focus on aggressiveness. The results indicate that LLMs assess the aggressiveness of images in a manner similar to humans and output more aggressive negotiation behaviors when prompted with an aggressive visual persona. Interestingly, the LLM exhibited more aggressive negotiation behaviors when the opponents image appeared less aggressive than their own, and less aggressive behaviors when the opponents image appeared more aggressive.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Reasoning or a Semblance of it? A Diagnostic Study of Transitive Reasoning in LLMs

Houman Mehravarin, Arash Esghli, Ioannis Konstas

Evaluating Large Language Models (LLMs) on reasoning benchmarks demonstrates their ability to solve compositional questions. However, little is known of whether these models engage in genuine logical reasoning or simply rely on implicit cues to generate answers. In this paper, we investigate the transitive reasoning capabilities of two distinct LLM architectures, LLaMA 2 and Flan-T5, by manipulating facts within two compositional datasets: QASC and Bamboogle. We controlled for potential cues that might influence the models' performance, including (a) word/phrase overlaps across sections of test input; (b) models' inherent knowledge during pre-training or fine-tuning; and (c) Named Entities. Our findings reveal that while both models leverage (a), Flan-T5 shows more resilience to experiments (b and c), having less variance than LLaMA 2. This suggests that models may develop an understanding of transitivity through fine-tuning on knowingly relevant datasets, a hypothesis we leave to future work.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Rationalizing Transformer Predictions via End-To-End Differentiable Self-Training

Marc Felix Brinner, Sina Zarriess

We propose an end-to-end differentiable training paradigm for stable training of a rationalized transformer classifier. Our approach results in a single model that simultaneously classifies a sample and scores input tokens based on their relevance to the classification. To this end, we build on the widely-used three-player-game for training rationalized models, which typically relies on training a rationale selector, a classifier and a complement classifier. We simplify this approach by making a single model fulfill all three roles, leading to a more efficient training paradigm that is not susceptible to the common training instabilities that plague existing approaches. Further, we extend this paradigm to produce class-wise rationales while incorporating recent advances in parameterizing and regularizing the resulting rationales, thus leading to substantially improved and state-of-the-art alignment with human annotations without any explicit supervision.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Consecutive Batch Model Editing with HooK Layers

Shuaiyi Li, Yang Deng, Deng Cai, Hongyuan Lu, Liang CHEN, Wai Lam

As the typical retraining paradigm is unacceptably time- and resource-consuming, researchers are turning to model editing to find an effective way that supports both consecutive and batch scenarios to edit the model behavior directly. Despite all these practical expectations, existing model editing methods fail to realize all of them. Furthermore, the memory demands for such sequential model editing approaches tend to be prohibitive, frequently necessitating an external memory that grows incrementally over time. To cope with these challenges, we propose CoachHooK, a model editing method that simultaneously supports sequential and batch editing. CoachHooK is memory-friendly as it only needs a small amount of it to store several hook layers whose size remains unchanged over time. Experimental results demonstrate the superiority of our method over other batch-supportive model editing methods under both single-round and consecutive batch editing scenarios. Extensive analyses of CoachHooK have been conducted to verify the stability of our method over a number of consecutive steps.

Nov 14 (Thu) 10:30-12:00 - Jasmine

CONTESTS: a Framework for Consistency Testing of Span Probabilities in Language Models

Eitan Wagner, Yuli Slavutsky, Omri Abend

Although language model scores are often treated as probabilities, their reliability as probability estimators has mainly been studied through calibration, overlooking other aspects. In particular, it is unclear whether language models produce the same value for different ways of assign-

⁷<https://lmexplainer.github.io/xplainllm>

ing joint probabilities to word spans. Our work introduces a novel framework, ConTestS (Consistency Testing over Spans), involving statistical tests to assess score consistency across interchangeable completion and conditioning orders. We conduct experiments on post-release real and synthetic data to eliminate training effects. Our findings reveal that both Masked Language Models (MLMs) and autoregressive models exhibit inconsistent predictions, with autoregressive models showing larger discrepancies. Larger MLMs tend to produce more consistent predictions, while autoregressive models show the opposite trend. Moreover, for both model types, prediction entropies offer insights into the true word span likelihood and therefore can aid in selecting optimal decoding strategies. The inconsistencies revealed by our analysis, as well as their connection to prediction entropies and differences between model types, can serve as useful guides for future research on addressing these limitations.

Nov 14 (Thu) 10:30-12:00 - Jasmine

The Illusion of Competence: Evaluating the Effect of Explanations on Users' Mental Models of Visual Question Answering Systems

Judith Sieker, Simeon Jenker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, Sina Zarrieß

We examine how users perceive the limitations of an AI system when it encounters a task that it cannot perform perfectly and whether providing explanations alongside its answers aids users in constructing an appropriate mental model of the system's capabilities and limitations. We employ a visual question answer and explanation task where we control the AI system's limitations by manipulating the visual inputs: during inference, the system either processes full-color or grayscale images. Our goal is to determine whether participants can perceive the limitations of the system. We hypothesize that explanations will make limited AI capabilities more transparent to users. However, our results show that explanations do not have this effect. Instead of allowing users to more accurately assess the limitations of the AI system, explanations generally increase users' perceptions of the system's competence – regardless of its actual performance.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Adaptation Odyssey in LLMs: Why Does Additional Pretraining Sometimes Fail to Improve?

Frai Öncel, Matthias Bethge, Beyza Ermiş, Mirco Ravanelli, Cem Subakan, Çaataý Yıldız

In the last decade, the generalization and adaptation abilities of deep learning models were typically evaluated on fixed training and test distributions. Contrary to traditional deep learning, large language models (LLMs) are (i) even more overparameterized, (ii) trained on unlabeled text corpora curated from the Internet with minimal human intervention, and (iii) trained in an online fashion. These stark contrasts prevent researchers from transferring lessons learned on model generalization and adaptation in deep learning contexts to LLMs. To this end, our short paper introduces empirical observations that aim to shed light on further training of already pretrained language models. Specifically, we demonstrate that training a model on a text domain could degrade its perplexity on the test portion of the same domain. We observe with our subsequent analysis that the performance degradation is positively correlated with the similarity between the additional and the original pretraining dataset of the LLM. Our further token-level perplexity analysis reveals that the perplexity degradation is due to a handful of tokens that are not informative about the domain. We hope these findings will guide us in determining when to adapt a model vs when to rely on its foundational capabilities.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Can LLMs Learn Uncertainty on Their Own? Expressing Uncertainty Effectively in A Self-Training Manner

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, Min Zhang

Large language models (LLMs) often exhibit excessive, random, and uninformative uncertainty, rendering them unsuitable for decision-making in human-computer interactions. In this paper, we aim to instigate a heightened awareness of self-uncertainty in LLMs, enabling them to express uncertainty more effectively. To accomplish this, we propose an uncertainty-aware instruction tuning (UaIT) method, aligning LLMs' perception with the probabilistic uncertainty of the generation. We conducted experiments using LLaMA2 and Mistral on multiple free-form QA tasks. Experimental results revealed a surprising 45.2% improvement in the effectiveness of uncertainty expression by LLMs, accompanied by reasonably good out-of-domain generalization capabilities. Moreover, this uncertainty expression can serve as a valuable real-time basis for human decision-making, e.g., retrieving external documents and incorporating stronger LLMs.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Rebuilding ROME : Resolving Model Collapse during Sequential Model Editing

Akshat Gupta, Sidharth Baskaran, Gopala Anumanchipalli

Recent work using Rank-One Model Editing (ROME), a popular model editing method, has shown that there are certain facts that the algorithm is unable to edit without breaking the model. Such edits have previously been called disabling edits. These disabling edits cause immediate model collapse and limits the use of ROME for sequential editing. In this paper, we show that disabling edits are a artifact of irregularities in the implementation of ROME. With this paper, we provide a more stable implementation ROME, which we call r-ROME and show that model collapse is no longer observed when making large scale sequential edits with r-ROME, while further improving generalization and locality of model editing compared to the original implementation of ROME. We also provide a detailed mathematical explanation of the reason behind disabling edits.

Nov 14 (Thu) 10:30-12:00 - Jasmine

FASTTRACK: Reliable Fact Tracing via Clustering and LLM-Powered Evidence Validation

Si Chen, Feiyang Kang, Ning Yu, Ruoxi Jia

Fact tracing seeks to identify specific training examples that serve as the knowledge source for a given query. Existing approaches to fact tracing rely on assessing the similarity between each training sample and the query along a certain dimension, such as lexical similarity, gradient, or embedding space. However, these methods fall short of effectively distinguishing between samples that are merely relevant and those that actually provide supportive evidence for the information sought by the query. This limitation often results in suboptimal effectiveness. Moreover, these approaches necessitate the examination of the similarity of individual training points for each query, imposing significant computational demands and creating a substantial barrier for practical applications. This paper introduces FASTTRACK, a novel approach that harnesses the capabilities of Large Language Models (LLMs) to validate supportive evidence for queries and at the same time clusters the training database towards a reduced extent for LLMs to trace facts. Our experiments show that FASTTRACK substantially outperforms existing methods in both accuracy and efficiency, achieving more than 100% improvement in F1 score over the state-of-the-art methods while being x33 faster than TracIn.

Nov 14 (Thu) 10:30-12:00 - Jasmine

On the token distance modeling ability of higher RoPE attention dimension

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, Jie Zhou

Length extrapolation algorithms based on Rotary position embedding (RoPE) have shown promising results in extending the context length of language models. However, understanding how position embedding can capture longer-range contextual information remains elusive. Based on the intuition that different dimensions correspond to different frequency of changes in RoPE encoding, we conducted a dimension-level analysis to investigate the correlation between a hidden dimension of an attention head and its contribution to capturing long-distance dependencies. Using our correlation metric, we identified a particular type of attention heads, which we named Positional Heads, from various length-extrapolated models. These heads exhibit a strong focus on long-range information interaction and play a pivotal role in long input

processing, as evidence by our ablation. We further demonstrate the correlation between the efficiency of length extrapolation and the extension of the high-dimensional attention allocation of these heads. The identification of Positional Heads provides insights for future research in long-text comprehension.

Nov 14 (Thu) 10:30-12:00 - Jasmine

RippleCOT: Amplifying Ripple Effect of Knowledge Editing in Language Models via Chain-of-Thought In-Context Learning

Zihao Zhao, Yuchen Yang, Yijiang Li, Yinzh Cao

The ripple effect poses a significant challenge in knowledge editing for large language models. Namely, when a single fact is edited, the model struggles to accurately update the related facts in a sequence, which is evaluated by multi-hop questions linked to a chain of related facts. Recent strategies have moved away from traditional parameter updates to more flexible, less computation-intensive methods, proven to be more effective in the ripple effect. In-context learning (ICL) editing uses a simple demonstration *Imagine that + new fact* to guide LLMs, but struggles with complex multi-hop questions as the new fact alone fails to specify the chain of facts involved in such scenarios. Besides, memory-based editing maintains additional storage for all edits and related facts, requiring continuous updates to stay effective. As a result of the design limitations, the challenge remains, with the highest accuracy being only 33.8% on the MQuAKE-CF benchmarks for Vicuna-7B. To address this, we propose RippleCOT, a novel ICL editing approach integrating Chain-of-Thought (COT) reasoning. RippleCOT structures demonstrations as *new fact, question, thought, answer*, incorporating a *thought* component to identify and decompose the multi-hop logic within questions. This approach effectively guides the model through complex multi-hop questions with chains of related facts. Comprehensive experiments demonstrate that RippleCOT significantly outperforms the state-of-the-art in the ripple effect, achieving accuracy gains ranging from 7.8% to 87.1%.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Measuring Susceptibility to Irrelevant Context in Language Models

Tianyu Liu, Kevin Du, Mrimaya Sachan, Ryan Cotterell

One strength of modern language models is their ability to incorporate information from a user-input context when answering queries. However, they are not equally sensitive to the subtle changes in that context. To quantify this, Du et al. (2024) gives an information-theoretic metric to measure such sensitivity. Their metric, susceptibility, is defined as the degree to which contexts can influence a model's response to a query at a distributional level. However, exactly computing susceptibility is difficult and, thus, Du et al. (2024) falls back on a Monte Carlo approximation. Due to the large number of samples required, the Monte Carlo approximation is inefficient in practice. As a faster alternative, we propose Fisher susceptibility, an efficient method to estimate the susceptibility based on Fisher information. Empirically, we validate that Fisher susceptibility is comparable to Monte Carlo estimated susceptibility across a diverse set of query domains despite its being $70 \times$ faster. Exploiting the improved efficiency, we apply Fisher susceptibility to analyze factors affecting the susceptibility of language models. We observe that larger models are as susceptible as smaller ones.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Learning Semantic Structure through First-Order-Logic Translation

Akshay Chaturvedi, Nicholas Asher

In this paper, we study whether transformer-based language models can extract predicate argument structure from simple sentences. We firstly show that language models sometimes confuse which predicates apply to which objects. To mitigate this, we explore two tasks: question answering (Q/A), and first order logic (FOL) translation, and two regimes, prompting and finetuning. In FOL translation, we finetune several large language models on synthetic datasets designed to gauge their generalization abilities. For Q/A, we finetune encoder models like BERT and RoBERTa and use prompting for LLMs. The results show that FOL translation for LLMs is better suited to learn predicate argument structure.

Nov 14 (Thu) 10:30-12:00 - Jasmine

SynthEval: Hybrid Behavioral Testing of NLP Models with Synthetic Evaluation

Raoyuan Zhao, Abdullatif Kóksal, Yihong Liu, Leonie Weissweiler, Anna Korhonen, Hinrich Schütze

Traditional benchmarking in NLP typically involves using static, held-out test sets and calculating aggregated statistics based on diverse examples. However, this approach often results in an overestimation of performance and lacks the ability to offer comprehensive, interpretable, and dynamic assessments of NLP models. Recently, works like DynaBench and Checklist have addressed these limitations through behavioral testing of NLP models with test types generated by a multi-step human-annotated pipeline. Unfortunately, manually creating a variety of test types requires significant human labor, thus weakening efficiency. In this work, we propose SynthEval, a hybrid behavioral testing framework that leverages large language models (LLMs) to generate a wide range of test types for a comprehensive evaluation of NLP models. The SynthEval framework first generates sentences via LLMs using controlled generation, and then identifies challenging examples by comparing the predictions made by LLMs with task-specific NLP models. In the last stage, human experts investigate the challenging examples, manually design templates, and identify the types of failures the task-specific models consistently exhibit. We apply SynthEval to two classification tasks and show that our framework is effective in identifying weaknesses of strong models on these tasks.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Knowledge Mechanisms in Large Language Models: A Survey and Perspective

Mengru Huang, Yunzhi Yao, Ziwen Xu, Shuohei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, Ningyu Zhang

Understanding knowledge mechanisms in Large Language Models (LLMs) is crucial for advancing towards trustworthy AGI. This paper reviews knowledge mechanism analysis from a novel taxonomy including knowledge utilization and evolution. Knowledge utilization delves into the mechanism of memorization, comprehension and application, and creation. Knowledge evolution focuses on the dynamic progression of knowledge within individual and group LLMs. Moreover, we discuss what knowledge LLMs have learned, the reasons for the fragility of parametric knowledge, and the potential dark knowledge (hypothesis) that will be challenging to address. We hope this work can help understand knowledge in LLMs and provide insights for future research.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Activation Scaling for Attribution and Intervention in Language Models

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, Aaron Schein

Given the prompt "Rome is in", can we steer a language model to flip its prediction of an incorrect token "France" to a correct token "Italy" by only multiplying a few relevant activation vectors with scalars? We argue that successfully intervening on a model is a prerequisite for interpreting its internal workings. Concretely, we establish a three-term objective: a successful intervention should flip the correct with the wrong token and vice versa (effectiveness), and leave other tokens unaffected (faithfulness), all while being sparse (minimality). Using gradient-based optimization, this objective lets us learn (and later evaluate) a specific kind of efficient and interpretable intervention: activation scaling only modifies the signed magnitude of activation vectors to strengthen, weaken, or reverse the steering directions already encoded in the model. On synthetic tasks, this intervention performs comparably with steering vectors in terms of effectiveness and faithfulness, but is much more minimal allowing us to pinpoint interpretable model components. We evaluate activation scaling from different angles, com-

pare performance on different datasets, and make activation scalars a learnable function of the activation vectors themselves to generalize to varying-length prompts.

Nov 14 (Thu) 10:30-12:00 - Jasmine

On the Similarity of Circuits across Languages: a Case Study on the Subject-verb Agreement Task

Javier Ferrando, Marta R. Costa-jussà

Several algorithms implemented by language models have recently been successfully reversed-engineered. However, these findings have been concentrated on specific tasks and models, leaving it unclear how universal circuits are across different settings. In this paper, we study the circuits implemented by Gemma 2B for solving the subject-verb agreement task across two different languages, English and Spanish. We discover that both circuits are highly consistent, being mainly driven by a particular attention head writing a ‘subject number’ signal to the last residual stream, which is read by a small set of neurons in the final MLPs. Notably, this subject number signal is represented as a direction in the residual stream space, and is language-independent. Finally, we demonstrate this direction has a causal effect on the model predictions, effectively flipping the Spanish predicted verb number by intervening with the direction found in English.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Knowledge-based Consistency Testing of Large Language Models

Sai Sathiesh Rajan, Ezequiel Soremekun, Sudipta Chattopadhyay

In this work, we systematically expose and measure the inconsistency and knowledge gaps of Large Language Models (LLMs). Specifically, we propose an automated testing framework (called KONTEST) which leverages a knowledge graph to construct test cases. KONTEST probes and measures the inconsistencies in the LLMs knowledge of the world via a combination of semantically-equivalent queries and test oracles (metamorphic or ontological oracle). KONTEST further mitigates knowledge gaps via a weighted LLM model ensemble. Using four state-of-the-art LLMs (Falcon, Gemini, GPT3.5, and Llama2), we show that KONTEST generates 19.2% error inducing inputs (1917 errors from 9979 test inputs). It also reveals a 16.5% knowledge gap across all tested LLMs. A mitigation method informed by KONTEST's test suite reduces LLM knowledge gap by 32.48%. Our ablation study further shows that GPT3.5 is not suitable for knowledge-based consistency testing because it is only 60%-68% effective in knowledge construction.

Nov 14 (Thu) 10:30-12:00 - Jasmine

CERT-ED: Certifiable Robust Text Classification for Edit Distance

Zhuoguo Huang, Neil G Marchant, Olga Ohrimenko, Benjamin I. P. Rubinstein

With the growing integration of AI in daily life, ensuring the robustness of systems to inference-time attacks is crucial. Among the approaches for certifying robustness to such adversarial examples, randomized smoothing has emerged as highly promising due to its nature as a wrapper around arbitrary black-box models. Previous work on randomized smoothing in natural language processing has primarily focused on specific subsets of edit distance operations, such as synonym substitution or word insertion, without exploring the certification of all edit operations. In this paper, we adapt Randomized Deletion (Huang et al., 2023) and propose, CERTified Edit Distance defense (CERT-ED) for natural language classification. Through comprehensive experiments, we demonstrate that CERT-ED outperforms the existing Hamming distance method RanMASK (Zeng et al., 2023) in 4 out of 5 datasets in terms of both accuracy and the cardinality of the certificate. By covering various threat models, including 5 direct and 5 transfer attacks, our method improves empirical robustness in 38 out of 50 settings.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Functionality learning through specification instructions

Pedro Henrique Luz de Araújo, Benjamin Roth

Test suites assess natural language processing models' performance on specific functionalities: cases of interest involving model robustness, fairness, or particular linguistic capabilities. This paper introduces specification instructions: text descriptions specifying fine-grained task-specific behaviors. For each functionality in a suite, we generate an instruction that describes it. We combine the specification instructions to create specification-augmented prompts, which we feed to language models pre-trained on natural instruction data. We conduct experiments to measure how optimizing for some functionalities may negatively impact functionalities that are not covered by the specification set. Our analyses across four tasks and models of diverse sizes and families show that smaller models struggle to follow specification instructions. However, larger models (> 3 B params.) can benefit from specifications and—surprisingly—even generalize certain desirable behaviors across functionalities.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Scaling Laws for Fact Memorization of Large Language Models

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xipeng Qiu

Fact knowledge memorization is crucial for Large Language Models (LLM) to generate factual and reliable responses. However, the behaviors of LLM fact memorization remain under-explored. In this paper, we analyze the scaling laws for LLM's fact knowledge and LLMs' behaviors of memorizing different types of facts. We find that LLMs' fact knowledge capacity has a linear and negative exponential law relationship with model size and training epochs, respectively. Estimated by the built scaling law, memorizing the whole Wikidata's facts requires training an LLM with 1000B non-embed parameters for 100 epochs, suggesting that using LLMs to memorize all public facts is almost implausible for a general pre-training setting. Meanwhile, we find that LLMs can generalize on unseen fact knowledge and its scaling law is similar to general pre-training. Additionally, we analyze the compatibility and preference of LLMs' fact memorization. For compatibility, we find LLMs struggle with memorizing redundant facts in a unified way. Only when correlated facts have the same direction and structure, the LLM can compatibly memorize them. This shows the inefficiency of LLM memorization for redundant facts. For preference, the LLM pays more attention to memorizing more frequent and difficult facts, and the subsequent facts can overwrite prior facts' memorization, which significantly hinders low-frequency facts memorization. Our findings reveal the capacity and characteristics of LLMs' fact knowledge learning, which provide directions for LLMs' fact knowledge augmentation.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Gender Identity in Pretrained Language Models: An Inclusive Approach to Data Creation and Probing

Urban Knupke, Agnieszka Falenska, Filip Mileti

Pretrained language models (PLMs) have been shown to encode binary gender information of text authors, raising the risk of skewed representations and downstream harms. This effect is yet to be examined for transgender and non-binary identities, whose frequent marginalization may exacerbate harmful system behaviors. Addressing this gap, we first create TRANSCRIPT, a corpus of YouTube transcripts from transgender, cisgender, and non-binary speakers. Using this dataset, we probe various PLMs to assess if they encode the gender identity information, examining both frozen and fine-tuned representations as well as representations for inputs with author-specific words removed. Our findings reveal that PLM representations encode information for all gender identities but to different extents. The divergence is most pronounced for cis women and non-binary individuals, underscoring the critical need for gender-inclusive approaches to NLP systems.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Pre-trained Language Models Return Distinguishable Probability Distributions to Unfaithfully Hallucinated Texts

Taejun Cha, Donghun Lee

In this work, we show the pre-trained language models return distinguishable generation probability and uncertainty distribution to unfaithfully hallucinated texts, regardless of their size and structure. By examining 24 models on 6 data sets, we find out that 88-98% of cases return statistically significantly distinguishable generation probability and uncertainty distributions. Using this general phenomenon, we showcase a hallucination-reducing training algorithm. Our algorithm outperforms other baselines by achieving higher faithfulness metrics while maintaining sound general text quality measures.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Robust Text Classification: Analyzing Prototype-Based Networks

Zhiwar Sourati, Darshan Girish Deshpande, Filip Ilievski, Kiril Gashevski, Sascha Saralajew

Downstream applications often require text classification models to be accurate and robust. While the accuracy of state-of-the-art Language Models (LMs) approximates human performance, they often exhibit a drop in performance on real-world noisy data. This lack of robustness can be concerning, as even small perturbations in text, irrelevant to the target task, can cause classifiers to incorrectly change their predictions. A potential solution can be the family of Prototype-Based Networks (PBNs) that classifies examples based on their similarity to prototypical examples of a class (prototypes) and has been shown to be robust to noise for computer vision tasks. In this paper, we study whether the robustness properties of PBNs transfer to text classification tasks under both targeted and static adversarial attack settings. Our results show that PBNs, as a mere architectural variation of vanilla LMs, offer more robustness compared to vanilla LMs under both targeted and static settings. We showcase how PBNs' interpretability can help us understand PBNs' robustness properties. Finally, our ablation studies reveal the sensitivity of PBNs' robustness to the strictness of clustering and the number of prototypes in the training phase, as tighter clustering and a low number of prototypes result in less robust PBNs.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Skills-in-Context: Unlocking Compositionality in Large Language Models

Jiaoy Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, Jianshu Chen

We investigate how to elicit compositional generalization capabilities in large language models (LLMs). Compositional generalization empowers LLMs to solve complex problems by combining foundational skills, a critical reasoning ability akin to human intelligence. However, even the most advanced LLMs currently struggle with this form of reasoning. We examine this problem within the framework of in-context learning and find that demonstrating both foundational skills and compositional examples grounded in these skills within the same prompt context is crucial. We refer to this prompt structure as skills-in-context (SKiC). With as few as two exemplars, this in-context learning structure enables LLMs to tackle more challenging problems requiring innovative skill combinations, achieving near-perfect systematic generalization across a broad range of tasks. Intriguingly, SKiC also unlocks the latent potential of LLMs, allowing them to more actively utilize pre-existing internal skills acquired during earlier pretraining stages to solve complex reasoning problems. The SKiC structure is robust across different skill constructions and exemplar choices and demonstrates strong transferability to new tasks. Finally, inspired by our in-context learning study, we show that fine-tuning LLMs with SKiC-style data can elicit zero-shot weak-to-strong generalization, enabling the models to solve much harder problems directly with standard prompting.

Nov 14 (Thu) 10:30-12:00 - Jasmine

Performance Trade-offs of a Family of Text Watermarks

Anirudh Ajith, Sameer Singh, Danish Pruthi

Watermarking involves implanting an imperceptible signal into generated text that can later be detected via statistical tests. A prominent family of watermarking strategies for LLMs embeds this signal by upsampling a (pseudorandomly-chosen) subset of tokens at every generation step. However, such signals alter the model's output distribution and can have unintended effects on its downstream performance. In this work, we evaluate the performance of LLMs watermarking using three different strategies over a diverse suite of tasks including those cast as k-class classification (CLS), multiple choice question answering (MCQ), short-form generation (e.g., open-ended question answering) and long-form generation (e.g., translation) tasks. We find that watermarks (under realistic hyperparameters) can cause significant drops in LLMs' effective utility across all tasks. We observe drops of 10-20% in CLS tasks in the average case, which shoot up to 100% in the worst case. We notice degradations of about 7% in MCQ tasks, 10-15% in short-form generation, and 5-15% in long-form generation tasks. Our findings highlight the trade-offs that users should be cognizant of when using watermark models.

Nov 14 (Thu) 10:30-12:00 - Jasmine

From Internal Conflict to Contextual Adaptation of Language Models

Sara Vera Marjanovic, Haewn Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, Isabelle Augenstein

Knowledge-intensive language understanding tasks require Language Models (LMs) to integrate relevant context, mitigating their inherent weaknesses, such as incomplete or outdated knowledge. However, conflicting knowledge can be present in the LMs parameters, termed intra-memory conflict, which can affect a models propensity to accept contextual knowledge. To study the effect of intra-memory conflict on LMs ability to accept the relevant context, we utilise two knowledge conflict measures and a novel dataset containing inherently conflicting data, DYNAMICQA. This dataset includes facts with a temporal dynamic nature where facts can change over time and disputable dynamic facts, which can change depending on the viewpoint. DYNAMICQA is the first to include real-world knowledge conflicts and provide context to study the link between the different types of knowledge conflicts. We also evaluate several measures on their ability to reflect the presence of intra-memory conflict: semantic entropy and a novel coherent persuasion score. With our extensive experiments, we verify that LMs show a greater degree of intra-memory conflict with dynamic facts compared to facts that have a single truth value. Further, we reveal that facts with intra-memory conflict are harder to update with context, suggesting that retrieval-augmented generation will struggle with the most commonly adapted facts

Nov 14 (Thu) 10:30-12:00 - Jasmine

When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, David Jurgens

Prompting serves as the major way humans interact with Large Language Models (LLM). Commercial AI systems commonly define the role of the LLM in system prompts. For example, ChatGPT uses "You are a helpful assistant" as part of its default system prompt. Despite current practices of adding personas to system prompts, it remains unclear how different personas affect a model's performance on objective tasks. In this study, we present a systematic evaluation of personas in system prompts. We curate a list of 162 roles covering 6 types of interpersonal relationships and 8 domains of expertise. Through extensive analysis of 4 popular families of LLMs and 2,410 factual questions, we demonstrate that adding personas in system prompts does not improve model performance across a range of questions compared to the control setting where no persona is added. Nevertheless, further analysis suggests that the gender, type, and domain of the persona can all influence the resulting prediction accuracies. We further experimented with a list of persona search strategies and found that, while aggregating results from the best persona for each question significantly improves prediction accuracy, automatically identifying the best persona is challenging, with predictions often performing no better than random selection. Overall, our findings suggest that while adding a persona may lead to performance gains in certain settings, the effect of each persona can be largely random. Our results can help inform the design of system

prompts for AI systems. Code and data are available at <https://github.com/Jiaxin-Pei/Prompting-with-Social-Roles>.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

To Know or Not To Know? Analyzing Self-Consistency of Large Language Models under Ambiguity

Anastasia Sedova, Robert Litschko, Diego Frassinelli, Benjamin Roth, Barbara Plank

One of the major aspects contributing to the striking performance of large language models (LLMs) is the vast amount of factual knowledge accumulated during pre-training. Yet, many LLMs suffer from self-inconsistency, which raises doubt about their trustworthiness and reliability. This paper focused on entity type ambiguity, analyzing the proficiency and consistency of state-of-the-art LLMs in applying factual knowledge when prompted with ambiguous entities. To do so, we propose an evaluation protocol that disentangles knowing from applying knowledge, and test state-of-the-art LLMs on 49 ambiguous entities. Our experiments reveal that LLMs struggle with choosing the correct entity reading, achieving an average accuracy of only 85%, and as low as 75% with underspecified prompts. The results also reveal systematic discrepancies in LLM behavior, showing that while the models may possess knowledge, they struggle to apply it consistently, exhibit biases toward preferred readings, and display self-inconsistencies. This highlights the need to address entity ambiguity in the future for more trustworthy LLMs.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Faithful and Plausible Natural Language Explanations for Image Classification: A Pipeline Approach

Adam Wojciechowski, Mateusz Lango, Ondrej Dusek

Existing explanation methods for image classification struggle to provide faithful and plausible explanations. This paper addresses this issue by proposing a post-hoc natural language explanation method that can be applied to any CNN-based classifier without altering its training process or affecting predictive performance. By analysing influential neurons and the corresponding activation maps, the method generates a faithful description of the classifier's decision process in the form of a structured meaning representation, which is then converted into text by a language model. Through this pipeline approach, the generated explanations are grounded in the neural network architecture, providing accurate insight into the classification process while remaining accessible to non-experts. Experimental results show that the NLEs constructed by our method are significantly more plausible and faithful than baselines. In particular, user interventions in the neural network structure (masking of neurons) are three times more effective.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Enhancing Healthcare LLM Trust with Atypical Presentations Recalibration

Jeremy Qin, Bang Liu, Quoc Dinh Nguyen

Black-box large language models (LLMs) are increasingly deployed in various environments, making it essential for these models to effectively convey their confidence and uncertainty, especially in high-stakes settings. However, these models often exhibit overconfidence, leading to potential risks and misjudgments. Existing techniques for eliciting and calibrating LLM confidence have primarily focused on general reasoning datasets, yielding only modest improvements. Accurate calibration is crucial for informed decision-making and preventing adverse outcomes but remains challenging due to the complexity and variability of tasks these models perform. In this work, we investigate the miscalibration behavior of black-box LLMs within the healthcare setting. We propose a novel method, Atypical Presentations Recalibration, which leverages atypical presentations to adjust the model's confidence estimates. Our approach significantly improves calibration, reducing calibration errors by approximately 60% on three medical question answering datasets and outperforming existing methods such as vanilla verbalized confidence, CoT verbalized confidence and others. Additionally, we provide an in-depth analysis of the role of atypicality within the recalibration framework.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Diciphering the Factors Influencing the Efficacy of Chain-of-Thought: Probability, Memorization, and Noisy Reasoning

Akshara Prabhakar, Thomas L. Griffiths, R. Thomas McCoy

Chain-of-Thought (CoT) prompting has been shown to enhance the multi-step reasoning capabilities of Large Language Models (LLMs). However, debates persist about whether LLMs exhibit *abstract generalization* or rely on *shallow heuristics* when given CoT prompts. To understand the factors influencing CoT reasoning we provide a detailed case study of the symbolic reasoning task of decoding shift ciphers, where letters are shifted forward some number of steps in the alphabet. We analyze the pattern of results produced by three LLMs—GPT-4, Claude 3, and Llama 3.1—performing this task using CoT prompting. By focusing on a single relatively simple task, we are able to identify three factors that systematically affect CoT performance: the probability of the task's expected output (probability), what the model has implicitly learned during pre-training (memorization), and the number of intermediate operations involved in reasoning (noisy reasoning). We show that these factors can drastically influence task accuracy across all three LLMs; e.g., when tested with GPT-4, varying the output's probability of occurrence shifts accuracy from 26% to 70%. Overall, we conclude that CoT prompting performance reflects both memorization and a probabilistic version of genuine reasoning.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

The Fall of ROME: Understanding the Collapse of LLMs in Model Editing

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, Huawei Shen

Despite significant progress in model editing methods, their application in real-world scenarios remains challenging as they often cause large language models (LLMs) to collapse. Among them, ROME is particularly concerning, as it could disrupt LLMs with only a single edit. In this paper, we study the root causes of such collapse. Through extensive analysis, we identify two primary factors that contribute to the collapse: i) inconsistent handling of prefixed and unprefixed keys in the parameter update equation may result in very small denominators, causing excessively large parameter updates; ii) the subject of collapse cases is usually the first token, whose unprefixed key distribution significantly differs from the prefixed key distribution in autoregressive transformers, causing the aforementioned issue to materialize. To validate our findings, we propose a simple yet effective approach: uniformly using prefixed keys during editing phase and adding prefixes during testing phase to ensure the consistency between training and testing. The experimental results show that the proposed solution can prevent model collapse while maintaining the effectiveness of the edits.

Nov 14 (Thu) 10:30-12:00 - *Jasmine*

Axis Tour: Word Tour Determines the Order of Axes in ICA-transformed Embeddings

Hiroyuki Yamagishi, Yusuke Takase, Hidetoshi Shimodaira

Word embedding is one of the most important components in natural language processing, but interpreting high-dimensional embeddings remains a challenging problem. To address this problem, Independent Component Analysis (ICA) is identified as an effective solution. ICA-transformed word embeddings reveal interpretable semantic axes; however, the order of these axes are arbitrary. In this study, we focus on this property and propose a novel method, Axis Tour, which optimizes the order of the axes. Inspired by Word Tour, a one-dimensional word embedding method, we aim to improve the clarity of the word embedding space by maximizing the semantic continuity of the axes. Furthermore, we show through experiments on downstream tasks that Axis Tour yields better or comparable low-dimensional embeddings compared to both PCA and ICA.

Machine Learning for NLP 3

Nov 14 (Thu) 10:30-12:00 - Room: Riverfront Hall

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training

Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, Ivan P. Yamschikov

Language models can greatly benefit from efficient tokenization. However, they still mostly utilize the classical Byte-Pair Encoding (BPE) algorithm, a simple and reliable method. BPE has been shown to cause such issues as under-trained tokens and sub-optimal compression that may affect the downstream performance. We introduce PickyBPE, a modified BPE algorithm that carries out vocabulary refinement during tokenizer training by removing merges that leave intermediate "junk" tokens. Our method improves vocabulary efficiency, eliminates under-trained tokens, and does not compromise text compression. Our experiments show that this method either improves downstream performance or does not harm it.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

BPO: Supercharging Online Preference Learning by Adhering to the Proximity of Behavior LLM

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li

Direc alignment from preferences (DAP) has emerged as a promising paradigm for aligning large language models (LLMs) to human desiderata from pre-collected, offline preference datasets. While recent indications indicate that existing offline DAP methods can directly benefit from online training samples, we highlight the need to develop specific online DAP algorithms to fully harness the power of online training. Specifically, we identify that the learned LLM should adhere to the proximity of the behavior LLM, which collects the training samples. To this end, we propose online Preference Optimization in proximity to the Behavior LLM (BPO), emphasizing the importance of constructing a proper trust region for LLM alignment. We conduct extensive experiments to validate the effectiveness and applicability of our approach by integrating it with various DAP methods, resulting in significant performance improvements across a wide range of tasks when training with the same amount of preference data. Even when only introducing one additional data collection phase, our online BPO improves its offline DAP baseline from 72.0% to 80.2% on TL:DR and from 82.2% to 89.1% on Anthropic Helpfulness in terms of win rate against human reference text.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Where Am I From? Identifying Origin of LLM-generated Content

Liyang Li, Yihuan Bai, Minhao Cheng

Generative models, particularly large language models (LLMs), have achieved remarkable success in producing natural and high-quality content. However, their widespread adoption raises concerns regarding copyright infringement, privacy violations, and security risks associated with AI-generated content. To address these concerns, we propose a novel digital forensics framework for LLMs, enabling the tracing of AI-generated content back to its source. This framework embeds a secret watermark directly into the generated output, eliminating the need for model retraining. To enhance traceability, especially for short outputs, we introduce a "depth watermark" that strengthens the link between content and generator. Our approach ensures accurate tracing while maintaining the quality of the generated content. Extensive experiments across various settings and datasets validate the effectiveness and robustness of our proposed framework.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

The Mystery of the Pathological Path-star Task for Language Models

Arvid Frydenlund

The recently introduced path-star task is a minimal task designed to exemplify limitations to the abilities of language models (Bachmann and Nagarajan, 2024). It involves a path-star graph where multiple arms radiate from a single starting node and each node is unique. Given the start node and a specified target node that ends an arm, the task is to generate the arm containing that target node. This is straightforward for a human but surprisingly difficult for language models, which did not outperform the random baseline. The authors hypothesized this is due to a deficiency in teacher-forcing and the next-token prediction paradigm. We demonstrate the task is learnable using teacher-forcing in alternative settings and that the issue is partially due to representation. We introduce a regularization method using structured samples of the same graph but with differing target nodes, improving results across a variety of model types. We provide RASP proofs showing the task is theoretically solvable. Finally, we find settings where an encoder-only model can consistently solve the task.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Bridging Local Details and Global Context in Text-Attributed Graphs

Yaoke Wang, Yun Zhu, Wengqiao Zhang, Yuetong Zhuang, liyunfei, Siliang Tang

Representation learning on text-attributed graphs (TAGs) is vital for real-world applications, as they combine semantic textual and contextual structural information. Research in this field generally consist of two main perspectives: local-level encoding and global-level aggregating, respectively refer to textual node information unification (*e.g.*, using Language Models) and structure-augmented modeling (*e.g.*, using Graph Neural Networks). Most existing works focus on combining different information levels but overlook the interconnections, *i.e.*, the contextual textual information among nodes, which provides semantic insights to bridge local and global levels. In this paper, we propose GraphBridge, a *multi – granularity integration* framework that bridges local and global perspectives by leveraging contextual textual information, enhancing fine-grained understanding of TAGs. Besides, to tackle scalability and efficiency challenges, we introduce a graph-aware token reduction module. Extensive experiments across various models and datasets show that our method achieves state-of-the-art performance, while our graph-aware token reduction module significantly enhances efficiency and solves scalability issues. Codes are available at <https://github.com/wykk00/GraphBridge>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Improving Discriminative Capability of Reward Models in RLHF Using Contrastive Learning

Lu Chen, Rui Zheng, Binghai Wang, Senjie Jin, Caishuang Huang, Junjie Ye, Zhihao Zhang, Yuhao Zhou, Zhiheng Xi, Tao Gui, Qi Zhang, Xuanjing Huang

Reinforcement Learning from Human Feedback (RLHF) is a crucial approach to aligning language models with human values and intentions. A fundamental challenge in this method lies in ensuring that the reward model accurately understands and evaluates human preferences. Current methods rely on ranking losses to teach the reward model to assess preferences, but they are susceptible to noise and ambiguous data, often failing to deeply understand human intentions. To address this issue, we introduce contrastive learning into the reward modeling process. In addition to supervised ranking loss, we introduce an unsupervised contrastive loss to enable the reward model to fully capture the distinctions in contrastive data. Experimental results demonstrate that the proposed contrastive learning-based reward modeling method effectively enhances the generalization of the reward model, stabilizes the reinforcement learning training process, and improves the final

alignment with human preferences.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Unlocking Anticipatory Text Generation: A Constrained Approach for Large Language Models Decoding

Lifu Tu, Semih Yavuz, Jin Qu, Jiacheng Xu, Rui Meng, Caiming Xiong, Yingbo Zhou

Large Language Models (LLMs) have demonstrated a powerful ability for text generation. However, achieving optimal results with a given prompt or instruction can be challenging, especially for billion-sized models. Additionally, undesired behaviors such as toxicity or hallucinations can manifest. While much larger models (e.g., ChatGPT) may demonstrate strength in mitigating these issues, there is still no guarantee of complete prevention. In this work, we propose formalizing text generation as a future-constrained generation problem to minimize undesirable behaviors and enforce faithfulness to instructions. The estimation of future constraint satisfaction, accomplished using LLMs, guides the text generation process. Our extensive experiments demonstrate the effectiveness of the proposed approach across three distinct text generation tasks: keyword-constrained generation (Lin et al., 2020), toxicity reduction (Gehman et al., 2020), and factual correctness in question-answering (Gao et al., 2023).

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Context-Aware Adapter Tuning for Few-Shot Relation Learning in Knowledge Graphs

LIU Ran, Zhongzhou Liu, Xiaoli Li, Yuan Fang

Knowledge graphs (KGs) are instrumental in various real-world applications, yet they often suffer from incompleteness due to missing relations. To predict instances for novel relations with limited training examples, few-shot relation learning approaches have emerged, utilizing techniques such as meta-learning. However, the assumption is that novel relations in meta-testing and base relations in meta-training are independently and identically distributed, which may not hold in practice. To address the limitation, we propose RelAdapter, a context-aware adapter for few-shot relation learning in KGs designed to enhance the adaptation process in meta-learning. First, RelAdapter is equipped with a lightweight adapter module that facilitates relation-specific, tunable adaptation of meta-knowledge in a parameter-efficient manner. Second, RelAdapter is enriched with contextual information about the target relation, enabling enhanced adaptation to each distinct relation. Extensive experiments on three benchmark KGs validate the superiority of RelAdapter over state-of-the-art methods.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Calibrating Language Models with Adaptive Temperature Scaling

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, Chelsea Finn

The effectiveness of large language models (LLMs) is not only measured by their ability to generate accurate outputs but also by their calibration—how well their confidence scores reflect the probability of their outputs being correct. While unsupervised pre-training has been shown to yield LLMs with well-calibrated conditional probabilities, recent studies have shown that after fine-tuning with reinforcement learning from human feedback (RLHF), the calibration of these models degrades significantly. In this work, we introduce Adaptive Temperature Scaling (ATS), a post-hoc calibration method that predicts a temperature scaling parameter for each token prediction. The predicted temperature values adapt based on token-level features and are fit over a standard supervised fine-tuning (SFT) dataset. The adaptive nature of ATS addresses the varying degrees of calibration shift that can occur after RLHF fine-tuning. ATS improves calibration by over 10-50% across three downstream natural language evaluation benchmarks compared to prior calibration methods and does not impede performance improvements from RLHF.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Instruction Matters, a Simple yet Effective Task Selection Approach in Instruction Tuning for Specific Tasks

Changho Lee, Janghoon Han, Seonghyeon Ye, Stanley Jungkyu Choi, Honglak Lee, Kyunghoon Bae

Instruction tuning has been proven effective in enhancing zero-shot generalization across various tasks and in improving the performance of specific tasks. For task-specific improvements, strategically selecting and training on related tasks that provide meaningful supervision is crucial, as this approach enhances efficiency and prevents performance degradation from learning irrelevant tasks. In this light, we introduce a simple yet effective task selection method that leverages instruction information alone to identify relevant tasks, optimizing instruction tuning for specific tasks. Our method is significantly more efficient than traditional approaches, which require complex measurements of pairwise transferability between tasks or the creation of data samples for the target task. Additionally, by aligning the model with the unique instructional template style of the meta-dataset, we enhance its ability to granularly discern relevant tasks, leading to improved overall performance. Experimental results demonstrate that training on a small set of tasks, chosen solely based on the instructions, results in substantial improvements in performance on benchmarks such as P3, Big-Bench, NIV2, and Big-Bench Hard. Significantly, these improvements surpass those achieved by prior task selection methods, highlighting the superiority of our approach.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Are Data Augmentation Methods in Named Entity Recognition Applicable for Uncertainty Estimation?

Wataru Hashimoto, Hidetaka Kamigaito, Taro Watanabe

This work investigates the impact of data augmentation on confidence calibration and uncertainty estimation in Named Entity Recognition (NER) tasks. For the future advance of NER in safety-critical fields like healthcare and finance, it is essential to achieve accurate predictions with calibrated confidence when applying Deep Neural Networks (DNNs), including Pre-trained Language Models (PLMs), as a real-world application. However, DNNs are prone to miscalibration, which limits their applicability. Moreover, existing methods for calibration and uncertainty estimation are computational expensive. Our investigation in NER found that data augmentation improves calibration and uncertainty in cross-genre and cross-lingual setting, especially in-domain setting. Furthermore, we showed that the calibration for NER tends to be more effective when the perplexity of the sentences generated by data augmentation is lower, and that increasing the size of the augmentation further improves calibration and uncertainty.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Contrastive Policy Gradient: Aligning LLMs on sequence-level scores in a supervised-friendly fashion

Yannish Flet-Berliac, Nathan Grinstajn, Florian Strub, Eugene Choi, Bill Wu, Chris Cremer, Arash Ahmadian, Yash Chandak, Mohammad Gheshlaghi Azar, Olivier Pietquin, Matthieu Geist

Reinforcement Learning (RL) has been used to finetune Large Language Models (LLMs) using a reward model trained from preference data, to better align with human judgment. The recently introduced direct alignment methods, which are often simpler, more stable, and computationally lighter, can more directly achieve this. However, these approaches cannot optimize arbitrary rewards, and the preference-based ones are not the only rewards of interest for LLMs (eg. unit tests for code generation or textual entailment for summarization, among others). RL-finetuning is usually done with a variation of policy gradient, which calls for on-policy or near-on-policy samples, requiring costly generations. We introduce “Contrastive Policy Gradient”, or CoPG, a simple and mathematically principled new RL algorithm that can estimate the optimal policy even from off-policy data. It can be seen as an off-policy policy gradient approach that does not rely on important sampling techniques and highlights the importance of using (the right) state baseline. We show this approach to generalize the direct alignment method IPO (identity preference optimization) and classic policy gradient. We experiment with the proposed CoPG on a toy bandit problem to illustrate its properties, as well as for finetuning LLMs on a summarization task, using a learned reward function considered as ground truth for

the purpose of the experiments.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

On the Fragility of Active Learners for Text Classification

Abhishek Ghose, Emma Thuong Nguyen

Active learning (AL) techniques optimally utilize a labeling budget by iteratively selecting instances that are most valuable for learning. However, they lack "prerequisite checks", i.e., there are no prescribed criteria to pick an AL algorithm best suited for a dataset. A practitioner must pick a technique they trust would beat random sampling, based on prior reported results, and hope that it is resilient to the many variables in their environment: dataset, labeling budget and prediction pipelines. The important questions then are: how often on average, do we expect any AL technique to reliably beat the computationally cheap and easy-to-implement strategy of random sampling? Does it at least make sense to use AL in an "Always ON" mode in a prediction pipeline, so that while it might not always help, it never under-performs random sampling? How much of a role does the prediction pipeline play in AL's success? We examine these questions in detail for the task of text classification using pre-trained representations, which are ubiquitous today. Our primary contribution here is a rigorous evaluation of AL techniques, old and new, across setups that vary wrt datasets, text representations and classifiers. This unlocks multiple insights around warm-up times, i.e., number of labels before gains from AL are seen, viability of an "Always ON" mode and the relative significance of different factors. Additionally, we release a framework for rigorous benchmarking of AL techniques for text classification.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Immunization against harmful fine-tuning attacks

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszce, Hassan Sajjad, Frank Rudzicz

Large Language Models (LLMs) are often trained with safety guards intended to prevent harmful text generation. However, such safety training can be removed by fine-tuning the LLM on harmful datasets. While this emerging threat (harmful fine-tuning attacks) has been characterized by previous work, there is little understanding of how we should proceed in constructing and validating defenses against these attacks especially in the case where defenders would not have control of the fine-tuning process. We introduce a formal framework based on the training budget of an attacker which we call "Immunization" conditions. Using a formal characterisation of the harmful fine-tuning problem, we provide a thorough description of what a successful defense must comprise of and establish a set of guidelines on how rigorous defense research that gives us confidence should proceed.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Enhancing Large Language Model Based Sequential Recommender Systems with Pseudo Labels Reconstruction

Hyunsoo Na, Minseok Gang, Youngrok Ko, Jinseok Seol, Sang-goo Lee

Large language models (LLMs) are utilized in various studies, and they also demonstrate a potential to function independently as a recommendation model. Nevertheless, training sequences and text labels modifies LLMs' pre-trained weights, diminishing their inherent strength in constructing and comprehending natural language sentences. In this study, we propose a reconstruction-based LLM recommendation model (ReLRec) that harnesses the feature extraction capability of LLMs, while preserving LLMs' sentence generation abilities. We reconstruct the user and item pseudo-labels generated from user reviews, while training on sequential data, aiming to exploit the key features of both users and items. Experimental results demonstrate the efficacy of label reconstruction in sequential recommendation tasks.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Semi-Supervised Reward Modeling via Iterative Self-Training

Yifei He, Haoxiang Wang, Ziyuan Jiang, Alexandros Papangelis, Han Zhao

Reward models (RM) capture the values and preferences of humans and play a central role in Reinforcement Learning with Human Feedback (RLHF) to align pretrained large language models (LLMs). Traditionally, training these models relies on extensive human-annotated preference data, which poses significant challenges in terms of scalability and cost. To overcome these limitations, we propose Semi-Supervised Reward Modeling (SSRM), an approach that enhances RM training using unlabeled data. Given an unlabeled dataset, SSRM involves three key iterative steps: pseudo-labeling unlabeled examples, selecting high-confidence examples through a confidence threshold, and supervised finetuning on the refined dataset. Across extensive experiments on various model configurations, we demonstrate that SSRM significantly improves reward models without incurring additional labeling costs. Notably, SSRM can achieve performance comparable to models trained entirely on labeled data of equivalent volumes. Overall, SSRM substantially reduces the dependency on large volumes of human-annotated data, thereby decreasing the overall cost and time involved in training effective reward models.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

BIPEFT: Budget-Guided Iterative Search for Parameter Efficient Fine-Tuning of Large Pretrained Language Models

Aofei Chang, Jiapi Wang, Han Liu, Parminder Bharia, Cao Xiao, Ting Wang, Fenglong Ma

Parameter Efficient Fine-Tuning (PEFT) offers an efficient solution for fine-tuning large pretrained language models for downstream tasks. However, most PEFT strategies are manually designed, often resulting in suboptimal performance. Recent automatic PEFT approaches aim to address this but face challenges such as search space entanglement, inefficiency, and lack of integration between parameter budgets and search processes. To overcome these issues, we introduce a novel Budget-guided Iterative search strategy for automatic PEFT (BIPEFT), significantly enhancing search efficiency. BIPEFT employs a new iterative search strategy to disentangle the binary module and rank dimension search spaces. Additionally, we design early selection strategies based on parameter budgets, accelerating the learning process by gradually removing unimportant modules and fixing rank dimensions. Extensive experiments on public benchmarks demonstrate the superior performance of BIPEFT in achieving efficient and effective PEFT for downstream tasks with a low parameter budget.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Editing the Mind of Giants: An In-Depth Exploration of Pitfalls of Knowledge Editing in Large Language Models

Cheng-Hsuan Hsueh, Paul Kuo-Ming Huang, Tzu-Han Lin, CHE WEI LIAO, Hung-Chieh Fang, Chao-Wei Huang, Yun-Nung Chen

Knowledge editing is a rising technique for efficiently updating factual knowledge in large language models (LLMs) with minimal alteration of parameters. However, recent studies have identified side effects, such as knowledge distortion and the deterioration of general abilities, that have emerged after editing. Despite these findings, evaluating the pitfalls of knowledge editing often relies on inconsistent metrics and benchmarks, lacking a uniform standard. In response, this survey presents a comprehensive study of these side effects, providing a unified perspective on the challenges of knowledge editing in LLMs by conducting experiments with consistent metrics and benchmarks. Additionally, we review related works and outline potential research directions to address these limitations. Our survey highlights the limitations of current knowledge editing methods, emphasizing the need for a deeper understanding of the inner knowledge structures of LLMs and improved knowledge editing methods. To foster future research, we have released the complementary materials publicly ([https://github.com/MiuLab/EditLLM-Survey](https://github.com/MiuLab>EditLLM-Survey)).

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access

Jiayuan Su, Jing Luo, Hongwei Wang, Lu Cheng

This study aims to address the pervasive challenge of quantifying uncertainty in large language models (LLMs) with black-box API access. Conformal Prediction (CP), known for its model-agnostic and distribution-free features, is a desired approach for various LLMs and data distributions. However, existing CP methods for LLMs typically assume access to the logits, which are unavailable for some API-only LLMs. In addition, logits are known to be miscalibrated, potentially leading to degraded CP performance. To tackle these challenges, we introduce a novel CP method that (1) is tailored for API-only LLMs without logit-access; (2) minimizes the size of prediction sets; and (3) ensures a statistical guarantee of the user-defined coverage. The core idea of this approach is to formulate nonconformity measures using both coarse-grained (i.e., sample frequency) and fine-grained uncertainty notions (e.g., semantic similarity). Experimental results on both close-ended and open-ended Question Answering tasks show our approach can mostly outperform the logit-based CP baselines.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Position Paper: Data-Centric AI in the Age of Large Language Models

Xinyi Xu, Xiaoxuan Wu, Rui Qiao, Arun Verma, Yao Shu, Jingtan Wang, Xinyuan Niu, Zhenfeng He, Jiangwei Chen, Zijian Zhou, Gregory Kang Ruey Lau, Hieu Dao, Lucas Agusurja, Rachael Hwee Ling Sim, Xiaoqiang Lin, Wenyang Hu, Zhongxiang Dai, Pang Wei Koh, Bryan Kian Hsiang Low

This position paper proposes a data-centric viewpoint of AI research, focusing on large language models (LLMs). We start by making a key observation that data is instrumental in the developmental (e.g., pretraining and fine-tuning) and inferential stages (e.g., in-context learning) of LLMs, and advocate that data-centric research should receive more attention from the community. We identify four specific scenarios centered around data, covering data-centric benchmarks and data curation, data attribution, knowledge transfer, and inference contextualization. In each scenario, we underscore the importance of data, highlight promising research directions, and articulate the potential impacts on the research community and, where applicable, the society as a whole. For instance, we advocate for a suite of data-centric benchmarks tailored to the scale and complexity of data for LLMs. These benchmarks can be used to develop new data curation methods and document research efforts and results, which can help promote openness and transparency in AI and LLM research.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Efficient Pointwise-Pairwise Learning-to-Rank for News Recommendation

Nithish Kannan, Yao Ma, Gerrit J.J. Van den Burg, Jean Baptiste Faddoul

News recommendation is a challenging task that involves personalization based on the interaction history and preferences of each user. Recent works have leveraged the power of pretrained language models (PLMs) to directly rank news items by using inference approaches that predominately fall into three categories: pointwise, pairwise, and listwise learning-to-rank. While pointwise methods offer linear inference complexity, they fail to capture crucial comparative information between items that is more effective for ranking tasks. Conversely, pairwise and listwise approaches excel at incorporating these comparisons but suffer from practical limitations: pairwise approaches are either computationally expensive or lack theoretical guarantees and listwise methods often perform poorly in practice. In this paper, we propose a novel framework for PLM-based news recommendation that integrates both pointwise relevance prediction and pairwise comparisons in a scalable manner. We present a rigorous theoretical analysis of our framework, establishing conditions under which our approach guarantees improved performance. Extensive experiments show that our approach outperforms the state-of-the-art methods on the MIND and Adressa news recommendation datasets.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Out-of-Distribution Detection through Soft Clustering with Non-Negative Kernel Regression

Aryan Gulati, Xingjian Dong, Carlos Hurtado, Sarah Shekhhizbar, Swabha Swamyamdipta, Antonio Ortega

As language models become more general purpose, increased attention needs to be paid to detecting out-of-distribution (OOD) instances, i.e., those not belonging to any of the distributions seen during training. Existing methods for detecting OOD data are computationally complex and storage-intensive. We propose a novel soft clustering approach for OOD detection based on non-negative kernel regression. Our approach greatly reduces computational and space complexities (up to 11 x improvement in inference time and 87% reduction in storage requirements). It outperforms existing approaches by up to 4 AUROC points on four benchmarks. We also introduce an entropy-constrained version of our algorithm, leading to further reductions in storage requirements (up to 97% lower than comparable approaches) while retaining competitive performance. Our soft clustering approach for OOD detection highlights its potential for detecting tail-end phenomena in extreme-scale data settings. Our source code is available on Github.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Better Alignment with Instruction Back-and-Forth Translation

Thao Nguyen, Jeffrey Li, Sewoong Oh, Ludwig Schmidt, Jason E Weston, Luke Zettlemoyer, Xian Li

We propose a new method, instruction back-and-forth translation, to improve the quality of instruction-tuning data used for aligning large language models (LLMs). Given preprocessed texts from an initial web corpus (e.g., Dolma (Soldaini et al., 2024)), we generate synthetic instructions using the backtranslation approach proposed by Li et al., (2023), filter the generated data and rewrite the responses to improve their quality further based on the initial texts. Given similar quantities of instructions, fine-tuning Llama-2 on our (synthetic instruction, rewritten response) pairs yields better AlpacaEval win rates than using other common instruction datasets such as Humpback, ShareGPT, OpenOrca, Alpaca-GPT4 and Self-instruct, at both 7B and 70B parameter scales. We also demonstrate that rewriting the responses with an LLM is different from direct distillation: the former process yields better win rate at 70B scale, and the two text distributions exhibit significant distinction in the embedding space. Besides, we provide analyses showing that our backtranslated instructions are of higher quality than other sources of synthetic instructions, while our responses are more diverse and complex than what can be obtained from distillation. Overall we find that instruction back-and-forth translation combines the best of both worlds—making use of the information diversity and quantity found on the web, while ensuring the quality of the responses which is necessary for effective alignment.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Attribute Controlled Fine-tuning for Large Language Models: A Case Study on Detoxification

Tao Meng, Nisareh Mehrabi, Palash Goyal, Anil Ramakrishna, Aram Galstyan, Richard Zemel, Kai-Wei Chang, Rahul Gupta, Charith Peris We propose a constraint learning schema for fine-tuning Large Language Models (LLMs) with attribute control. Given a training corporus and control criteria formulated as a sequence-level constraint on model outputs, our method fine-tunes the LLM on the training corpus while enhancing constraint satisfaction with minimal impact on its utility and generation quality. Specifically, our approach regularizes the LLM training by penalizing the KL divergence between the desired output distribution, which satisfies the constraints, and the LLM's posterior. This regularization term can be approximated by an auxiliary model trained to decompose sequence-level constraints into token-level guidance, allowing the term to be measured by a closed-form formulation. To further improve efficiency, we design a parallel scheme for concurrently updating both the LLM and the auxiliary model. We evaluate the empirical performance of our approach by controlling rhetotoxicity when training an LLM. We show that our approach leads to an LLM that produces fewer inappropriate responses while achieving competitive performance on benchmarks and toxicity detection task

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

A Study of Parameter Efficient Fine-tuning by Learning to Efficiently Fine-Tune

Taha Ceritli, Savas Ozkan, Jeongwon Min, Eunchung Noh, Cho Jung Min, Mete Ozay

The growing size of large language models (LLMs) requires parameter-efficient fine-tuning (PEFT) methods for their adaptation to new tasks. Existing methods, such as Low-Rank Adaptation (LoRA), typically involve model adaptation by training the PEFT parameters. One open problem required to be solved to effectively employ these methods is the identification of PEFT parameters. More precisely, related works identify PEFT parameters by projecting high dimensional parameters of LLMs onto low dimensional parameter manifolds with predefined projections, or identifying PEFT parameters as projections themselves. To study this problem, we propose a new approach called Learning to Efficiently Fine-tune (LEFT) where we aim to learn spaces of PEFT parameters from data. In order to learn how to generate the PEFT parameters on a learned parameter space while fine-tuning the LLMs, we propose the Parameter Generation (PG) method. In the experimental analyses, we examine the effectiveness of our solutions exploring accuracy of fine-tuned LLMs and characteristics of PEFT parameters on benchmark GLUE tasks.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Navigating Noisy Feedback: Enhancing Reinforcement Learning with Error-Prone Language Models

Muhan Lin, Shuyang Shi, Yue Guo, Behdad Chalaki, Vaishnav Tadiparthi, Ehsan Moradi Pari, Simon Stepputis, Joseph Campbell, Katia P. Sycara

The correct specification of reward models is a well-known challenge in reinforcement learning. Hand-crafted reward functions often lead to inefficient or suboptimal policies and may not be aligned with user values. Reinforcement learning from human feedback is a successful technique that can mitigate such issues, however, the collection of human feedback can be laborious. Recent works have solicited feedback from pre-trained large language models rather than humans to reduce or eliminate human effort, however, these approaches yield poor performance in the presence of hallucination and other errors. This paper studies the advantages and limitations of reinforcement learning from large language model feedback and proposes a simple yet effective method for soliciting and applying feedback as a potential-based shaping function. We theoretically show that inconsistent rankings which approximate ranking errors lead to uninformative rewards with our approach. Our method empirically improves convergence speed and policy returns over commonly used baselines even with significant ranking errors, and eliminates the need for complex post-processing of reward functions.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Gradient Localization Improves Lifelong Pretraining of Language Models

Jared Fernández, Yonatan Bisk, Emma Strubell

Large Language Models (LLMs) trained on web-scale text corpora have been shown to capture world knowledge in their parameters. However, the mechanism by which language models store different types of knowledge is poorly understood. In this work, we examine two types of knowledge relating to temporally sensitive entities and demonstrate that each type is localized to different sets of parameters within the LLMs. We hypothesize that the lack of consideration of the locality of knowledge in existing continual learning methods contributes to both: the failed uptake of new information, and catastrophic forgetting of previously learned information. We observe that sequences containing references to updated and newly mentioned entities exhibit larger gradient norms in a subset of layers. We demonstrate that targeting parameter updates to these relevant layers can improve the performance of continually pretraining on language containing temporal drift.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Hop, skip, jump to Convergence: Dynamics of Learning Rate Transitions for Improved Training of Large Language Models

Vignesh Ganapathiraman, Shreyas Subramanian, Corey D Barrett

Various types of learning rate (LR) schedulers are being used for training or fine tuning of Large Language Models today. In practice, several mid-flight changes are required in the LR schedule either manually, or with careful choices around warmup steps, peak LR, type of decay and restarts. To study this further, we consider the effect of switching the learning rate at a predetermined time during training, which we refer to as "SkipLR". We model SGD as a stochastic gradient flow and show that when starting from the same initial parameters, switching the learning rate causes the loss curves to contract towards each other. We demonstrate this theoretically for some simple cases, and empirically on large language models. Our analysis provides insight into how learning rate schedules affect the training dynamics, and could inform the design of new schedules to accelerate convergence.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Conditioned Language Policy: A General Framework For Steerable Multi-Objective Finetuning

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andreia Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Kumar Avinava Dubey, Alexandre Rame, Johan Ferret, Geoffrey Cidener, Le Hou, Hongkun Yu, Amr Ahmed, Aranyak Mehta, Leonard Hussinet, Olivier Bachem, Edouard Leurent

Reward-based finetuning is crucial for aligning language policies with intended behaviors (*e.g.* creativity and safety). A key challenge is to develop steerable language models that trade-off multiple (conflicting) objectives in a flexible and efficient manner. This paper presents Conditional Language Policy (CLP), a general framework for finetuning language models on multiple objectives. Building on techniques from multi-task training and parameter-efficient finetuning, CLP learn steerable models that effectively trade-off conflicting objectives at *inference time*. Notably, this does not require training or maintaining multiple models to achieve different trade-offs between the objectives. Through extensive experiments and ablations on two summarization datasets, we show that CLP learns steerable language models that outperform and Pareto-dominate the existing approaches for multi-objective

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Data Diversity Matters for Robust Instruction Tuning

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, Haoming Jiang

Recent works have shown that by curating high quality and diverse instruction tuning datasets, we can significantly improve instruction-following capabilities. However, creating such datasets is difficult and most works rely on manual curation or proprietary language models. Automatic data curation is difficult as it is still not clear how we can define diversity for instruction tuning, how diversity and quality depend on one other, and how we can optimize dataset quality and diversity. To resolve these issue, we propose a new algorithm, Quality-Diversity Instruction Tuning (QDIT). QDIT provides a simple method to simultaneously control dataset diversity and quality, allowing us to conduct an in-depth study on the effect of diversity and quality on instruction tuning performance. From this study we draw two key insights (1) there is a natural tradeoff between data diversity and quality and (2) increasing data diversity significantly improves the worst case instruction following performance, therefore improving robustness. We validate the performance of QDIT on several large scale instruction tuning datasets, where we find it can substantially improve worst and average case performance compared to quality-driven data selection.

Resources and Evaluation 5

Nov 14 (Thu) 10:30-12:00 - Room: Riverfront Hall

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

AXCEL: Automated eXplainable Consistency Evaluation using LLMs

P Aditya Sreekar, Sahil Verma, Suransh Chopra, Abhishek Persad, Sarik Ghazarian, Narayanan Sadagopan

Large Language Models (LLMs) are widely used in both industry and academia for various tasks, yet evaluating the consistency of generated text responses continues to be a challenge. Traditional metrics like ROUGE and BLEU show a weak correlation with human judgment. More sophisticated metrics using Natural Language Inference (NLI) have shown improved correlations but are complex to implement, require domain-specific training due to poor cross-domain generalization, and lack explainability. More recently, prompt-based metrics using LLMs as evaluators have emerged; while they are easier to implement, they still lack explainability and depend on task-specific prompts, which limits their generalizability. This work introduces Automated eXplainable Consistency Evaluation using LLMs (AXCEL), a prompt-based consistency metric which offers explanations for the consistency scores by providing detailed reasoning and pinpointing inconsistent text spans. AXCEL is also a generalizable metric which can be adopted to multiple tasks without changing the prompt. AXCEL outperforms both non-prompt and prompt-based state-of-the-art (SOTA) metrics in detecting inconsistencies across summarization by 8.7%, free text generation by 6.2%, and data-to-text conversion tasks by 29.4%. We also evaluate the influence of underlying LLMs on prompt based metric performance and recalibrate the SOTA prompt-based metrics with the latest LLMs for fair comparison. Further, we show that AXCEL demonstrates strong performance using open source LLMs.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Developing a Pragmatic Benchmark for Assessing Korean Legal Language Understanding in Large Language Models

Kimyeoun, Choi Youngrok, Eunkyoung Choi, JinHwan Choi, Hai Jin Park, Wonseok Hwang

Large language models (LLMs) have demonstrated remarkable performance in the legal domain, with GPT-4 even passing the Uniform Bar Exam in the U.S. However their efficacy remains limited for non-standardized tasks and tasks in languages other than English. This underscores the need for careful evaluation of LLMs within each legal system before application. Here, we introduce KBL, a benchmark for assessing the Korean legal language understanding of LLMs, consisting of (1) 7 legal knowledge tasks (510 examples), (2) 4 legal reasoning tasks (288 examples), and (3) the Korean bar exam (4 domains, 53 tasks, 2,510 examples). First two datasets were developed in close collaboration with lawyers to evaluate LLMs in practical scenarios in a certified manner. Furthermore, considering legal practitioners' frequent use of extensive legal documents for research, we assess LLMs in both a closed book setting, where they rely solely on internal knowledge, and a retrieval-augmented generation (RAG) setting, using a corpus of Korean statutes and precedents. The results indicate substantial room and opportunities for improvement.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation

Juhwan Choi, JungMin Yun, Kyohooin Jin, YoungBin Kim

The quality of the dataset is crucial for ensuring optimal performance and reliability of downstream task models. However, datasets often contain noisy data inadvertently included during the construction process. Numerous attempts have been made to correct this issue through human annotators. However, hiring and managing human annotators is expensive and time-consuming. As an alternative, recent studies are exploring the use of large language models (LLMs) for data annotation. In this study, we present a case study that extends the application of LLM-based data annotation to enhance the quality of existing datasets through a cleansing strategy. Specifically, we leverage approaches such as chain-of-thought and majority voting to imitate human annotation and classify unrelated documents from the Multi-News dataset, which is widely used for the multi-document summarization task. Through our proposed cleansing method, we introduce an enhanced Multi-News+. By employing LLMs for data cleansing, we demonstrate an efficient and effective approach to improving dataset quality without relying on expensive human annotation efforts.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection

Zekun Li, Baolin Peng, Pengcheng He, Xifeng Yan

Large Language Models (LLMs) have demonstrated exceptional proficiency in instruction-following, making them increasingly integral to various applications. However, this capability introduces the risk of prompt injection attacks, where malicious instructions are embedded in the input to trigger unintended actions or content. Understanding the robustness of LLMs against such attacks is critical for ensuring their safe deployment. In this work, we establish a benchmark to evaluate the robustness of instruction-following LLMs against prompt injection attacks, assessing their ability to discern which instructions to follow and which to disregard. Through extensive experiments with leading instruction-following LLMs, we reveal significant vulnerabilities, particularly in models that mis-follow injected instructions. Our results show that certain models are excessively inclined to prioritize embedded instructions in prompts, often focusing on the latter parts of the prompt without fully understanding the overall context. Conversely, models that exhibit stronger contextual understanding and instruction-following capabilities tend to be more easily compromised by injected instructions. These findings highlight the need to balance improving LLMs' instruction-following abilities with enhancing their overall comprehension of prompts, to prevent mis-following inappropriate instructions. We hope our analysis provides valuable insights into these vulnerabilities, contributing to the development of more robust solutions in the future.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

FLIRT: Feedback Loop In-context Red Teaming

Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, Rahul Gupta

Warning: this paper contains content that may be inappropriate or offensive. As generative models become available for public use in various applications, testing and analyzing vulnerabilities of these models has become a priority. In this work, we propose an automatic red teaming framework that evaluates a given black-box model and exposes its vulnerabilities against unsafe and inappropriate content generation. Our framework uses in-context learning in a feedback loop to red team models and trigger them into unsafe content generation. In particular, taking text-to-image models as target models, we explore different feedback mechanisms to automatically learn effective and diverse adversarial prompts. Our experiments demonstrate that even with enhanced safety features, Stable Diffusion (SD) models are vulnerable to our adversarial prompts, raising concerns on their robustness in practical uses. Furthermore, we demonstrate that the proposed framework is effective for red teaming text-to-text models.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Leveraging Conflicts in Social Media Posts: Unintended Offense Dataset

Che Wei Tsai, Yen-Hao Huang, Tsu-keng Liao, Didier Fernando Salazar Estrada, Retnani Latifah, Yi-Shin Chen

In multi-person communications, conflicts often arise. Each individual may have their own perspective, which can differ. Additionally, commonly referenced offensive datasets frequently neglect contextual information and are primarily constructed with a focus on intended offenses. This study suggests that conflicts are pivotal in revealing a broader range of human interactions, including instances of unintended offensive language. This paper proposes a conflict-based data collection method to utilize inter-conflict cues in multi-person communications.

By focusing on specific cue posts within conversation threads, our proposed approach effectively identifies relevant instances for analysis. Detailed analyses are provided to showcase the proposed approach efficiently gathers data on subtly offensive content. The experimental results indicate that incorporating elements of conflict into data collection significantly enhances the comprehensiveness and accuracy of detecting offensive language but also enriches our understanding of conflict dynamics in digital communication.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Forgetting Curve: A Reliable Method for Evaluating Memorization Capability for Long-Context Models

Xinyu Liu, Runsong Zhao, Pengcheng Huang, Chunyang Xiao, Bei Li, Jingang Wang, Tong Xiao, JingBo Zhu

Numerous recent works target to extend effective context length for language models and various methods, tasks and benchmarks exist to measure model's effective memory length. However, through thorough investigations, we find limitations for currently existing evaluations on model's memory. We provide an extensive survey for limitations in this work and propose a new method called forgetting curve to measure the memorization capability of long-context models. We show that forgetting curve has the advantage of being robust to the tested corpus and the experimental settings, of not relying on prompt and can be applied to any model size. We apply our forgetting curve to a large variety of models involving both transformer and RNN/SSM based architectures. Our measurement provides empirical evidence for the effectiveness of transformer extension techniques while raises questions for the effective length of RNN/SSM based models. We also examine the difference between our measurement and existing benchmarks as well as popular metrics for various models.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Data Advisor: Data Curation with Fore sight for Safety Alignment of Large Language Models

Fei Wang, Nihareh Mehrabi, Palash Goyal, Rahul Gupta, Kai-Wei Chang, Aram Galstyan

Data are crucial element in large language model (LLM) alignment. Recent studies have explored using LLMs for efficient data collection. However, LLM-generated data often suffers from quality issues, with underrepresented or absent aspects and low-quality datapoints. To address these problems, we propose Data Advisor, an enhanced LLM-based method for generating data that takes into account the characteristics of the desired dataset. Starting from a set of pre-defined principles in hand, Data Advisor monitors the status of the generated data, identifies weaknesses in the current dataset, and advises the next iteration of data generation accordingly. Data Advisor can be easily integrated into existing data generation methods to enhance data quality and coverage. Experiments on safety alignment of three representative LLMs (i.e., Mistral, Llama2, and Falcon) demonstrate the effectiveness of Data Advisor in enhancing model safety against various fine-grained safety issues without sacrificing model utility.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Detecting Errors through Ensembling Prompts (DEEP): An End-to-End LLM Framework for Detecting Factual Errors

Alex Chandler, Devesh Surve, Hui Su

Accurate text summarization is one of the most common and important tasks performed by Large Language Models, where the costs of human review for an entire document may be high, but the costs of errors in summarization may be even greater. We propose Detecting Errors through Ensembling Prompts (DEEP) - an end-to-end large language model framework for detecting factual errors in text summarization. Our framework uses a diverse set of LLM prompts to identify factual inconsistencies, treating their outputs as binary features, which are then fed into ensembling models. We then calibrate the ensembled models to produce empirically accurate probabilities that a text is factually consistent or free of hallucination. We demonstrate that prior models for detecting factual errors in summaries perform significantly worse without optimizing the thresholds on subsets of the evaluated dataset. Our framework achieves state-of-the-art (SOTA) balanced accuracy on the AggreFact-XSUM FTSOTA, TofuEval Summary-Level, and HaluEval Summarization benchmarks in detecting factual errors within transformer-generated text summaries. It does so without any fine-tuning of the language model or reliance on thresholding techniques not available in practical settings.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

FAC²E: Better Understanding Large Language Model Capabilities by Dissociating Language and Cognition

Xiaoguang Wang, Lingfei Wu, Tengfei Ma, Bang Liu

Large language models (LLMs) are primarily evaluated by overall performance on various text understanding and generation tasks. However, such a paradigm fails to comprehensively differentiate the fine-grained language and cognitive skills, rendering the lack of sufficient interpretation to LLMs' capabilities. In this paper, we present FAC²E, a framework for Fine-grained and Cognition-grounded LLMs' Capability Evaluation. Specifically, we formulate LLMs' evaluation in a multi-dimensional and explainable manner by dissociating the language-related capabilities and the cognition-related ones. Besides, through extracting the intermediate reasoning from LLMs, we further break down the process of applying a specific capability into three sub-steps: recalling relevant knowledge, utilizing knowledge, and solving problems. Finally, FAC²E evaluates each sub-step of each fine-grained capability, providing a two-faced diagnosis for LLMs. Utilizing FAC²E, we identify a common shortfall in knowledge utilization among models and propose a straightforward, knowledge-enhanced method to mitigate this issue. Our results not only showcase promising performance enhancements but also highlight a direction for future LLM advancements.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Rethinking the Evaluation of In-Context Learning for LLMs

Guoixin Yu, Lemao Liu, Mo Yu, Yue Yu, Xiang Ao

In-context learning (ICL) has demonstrated excellent performance across various downstream NLP tasks, especially when synergized with powerful large language models (LLMs). Existing studies evaluate ICL methods primarily based on downstream task performance. This evaluation protocol overlooks the significant cost associated with the demonstration configuration process, i.e., tuning the demonstration as the ICL prompt. However, in this work, we point out that the evaluation protocol leads to unfair comparisons and potentially biased evaluation, because we surprisingly find the correlation between the configuration costs and task performance. Then we call for a two-dimensional evaluation paradigm that considers both of these aspects, facilitating a fairer comparison. Finally, based on our empirical finding that the optimized demonstration on one language model generalizes across language models of different sizes, we introduce a simple yet efficient strategy that can be applied to any ICL method as a plugin, yielding a better trade-off between the two dimensions according to the proposed evaluation paradigm.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Sequential API Function Calling Using GraphQL Schema

Avirup Saha, Lakshmi Mandal, Balaji Ganeshan, Sambit Ghosh, Renuka Sindhwatta, Carlos Eberhardt, Dan Debrunner, Sameep Mehta

Function calling using Large Language Models (LLMs) is an active research area that aims to empower LLMs with the ability to execute APIs to perform real-world tasks. However, sequential function calling using LLMs with interdependence between functions is still under-explored. To this end, we introduce GraphQlRestBench, a dataset consisting of natural language utterances paired with function call sequences representing real-world REST API calls with variable mapping between functions. In order to represent the response structure of the functions in the LLM prompt, we use the GraphQl schema of the REST APIs. We also introduce a custom evaluation framework for our dataset consisting of four specially designed metrics. We evaluate various open-source LLMs on our dataset using few-shot Chain-of-Thought and

ReAct prompting to establish a reasonable baseline.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, Kam-Fai Wong

Large language models (LLMs) are increasingly used for complex multi-turn conversations across diverse real-world applications. However, existing benchmarks mainly focus on single-turn evaluations, overlooking the models' capabilities in multi-turn interactions. To address this gap, we introduce a comprehensive benchmark to evaluate the multi-turn conversational abilities of LLMs. By analyzing human-LLM conversations, we categorize interaction patterns into four types: recollection, expansion, refinement, and follow-up. We construct multi-turn queries for each category either by augmenting existing datasets or creating new examples using GPT-4 with a human-in-the-loop process to avoid data leakage. To study the factors impacting multi-turn abilities, we create single-turn versions of the 1170 multi-turn queries and compare performance. Our evaluation of 10 well-known LLMs shows that while closed-source models generally surpass open-source ones, certain open-source models exceed GPT-3.5-Turbo in specific tasks. We observe significant performance degradation in multi-turn settings compared to single-turn settings in most models, which is not correlated with the models' fundamental capabilities. Moreover, we identify the distance to relevant content and susceptibility to error propagation as the key factors influencing multi-turn performance.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

CmdCaliper: A Semantic-Aware Command-Line Embedding Model and Dataset for Security Research

Sian-Yao Huang, Cheng-Lin Yang, Che-Yu Lin, Chun-Ying Huang

This research addresses command-line embedding in cybersecurity, a field obstructed by the lack of comprehensive datasets due to privacy and regulation concerns. We propose the first dataset of similar command lines, named CyPER, for training and unbiased evaluation. The training set is generated using a set of large language models (LLMs) comprising 28,520 similar command-line pairs. Our testing dataset consists of 2,807 similar command-line pairs sourced from authentic command-line data. In addition, we propose a command-line embedding model named CmdCaliper, enabling the computation of semantic similarity with command lines. Performance evaluations demonstrate that the smallest version of CmdCaliper (30 million parameters) suppresses state-of-the-art (SOTA) sentence embedding models with ten times more parameters across various tasks (e.g., malicious command-line detection and similar command-line retrieval). Our study explores the feasibility of data generation using LLMs in the cybersecurity domain. Furthermore, we release our proposed command-line dataset, embedding models weight and all program codes to the public. This advancement paves the way for more effective command-line embedding for future researchers.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Multi-LogiEval: Towards Evaluating Multi-Step Logical Reasoning Ability of Large Language Models

Nisarg Patel, Mohit Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, Chitta Baral

As Large Language Models (LLMs) continue to exhibit remarkable performance in natural language understanding tasks, there is a crucial need to measure their ability for human-like multi-step logical reasoning. Existing logical reasoning evaluation benchmarks often focus primarily on simplistic single-step or multi-step reasoning with a limited set of inference rules. Furthermore, the lack of datasets for evaluating non-monotonic reasoning represents a crucial gap since it aligns more closely with human-like reasoning. To address these limitations, we propose Multi-LogiEval, a comprehensive evaluation dataset encompassing multi-step logical reasoning with various inference rules and depths. Multi-LogiEval covers three logic types propositional, first-order, and non-monotonic consisting of more than 30 inference rules and more than 60 of their combinations with various depths. Leveraging this dataset, we conduct evaluations on a range of LLMs such as GPT-4, ChatGPT, Gemini-Pro, Orca, and Mistral, employing a zero-shot chain-of-thought. Experimental results show that there is a significant drop in the performance of LLMs as the reasoning steps/depth increases (average accuracy of 68% at depth-1 to 43% at depth-5). We further conduct a thorough investigation of reasoning chains generated by LLMs which reveals several important findings. We believe that Multi-LogiEval facilitates future research for evaluating and enhancing the logical reasoning ability of LLMs.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

Youssef Mohamed, Ranjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, Mohamed Elhoseiny

Research in vision and language has made considerable progress thanks to benchmarks such as COCO. COCO captions focused on unambiguous facts in English; ArtEmiss introduced subjective emotions and ArtELingo introduced some multilinguality (Chinese and Arabic). However we believe there should be more multilinguality. Hence, we present ArtELingo-28, a vision-language benchmark that spans 28 languages and encompasses approximately 200,000 annotations (140 annotations per image). Traditionally, vision research focused on unambiguous class labels, whereas ArtELingo-28 emphasizes diversity of opinions over languages and cultures. The challenge is to build machine learning systems that assign emotional captions to images. Baseline results will be presented for three novel conditions: Zero-Shot, Few-Shot and One-vs-All Zero-Shot. We find that cross-lingual transfer is more successful for culturally-related languages. Data and code will be made publicly available.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Assessing and Verifying Task Utility in LLM-Powered Applications

Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Hassan Awadallah, Charles L. A. Clarke, Julia Kiseleva

The rapid development of Large Language Models (LLMs) has led to a surge in applications that facilitate collaboration among multiple agents, assisting humans in their daily tasks. However, a significant gap remains in assessing to what extent LLM-powered applications genuinely enhance user experience and task execution efficiency. This highlights the need to verify utility of LLM-powered applications, particularly by ensuring alignment between the application's functionality and end-user needs. We introduce AgentEval, a novel framework designed to simplify the utility verification process by automatically proposing a set of criteria tailored to the unique purpose of any given application. This allows for a comprehensive assessment, quantifying the utility of an application against the suggested criteria. We present a comprehensive analysis of the effectiveness and robustness of AgentEval for two open source datasets including Math Problem solving and ALFWORLD House-hold related tasks. For reproducibility purposes, we make the data, code and all the logs publicly available at <https://github.com/Narabzad/AgentEval>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

CELLO: Causal Evaluation of Large Vision-Language Models

Meiqi Chen, Bo Peng, Yan Zhang, Chaochao Lu

Causal reasoning is fundamental to human intelligence and crucial for effective decision-making in real-world environments. Despite recent advancements in large vision-language models (LVLMs), their ability to comprehend causality remains unclear. Previous work typically focuses on commonsense causality between events and/or actions, which is insufficient for applications like embodied agents and lacks the explicitly defined causal graphs required for formal causal reasoning. To overcome these limitations, we introduce a fine-grained and unified definition of causality involving interactions between humans and/or objects. Building on the definition, we construct a novel dataset, CELLO, consisting of 14,094 causal questions across all four levels of causality: discovery, association, intervention, and counterfactual. This dataset

surpasses traditional commonsense causality by including explicit causal graphs that detail the interactions between humans and objects. Extensive experiments on CELLO reveal that current LVMs still struggle with causal reasoning tasks, but they can benefit significantly from our proposed CELLO-CoT, a causally inspired chain-of-thought prompting strategy. Both quantitative and qualitative analyses from this study provide valuable insights for future research. Our project page is at <https://github.com/OpenCausalLab/CELLO>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Adversarial Math Word Problem Generation

Roy Xie, Chengxuan Huang, Junlin Wang, Bhuvan Dhingra

Large language models (LLMs) have significantly transformed the educational landscape. As current plagiarism detection tools struggle to keep pace with LLMs' rapid advancements, the educational community faces the challenge of assessing students' true problem-solving abilities in the presence of LLMs. In this work, we explore a new paradigm for ensuring fair evaluation—generating adversarial examples which preserve the structure and difficulty of the original questions aimed for assessment, but are unsolvable by LLMs. Focusing on the domain of math word problems, we leverage abstract syntax trees to structurally generate adversarial examples that cause LLMs to produce incorrect answers by simply editing the numeric values in the problems. We conduct experiments on various open- and closed-source LLMs, quantitatively and qualitatively demonstrating that our method significantly degrades their math problem-solving ability. We identify shared vulnerabilities among LLMs and propose a cost-effective approach to attack high-cost models. Additionally, we conduct automatic analysis to investigate the cause of failure, providing further insights into the limitations of LLMs.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

I Never Said That: A dataset, taxonomy and baselines on response clarity classification

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, Giorgos Stamou

Equivocation and ambiguity in public speech are well-studied discourse phenomena, especially in political science and analysis of political interviews. Inspired by the well-grounded theory on equivocation, we aim to resolve the closely related problem of response clarity in questions extracted from political interviews, leveraging the capabilities of Large Language Models (LLMs) and human expertise. To this end, we introduce a novel taxonomy that frames the task of detecting and classifying response clarity and a corresponding clarity classification dataset which consists of question-answer (QA) pairs drawn from political interviews and annotated accordingly. Our proposed two-level taxonomy addresses the clarity of a response in terms of the information provided for a given question (high-level) and also provides a fine-grained taxonomy of evasion techniques that relate to unclear, ambiguous responses (lower-level). We combine ChatGPT and human annotators to collect, validate and annotate discrete QA pairs from political interviews, to be used for our newly introduced response clarity task. We provide a detailed analysis and conduct several experiments with different model architectures, sizes and adaptation methods to gain insights and establish new baselines over the proposed dataset and task.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Evaluation of Question Answer Generation for Portuguese: Insights and Datasets

Felipe Paula, CÁSSIANA ROBERTA LIZZONI MICHELIN, Viviane Moreira

Automatic question generation is an increasingly important task that can be applied in different settings, including educational purposes, data augmentation for question-answering (QA), and conversational systems. More specifically, we focus on question answer generation (QAG), which produces question-answer pairs given an input context. We adapt and apply QAG approaches to generate question-answer pairs for different domains and assess their capacity to generate accurate, diverse, and abundant question-answer pairs. Our analyses combine both qualitative and quantitative evaluations that allow insights into the quality and types of errors made by QAG methods. We also look into strategies for error filtering and their effects. Our work concentrates on Portuguese, a widely spoken language that is underrepresented in natural language processing research. To address the pressing need for resources, we generate and make available human-curated extractive QA datasets in three diverse domains.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

From Generation to Selection Findings of Converting Analogical Problem-Solving into Multiple-Choice Questions

Donghyeon Shin, Seungpil Lee, Klea Lena Kovacec, Sundong Kim

As artificial intelligence reasoning abilities gain prominence, generating reliable benchmarks becomes crucial. The Abstract and Reasoning Corpus (ARC) offers challenging problems yet unsolved by AI. While ARC effectively assesses reasoning, its generation-based evaluation overlooks other assessment aspects. Bloom's Taxonomy suggests evaluating six cognitive stages: Remember, Understand, Apply, Analyze, Evaluate, and Create. To extend ARC's focus beyond the *Create* stage, we developed MC-LARC, a multiple-choice format suitable for assessing stages like Understand and Apply in Large Language Models (LLMs). Our evaluation of ChatGPT4V's analogical reasoning using MC-LARC confirmed that this format supports LLMs' reasoning capabilities and facilitates evidence analysis. However, we observed LLMs using shortcuts in MC-LARC tasks. To address this, we propose a self-feedback framework where LLMs identify issues and generate improved options. MC-LARC is available at <https://mc-larc.github.io/>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

CERD: A Comprehensive Chinese Rhetoric Dataset for Rhetorical Understanding and Generation in Essays

Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, Yan Xia

Existing rhetoric understanding and generation datasets or corpora primarily focus on single coarse-grained categories or fine-grained categories, neglecting the common interrelations between different rhetorical devices by treating them as independent sub-tasks. In this paper, we propose the Chinese Essay Rhetoric Dataset (CERD), consisting of 4 commonly used coarse-grained categories including metaphor, personification, hyperbole and parallelism and 23 fine-grained categories across both form and content levels. CERD is a manually annotated and comprehensive Chinese rhetoric dataset with five interrelated sub-tasks. Unlike previous work, our dataset aids in understanding various rhetorical devices, recognizing corresponding rhetorical components, and generating rhetorical sentences under given conditions, thereby improving the author's writing proficiency and language usage skills. Extensive experiments are conducted to demonstrate the interrelations between multiple tasks in CERD, as well as to establish a benchmark for future research on rhetoric. The experimental results indicate that Large Language Models achieve the best performance across most tasks, and jointly fine-tuning with multiple tasks further enhances performance.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting

Marco Naguib, Xavier Tannier, Aurélie Névéol

Large language models (LLMs) have become the preferred solution for many natural language processing tasks. In low-resource environments such as specialized domains, their few-shot capabilities are expected to deliver high performance. Named Entity Recognition (NER) is a critical task in information extraction that is not covered in recent LLM benchmarks. There is a need for better understanding the performance of LLMs for NER in a variety of settings including languages other than English. This study aims to evaluate generative LLMs, employed through prompt engineering, for few-shot clinical NER. We compare 13 auto-regressive models using prompting and 16 masked models using

fine-tuning on 14 NER datasets covering English, French and Spanish. While prompt-based auto-regressive models achieve competitive F1 for general NER, they are outperformed within the clinical domain by lighter biLSTM-CRF taggers based on masked models. Additionally, masked models exhibit lower environmental impact compared to auto-regressive models. Findings are consistent across the three languages studied, which suggests that LLM prompting is not yet suited for NER production in the clinical domain.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

TurkishMMLU: Measuring Massive Multitask Language Understanding in Turkish

Arda Yüksel, Abdullatif Koksal, Lütfi Kerem Senel, Anna Korhonen, Hinrich Schütze

Multiple choice question answering tasks evaluate the reasoning, comprehension, and mathematical abilities of Large Language Models (LLMs). While existing benchmarks employ automatic translation for multilingual evaluation, this approach is error-prone and potentially introduces culturally biased questions, especially in social sciences. We introduce the first multitask, multiple-choice Turkish QA benchmark, TurkishMMLU, to evaluate LLMs' understanding of the Turkish language. TurkishMMLU includes over 10,000 questions, covering 9 different subjects from Turkish high-school education curricula. These questions are written by curriculum experts, suitable for the high-school curricula in Turkey, covering subjects ranging from natural sciences and math questions to more culturally representative topics such as Turkish Literature and the history of the Turkish Republic. We evaluate over 20 LLMs, including multilingual open-source (e.g., Gemma, Llama, MT5), closed-source (GPT 4o, Claude, Gemini), and Turkish-adapted (e.g., Trendyol) models. We provide an extensive evaluation, including zero-shot and few-shot evaluation of LLMs, chain-of-thought reasoning, and question difficulty analysis along with model performance. We provide an in-depth analysis of the Turkish capabilities and limitations of current LLMs to provide insights for future LLMs for the Turkish language. We publicly release our code for the dataset and evaluation: <https://github.com/ArdaYueksel/TurkishMMLU>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Generalists vs. Specialists: Evaluating Large Language Models for Urdu

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, Awais Athar

In this paper, we compare general-purpose models, GPT-4-Turbo and Llama-3-8b, with special-purpose models XLM-Roberta-large, mT5-large, and Llama-3-8b that have been fine-tuned on specific tasks. We focus on seven classification and seven generation tasks to evaluate the performance of these models on Urdu language. Urdu has 70 million native speakers, yet it remains underrepresented in Natural Language Processing (NLP). Despite the frequent advancements in Large Language Models (LLMs), their performance in low-resource languages, including Urdu, still needs to be explored. We also conduct a human evaluation for the generation tasks and compare the results with the evaluations performed by GPT-4-Turbo, Llama-3-8b and Claude 3.5 Sonnet. We find that special-purpose models consistently outperform general-purpose models across various tasks. We also find that the evaluation done by GPT-4-Turbo for generation tasks aligns more closely with human evaluation compared to the evaluation done by Llama-3-8b. This paper contributes to the NLP community by providing insights into the effectiveness of general and specific-purpose LLMs for low-resource languages.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

MedINST: Meta Dataset of Biomedical Instructions

Wenhan Han, Meng Fang, Zihan Zhang, Yu Yin, Zirui Song, Ling Chen, Mykola Pechenizkiy, Qingyu Chen

The integration of large language model (LLM) techniques in the field of medical analysis has brought about significant advancements, yet the scarcity of large, diverse, and well-annotated datasets remains a major challenge. Medical data and tasks, which vary in format, size, and other parameters, require extensive preprocessing and standardization for effective use in training LLMs. To address these challenges, we introduce MedINST, the Meta Dataset of Biomedical Instructions, a novel multi-domain, multi-task instructional meta-dataset. MedINST comprises 133 biomedical NLP tasks and over 7 million training samples, making it the most comprehensive biomedical instruction dataset to date. Using MedINST as the meta dataset, we curate MedINST32, a challenging benchmark with different task difficulties aiming to evaluate LLMs' generalization ability. We fine-tune several LLMs on MedINST and evaluate on MedINST32, showcasing enhanced cross-task generalization.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Is GPT-4V (isom) All You Need for Automating Academic Data Visualization? Exploring Vision-Language Models' Capability in Reproducing Academic Charts

Zhehao Zhang, Weicheng Ma, Soroush Vosoughi

While effective data visualization is crucial to present complex information in academic research, its creation demands significant expertise in both data management and graphic design. We explore the potential of using Vision-Language Models (VLMs) in automating the creation of data visualizations by generating code templates from existing charts. As the first work to systematically investigate this task, we first introduce AcademiaChart, a dataset comprising 2525 high-resolution data visualization figures with captions from a variety of AI conferences, extracted directly from source codes. We then conduct large-scale experiments with six state-of-the-art (SOTA) VLMs, including both closed-source and open-source models. Our findings reveal that SOTA closed-source VLMs can indeed be helpful in reproducing charts. On the contrary, open-source ones are only effective at reproducing much simpler charts but struggle with more complex ones. Interestingly, the application of Chain-of-Thought (CoT) prompting significantly enhances the performance of the most advanced model, GPT-4-V, while it does not work as well for other models. These results underscore the potential of VLMs in data visualization while also highlighting critical areas that need improvement for broader application.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

The Mystery of Compositional Generalization in Graph-based Generative Commonsense Reasoning

Xiyun Fu, Anette Frank

While LLMs have emerged as performant architectures for reasoning tasks, their compositional generalization capabilities have been questioned. In this work, we introduce a Compositional Generalization Challenge for Graph-based Commonsense Reasoning (CGGC) that goes beyond previous evaluations that are based on sequences or tree structures and instead involves a reasoning graph: It requires models to generate a natural sentence based on given concepts and a corresponding reasoning graph, where the presented graph involves a previously unseen combination of relation types. To master this challenge, models need to learn how to reason over relation tuples within the graph, and how to compose them when conceptualizing a verbalization. We evaluate seven well-known LLMs using in-context learning and find that performant LLMs still struggle in compositional generalization. We investigate potential causes of this gap by analyzing the structures of reasoning graphs, and find that different structures present varying levels of difficulty for compositional generalization. Arranging the order of demonstrations according to the structures difficulty shows that organizing samples in an easy-to-hard schema enhances the compositional generalization ability of LLMs.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

PEDANTS (Precise Evaluations of Diverse Answer Nominee Text for Skinflints): Use Evaluation Metrics Wisely Efficient Evaluation Analysis and Benchmarking for Open-Domain Question Answering

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, Jordan Lee Boyd-Graber

Question answering (QA) can only make progress if we know if an answer is correct, but current answer correctness (AC) metrics struggle with verbose, free-form answers from large language models (LLMs). There are two challenges with current short-form QA evaluations: a

lack of diverse styles of evaluation data and an over-reliance on expensive and slow LLMs. LLM-based scorers correlate better with humans, but this expensive task has only been tested on limited QA datasets. We rectify these issues by providing rubrics and datasets for evaluating machine QA adopted from the Trivia community. We also propose an efficient, and interpretable QA evaluation that is more stable than an exact match and neural methods (BERTScore).

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation

Irene Zubiaaga, Aitor Soroa, Rodrigo Agerri

This paper proposes a novel approach to evaluate Counter Narrative (CN) generation using a Large Language Model (LLM) as an evaluator. We show that traditional automatic metrics correlate poorly with human judgements and fail to capture the nuanced relationship between generated CNs and human perception. To alleviate this, we introduce a model ranking pipeline based on pairwise comparisons of generated CNs from different models, organized in a tournament-style format. The proposed evaluation method achieves a high correlation with human preference, with a score of 0.88. As an additional contribution, we leverage LLMs as zero-shot CN generators and provide a comparative analysis of chat, instruct, and base models, exploring their respective strengths and limitations. Through meticulous evaluation, including fine-tuning experiments, we elucidate the differences in performance and responsiveness to domain-specific data. We conclude that chat-aligned models in zero-shot are the best option for carrying out the task, provided they do not refuse to generate an answer due to security concerns.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Can LLMs Reason in the Wild with Programs?

Yuan Yang, Siyeng Xiong, Ali Payani, Ehsan Shareghi, Faramarz Fekri

Large Language Models (LLMs) have shown superior capability to solve reasoning problems with programs. While being a promising direction, most of such frameworks are trained and evaluated in settings with a prior knowledge of task requirements. However, as LLMs become more capable, it is necessary to assess their reasoning abilities in more realistic scenarios where many real-world problems are open-ended with ambiguous scope, and often require multiple formalisms to solve. To investigate this, we introduce the task of reasoning in the wild, where an LLM is tasked to solve a reasoning problem of unknown type by identifying the sub-problems and their corresponding formalisms, and writing a program to solve each sub-problem, guided by a tactic. We create a large tactic-guided trajectory dataset containing detailed solutions to a diverse set of reasoning problems, ranging from well-defined single-form reasoning (e.g., math, logic), to ambiguous and hybrid ones (e.g., commonsense, combined math and logic). This allows us to test various aspects of LLMs' reasoning at the fine-grained level such as the selection and execution of tactics, and the tendency to take undesired shortcuts. In experiments, we highlight that existing LLMs fail significantly on problems with ambiguous and mixed scope, revealing critical limitations and overfitting issues (e.g., accuracy on GSM8K drops by at least 50%). We further show the potential of finetuning a local LLM on the tactic-guided trajectories in achieving better performance. Project repo is available at <https://github.com/gblackout/Reason-in-the-Wild>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Towards Robust Evaluation of Unlearning in LLMs via Data Transformations

Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, Ashutosh Modi

Large Language Models (LLMs) have shown to be a great success in a wide range of applications ranging from regular NLP-based use cases to AI agents. LLMs have been trained on a vast corpus of texts from various sources; despite the best efforts during the data pre-processing stage while training the LLMs, they may pick some undesirable information such as personally identifiable information (PII). Consequently, in recent times research in the area of Machine Unlearning (MUL) has become active, the main idea is to force LLMs to forget (unlearn) certain information (e.g., PII) without suffering from performance loss on regular tasks. In this work, we examine the robustness of the existing MUL techniques for their ability to enable leakage-proof forgetting in LLMs. In particular, we examine the effect of data transformation on the forgetting, i.e., is an unlearned LLM able to recall forgotten information if there is a change in the format of the input? Our findings on the TOFU dataset highlight the necessity of using diverse data formats to quantify unlearning in LLMs more reliably.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

PizzaCommonSense: A Dataset for Commonsense Reasoning about Intermediate Steps in Cooking Recipes

Aisatour Diallo, Antonis Bikakis, Luke Dickens, Anthony Hunter, Rob Miller

Understanding procedural texts, such as cooking recipes, is essential for enabling machines to follow instructions and reason about tasks, a key aspect of intelligent reasoning. In cooking, these instructions can be interpreted as a series of modifications to a food preparation. For a model to effectively reason about cooking recipes, it must accurately discern and understand the inputs and outputs of intermediate steps within the recipe. We present a new corpus of cooking recipes enriched with descriptions of intermediate steps that describe the input and output for each step. PizzaCommonsense serves as a benchmark for the reasoning capabilities of LLMs because it demands rigorous explicit input-output descriptions to demonstrate the acquisition of implicit commonsense knowledge, which is unlikely to be easily memorized. GPT-4 achieves only 26% human-evaluated preference for generations, leaving room for future improvements.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

"Knowing When You Don't Know": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation

Nandan Thakur, Luiz Bonifácio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, Jimmy Lin

Retrieval-Augmented Generation (RAG) grounds Large Language Model (LLM) output by leveraging external knowledge sources to reduce factual hallucinations. However, prior work lacks a comprehensive evaluation of different language families, making it challenging to evaluate LLM robustness against errors in an external retrieved knowledge. To overcome this, we establish **NoMIRACL**, a human-annotated dataset for evaluating LLM robustness in RAG across 18 typologically diverse languages. NoMIRACL includes both a non-relevant and a relevant subset. Queries in the non-relevant subset contain passages judged as non-relevant, whereas queries in the relevant subset include at least a single judged relevant passage. We measure relevance assessment using: (i) *hallucination rate*, measuring model tendency to hallucinate when the answer is not present in passages in the non-relevant subset, and (ii) *error rate*, measuring model inaccuracy to recognize relevant passages in the relevant subset. In our work, we observe that most models struggle to balance the two capacities. Models such as LLAMA-2 and Orca-2 achieve over 88% hallucination rate on the non-relevant subset. Mistral and LLAMA-3 hallucinate less but can achieve up to a 74.9% error rate on the relevant subset. Overall, GPT-4 is observed to provide the best tradeoff on both subsets, highlighting future work necessary to improve LLM robustness. NoMIRACL dataset and evaluation code are available at: <https://github.com/project-mirac/miracrl>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

What is the value of templates? Rethinking Document Information Extraction Datasets for LLMs

Ran Zmigrod, Pranav Shetty, Mathieu Sibue, Zhigiang Ma, Armineh Nourbakhsh, Xiaomo Liu, Manuela Veloso

The rise of large language models (LLMs) for visually rich document understanding (VRDU) has kindled a need for prompt-response, document-based datasets. As annotating new datasets from scratch is labor-intensive, the existing literature has generated prompt-response datasets from available resources using simple templates. For the case of key information extraction (KIE), one of the most common VRDU

tasks, past work has typically employed the template "What is the value for the key?". However, given the variety of questions encountered in the wild, simple and uniform templates are insufficient for creating robust models in research and industrial contexts. In this work, we present K2Q, a diverse collection of five datasets converted from KIE to a prompt-response format using a plethora of bespoke templates. The questions in K2Q can span multiple entities and be extractive or boolean. We empirically compare the performance of seven baseline generative models on K2Q with zero-shot prompting. We further compare three of these models when training on K2Q versus training on simpler templates to motivate the need of our work. We find that creating diverse and intricate KIE questions enhances the performance and robustness of VRDU models. We hope this work encourages future studies on data quality for generative model training.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

On Leakage of Code Generation Evaluation Datasets

Alexander McMahon, Tom Sherborne, Dennis Aumiller, Elena Tommasone, Milad Alizadeh, Jingyi He, Raymond Ma, Maxime Voisin, Ellen Gilsemen-McMahon, Matthias Gallé

In this paper, we consider contamination by code generation test sets, in particular in their use in modern large language models. We discuss three possible sources of such contamination and show findings supporting each of them: (i) direct data leakage, (ii) indirect data leakage through the use of synthetic data and (iii) overfitting to evaluation sets during model selection. To address this, we release Less Basic Python (LBPP): an uncontaminated new benchmark of 161 prompts with their associated Python solutions. LBPP is released at <https://huggingface.co/datasets/CoherenceForAI/lbpp>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

TOWER: Tree Organized Weighting for Evaluating Complex Instructions

Noah Ziems, Zhihan Zhang, Meng Jiang

Evaluating the ability of large language models (LLMs) to follow complex human-written instructions is essential for their deployment in real-world applications. While benchmarks like Chatbot Arena use human judges to assess model performance, they are resource-intensive and time-consuming. Alternative methods using LLMs as judges, such as AlpacaEval, MT Bench, WildBench, and InfoBench offer improvements but still do not capture that certain complex instruction aspects are more important than others to follow. To address this gap, we propose a novel evaluation metric, TOWER, that incorporates human-judged importance into the assessment of complex instruction following. We show that human annotators agree with tree-based representations of these complex instructions nearly as much as they agree with other human annotators. We release tree-based annotations of the InfoBench dataset and the corresponding evaluation code to facilitate future research.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

BLADE: Benchmarking Language Model Agents for Data-Driven Science

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A Merrill, Jeffrey Heer, Tim Althoff

Data-driven scientific discovery requires the iterative integration of scientific domain knowledge, statistical expertise, and an understanding of data semantics to make nuanced analytical decisions, e.g., about which variables, transformations, and statistical models to consider. LLM-based agents equipped with planning, memory, and code execution capabilities have the potential to support data-driven science. However, evaluating agents on such open-ended tasks is challenging due to multiple valid approaches, partially correct steps, and different ways to express the same decisions. To address these challenges, we present BLADE, a benchmark to automatically evaluate agents' multifaceted approaches to open-ended research questions. BLADE consists of 12 datasets and research questions drawn from existing scientific literature, with ground truth collected from independent analyses by expert data scientists and researchers. To automatically evaluate agent responses, we developed corresponding computational methods to match different representations of analyses to this ground truth. Though language models possess considerable world knowledge, our evaluation shows that they are often limited to basic analyses. However, agents capable of interacting with the underlying data demonstrate improved, but still non-optimal, diversity in their analytical decision making. Our work enables the evaluation of agents for data-driven science and provides researchers deeper insights into agents' analysis approaches.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

LEGOBench: Scientific Leaderboard Generation Benchmark

Shruti Singh, Shoab Alam, Husain Malwat, Mayank Singh

The ever-increasing volume of paper submissions makes it difficult to stay informed about the latest state-of-the-art research. To address this challenge, we introduce LEGObench, a benchmark for evaluating systems that generate scientific leaderboards. LEGObench is curated from 22 years of preprint submission data on arXiv and more than 11k machine learning leaderboards on the PapersWithCode portal. We present a language model-based and four graph-based leaderboard generation task configuration. We evaluate popular encoder-only scientific language models as well as decoder-only large language models across these task configurations. State-of-the-art models showcase significant performance gaps in automatic leaderboard generation on LEGObench. The code is available on GitHub and the dataset is hosted on OSF.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

CHAmBi: A New Benchmark on Chinese Ambiguity Challenges for Large Language Models

Qin Zhang, Sihan Cai, Jiaxu Zhao, Mykola Pechenizkiy, Meng Fang

Ambiguity is an inherent feature of language, whose management is crucial for effective communication and collaboration. This is particularly true for Chinese, a language with extensive lexical-morphemic ambiguity. Despite the wide use of large language models (LLMs) in numerous domains and their growing proficiency in Chinese, there is a notable lack of datasets to thoroughly evaluate LLMs' ability to handle ambiguity in Chinese. To bridge this gap, we introduce the CHAmBi dataset, a specialized Chinese multi-label disambiguation dataset formatted in Natural Language Inference. It comprises 4,991 pairs of premises and hypotheses, including 824 examples featuring a wide range of ambiguities. In addition to the dataset, we develop a series of tests and conduct an extensive evaluation of pre-trained LLMs' proficiency in identifying and resolving ambiguity in the Chinese language. Our findings reveal that GPT-4 consistently delivers commendable performance across various evaluative measures, albeit with limitations in robustness. The performances of other LLMs, however, demonstrate variability in handling ambiguity-related tasks, underscoring the complexity of such tasks in the context of Chinese. The overall results highlight the challenge of ambiguity handling for current LLMs and underscore the imperative need for further enhancement in LLM capabilities for effective ambiguity resolution in the Chinese language.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

One-to-Many Testing for Code Generation from (Just) Natural Language

Mansi Uniyal, Mukul Singh, Gust Verbrugge, Sumit Gulwani, Yu Le

MBPP is a popular dataset for evaluating the task of code generation from natural language. Despite its popularity, there are three problems: (1) it relies on providing test cases to generate the right signature, (2) there is poor alignment between instruction and evaluation test cases, and (3) contamination of the exact phrasing being present in training datasets. We adapt MBPP to emphasize on generating code from just natural language by (1) removing ambiguity about the semantics of the task from the descriptions, and (2) evaluating generated code on multiple sets of assertions to account for ambiguity in the syntax. We compare popular open and closed weight models on the original (MBPP) and adapted

(MBUPP) datasets.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Probing the Capacity of Language Model Agents to Operationalize Disparate Experiential Context Despite Distraction

Sonny George, Chris Sypherd, Dylan Cashman

Large language model (LLM) agents show promise in an increasing number of domains. In many proposed applications, it is expected that the agent reasons over accumulated experience presented in an input prompt. We propose the OEDD (Operationalize Experience Despite Distraction) corpus, a human-annotator-validated body of scenarios with pre-scripted agent histories where the agent must make a decision based on disparate experiential information in the presence of a distractor. We evaluate three state-of-the-art LLMs (GPT-3.5 Turbo, GPT-4o, and Gemini 1.5 Pro) using a minimal chain-of-thought prompting strategy and observe that when (1) the input context contains over 1,615 tokens of historical interactions, (2) a crucially decision-informing premise is the rightful conclusion over two disparate environment premises, and (3) a trivial, but distracting red herring fact follows, all LLMs perform worse than random choice at selecting the better of two actions. Our code and test corpus are publicly available at: github.com/sonnygeorge/OEDD.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

MalAlgQA: A Pedagogical Approach for Evaluating Counterfactual Reasoning Abilities of Large Language Models

Shashank Sonkar, Naiming Liu, MyCo Le, Richard Baranik

This paper introduces MalAlgQA, a novel dataset designed to evaluate the counterfactual reasoning capabilities of Large Language Models (LLMs) through a pedagogical approach. The dataset comprises mathematics and reading comprehension questions, each accompanied by four answer choices and their corresponding rationales. At the heart of MalAlgQA are "malgorithms" - rationales behind incorrect answer choices that represent flawed yet logically coherent reasoning paths. These malgorithms serve as counterfactual scenarios, allowing us to assess an LLM's ability to identify and analyze flawed reasoning patterns. We propose the Malgorithm Identification task, where LLMs are assessed based on their ability to identify corresponding malgorithm given an incorrect answer choice. To evaluate the model performance, we introduce two metrics: Algorithm Identification Accuracy (AIA) for correct answer rationale identification, and Malgorithm Identification Accuracy (MIA) for incorrect answer rationale identification. Our experiments reveal that state-of-the-art LLMs exhibit significant performance drops in MIA compared to AIA, highlighting the challenges in counterfactual reasoning. Surprisingly, we find that the chain-of-thought prompting technique not only fails to consistently enhance MIA but can sometimes lead to underperformance compared to simple prompting. These findings have important implications for developing LLMs with improved counterfactual reasoning, particularly relevant for AI-powered tutoring systems, where identifying and addressing student misconceptions is essential. MalAlgQA dataset is available here.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Gazelle: An Instruction Dataset for Arabic Writing Assistance

Samar Mohamed Magdy, Fahrraddin Alwajih, Sang Yun Kwon, Reem Abdel-Salam, Muhammad Abdul-Mageed

Writing has long been considered a hallmark of human intelligence and remains a pinnacle task for artificial intelligence (AI) due to the intricate cognitive processes involved. Recently, rapid advancements in generative AI, particularly through the development of Large Language Models (LLMs), have significantly transformed the landscape of writing assistance. However, underrepresented languages like Arabic encounter significant challenges in the development of advanced AI writing tools, largely due to the limited availability of data. This scarcity constrains the training of effective models, impeding the creation of sophisticated writing assistance technologies. To address these issues, we present **Gazelle**^{*}, a comprehensive dataset for Arabic writing assistance. In addition, we offer an evaluation framework designed to enhance Arabic writing assistance tools. Our human evaluation of leading LLMs, including GPT-**4***, GPT-**4o***, Cohere Command R+, and Gemini **1.5** Pro, highlights their respective strengths and limitations in addressing the challenges of Arabic writing. Our findings underscore the need for continuous model training and dataset enrichment to manage the complexities of Arabic language processing, paving the way for more effective AI-powered Arabic writing tools.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

P-FOLIO: Evaluating and Improving Logical Reasoning with Abundant Human-Written Reasoning Chains

SIMENG HAN, Aaron Yu, Rui Shen, Zhenxing Qi, Martin Riddell, Wanfei Zhou, Yujie Qiao, Yilun Zhao, Semih Yavuz, Ye Liu, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, Dragomir Radev

Existing methods on understanding the capabilities of LLMs in logical reasoning rely on binary entailment classification or synthetically derived rationales, which are not sufficient for properly assessing model's capabilities. We present **P-FOLIO**, a human-annotated dataset consisting of diverse and complex reasoning chains for a set of realistic logical reasoning stories also written by humans. P-FOLIO is collected with an annotation protocol that facilitates humans to annotate well-structured natural language proofs for first-order logic reasoning problems in step-by-step manner. The number of reasoning steps in P-FOLIO span from 0 to 20. We further use P-FOLIO to evaluate and improve large-language-model (LLM) reasoning capabilities. We evaluate LLM reasoning capabilities at a fine granularity via single-step inference rule classification, with more diverse inference rules of more diverse and higher levels of complexities than previous works. Given that a single model-generated reasoning chain could take a completely different path than the human-annotated one, we sample multiple reasoning chains from a model and use F1_{eR} metrics for evaluating the quality of model-generated reasoning chains. We show that human-written reasoning chains significantly boost the logical reasoning capabilities of LLMs via many-shot prompting and fine-tuning. Furthermore, fine-tuning Llam3-7B on P-FOLIO improves the model performance by 10% or more on three other out-of-domain logical reasoning datasets.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Coping with Emotion Coping: A Corpus to Model Emotions in Text Based on Role Playing

Enrica Troiano, Sofie Labat, Marco Antonio Straniaci, Rossana Damiano, Viviana Patti, Roman Klinger

There is a mismatch between psychological and computational studies on emotions. Psychological research aims at explaining and documenting internal mechanisms of these phenomena, while computational work often simplifies them into labels. Many emotion fundamentals remain under-explored in natural language processing, particularly how emotions develop and how people cope with them. To help reduce this gap, we follow theories on coping, and treat emotions as strategies to cope with salient situations (i.e., how people deal with emotion-eliciting events). This approach allows us to investigate the link between emotions and behavior, which also emerges in language. We introduce the task of coping identification, together with a corpus to do so, constructed via role-playing. We find that coping strategies realize in text even though they are challenging to recognize, both for humans and automatic systems trained and prompted on the same task. We thus open up a promising research direction to enhance the capability of models to better capture emotion mechanisms from text.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, Kai Chen

Large language models (LLMs) have demonstrated impressive capabilities across various tasks, but their performance is highly sensitive to the prompts utilized. This variability poses challenges for accurate assessment and user satisfaction. Current research frequently overlooks instance-level prompt variations and their implications on subjective evaluations. To address these shortcomings, we introduce **ProSA**, a framework designed to evaluate and comprehend prompt sensitivity in LLMs. ProSA incorporates a novel sensitivity metric,

PromptSensiScore, and leverages decoding confidence to elucidate underlying mechanisms. Our extensive study, spanning multiple tasks, uncovers that prompt sensitivity fluctuates across datasets and models, with larger models exhibiting enhanced robustness. We observe that few-shot examples can alleviate this sensitivity issue, and subjective evaluations are also susceptible to prompt sensitivities, particularly in complex, reasoning-oriented tasks. Furthermore, our findings indicate that higher model confidence correlates with increased prompt robustness. We believe this work will serve as a helpful tool in studying prompt sensitivity of LLMs. The project is released at: <https://github.com/open-compass/ProSA>.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning

Yinzu Quan, Zefang Liu

In this paper, we introduce EconLogicQA, a rigorous benchmark designed to assess the sequential reasoning capabilities of large language models (LLMs) within the intricate realms of economics, business, and supply chain management. Diverging from traditional benchmarks that predict subsequent events individually, EconLogicQA poses a more challenging task: it requires models to discern and sequence multiple interconnected events, capturing the complexity of economic logics. EconLogicQA comprises an array of multi-event scenarios derived from economic articles, which necessitate an insightful understanding of both temporal and logical event relationships. Through comprehensive evaluations, we exhibit that EconLogicQA effectively gauges a LLM's proficiency in navigating the sequential complexities inherent in economic contexts. We provide a detailed description of EconLogicQA dataset and shows the outcomes from evaluating the benchmark across various leading-edge LLMs, thereby offering a thorough perspective on their sequential reasoning potential in economic contexts. Our benchmark dataset is available at https://huggingface.co/datasets/yinzu-quan/econ_logic_qa.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning

Shramay Palta, Nishant Balepur, Peter A. Rankel, Sarah Wiegreffe, Marine Carpuat, Rachel Rudinger

Questions involving commonsense reasoning about everyday situations often admit many *possible* or *plausible* answers. In contrast, multiple-choice question (MCQ) benchmarks for commonsense reasoning require a hard selection of a single correct answer, which, in principle, should represent the *most plausible* answer choice. On 250 MCQ items sampled from two commonsense reasoning benchmarks, we collect 5,000 independent plausibility judgments on answer choices. We find that for over 20% of the sampled MCQs, the answer choice rated most plausible does not match the benchmark gold answers; upon manual inspection, we confirm that this subset exhibits higher rates of problems like ambiguity or semantic mismatch between question and answer choices. Experiments with LLMs reveal low accuracy and high variation in performance on the subset, suggesting our plausibility criterion may be helpful in identifying more reliable benchmark items for commonsense evaluation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Language Models Still Struggle to Zero-shot Reason about Time Series

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartwigsen, Tim Althoff

Time series are critical for decision-making in fields like finance and healthcare. Their importance has driven a recent influx of works passing time series into language models, leading to non-trivial forecasting on some datasets. But it remains unknown whether non-trivial forecasting implies that language models can reason about time series. To address this gap, we generate a first-of-its-kind evaluation framework for time series reasoning, including formal tasks and a corresponding dataset of multi-scale time series paired with text captions across ten domains. Using these data, we probe whether language models achieve three forms of reasoning: (1) Etiological Reasoning—given an input time series, can the language model identify the scenario that most likely created it? (2) Question Answering—can a language model answer factual questions about time series? (3) Context-Aided Forecasting—does highly relevant textual context improve a language model's time series forecasts? We find that otherwise highly-capable language models demonstrate surprisingly limited time series reasoning: they score marginally above random on etiological and question answering tasks (up to 30 percentage points worse than humans) and show modest success in using context to improve forecasting. These weakness showcases that time series reasoning is an impactful, yet deeply underdeveloped direction for language model research. We also make our datasets public to support further research in this direction.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

RoQLlama: A Lightweight Romanian Adapted Language Model

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, Dumitru-Clementin Cercel

The remarkable achievements obtained by open-source large language models (LLMs) in recent years have predominantly been concentrated on tasks involving the English language. In this paper, we aim to advance the performance of Llama2 models on Romanian tasks. We tackle the problem of reduced computing resources by using QLoRA for training. We release RoQLlama-7b, a quantized LLM, which shows equal or improved results compared to its full-sized counterpart when tested on seven Romanian downstream tasks in the zero-shot setup. Also, it consistently achieves higher average scores across all few-shot prompts. Additionally, we introduce a novel Romanian dataset, namely RoMedQA, which contains single-choice medical questions in Romanian.

Special Theme: Efficiency in Model Algorithms, Training, and Inference 3

Nov 14 (Thu) 10:30-12:00 - Room: Riverfront Hall

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Change Is the Only Constant: Dynamic LLM Slicing based on Layer Redundancy

Razvan-Gabriel Dumitriu, Paul Ioan Clotan, Vikas Yadav, Darius Peteteala, Mihai Surdeanu

This paper introduces a novel model compression approach through dynamic layer-specific pruning in Large Language Models (LLMs), enhancing the traditional methodology established by SliceGPT. By transitioning from constant to dynamic slicing, our method leverages the newly proposed Layer Redundancy (LR) score, which assesses how much change each layer changes its input by measuring the cosine similarity of the input to the output of the layer. We use this score to prime parts of individual layers based on redundancy in such a way that the average pruned percentage for all layers is a fixed value. We conducted extensive experiments using models like Llama3-8B and Mistral-7B on multiple datasets, evaluating different slicing bases and percentages to determine optimal configurations that balance efficiency and performance. Our findings show that our dynamic slicing approach not only maintains but, in many cases, enhances model performance compared to the baseline established by constant slicing methods. For instance, in several settings, we see performance improvements of up to 5% over the SliceGPT baseline. Additionally, a perplexity decrease by as much as 7% was observed across multiple benchmarks, validating the effectiveness of our method. The code, model weights, and datasets are open-sourced at <https://github.com/RazvanDu/DynamicSlicing>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Prefixing Attention Sinks can Mitigate Activation Outliers for Large Language Model Quantization

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeon Kim, Jaeho Lee

Despite recent advances in LLM quantization, activation quantization remains to be challenging due to the activation outliers. Conventional remedies, e.g., mixing precisions for different channels, introduce extra overhead and reduce the speedup. In this work, we develop a simple yet effective strategy to facilitate per-tensor activation quantization by preventing the generation of problematic tokens. Precisely, we propose a method to find a set of key-value cache, coined `_CushionCache_`, which mitigates outliers in subsequent tokens when inserted as a prefix. `CushionCache` works in two steps: First, we greedily search for a prompt token sequence that minimizes the maximum activation values in subsequent tokens. Then, we further tune the token cache to regularize the activations of subsequent tokens to be more quantization-friendly. The proposed method successfully addresses activation outliers of LLMs, providing a substantial performance boost for per-tensor activation quantization methods. We thoroughly evaluate our method over a wide range of models and benchmarks and find that it significantly surpasses the established baseline of per-tensor W8A8 quantization and can be seamlessly integrated with the recent activation quantization method.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Unsupervised Human Preference Learning

Sunuk Shashidhar, Abhinav Chinta, Vaibhav Sahai, Dilek Hakkani Tur

Large language models demonstrate impressive reasoning abilities but struggle to provide personalized content due to their lack of individual user preference information. Existing methods, such as in-context learning and parameter-efficient fine-tuning, fall short in capturing the complexity of human preferences, especially given the small, personal datasets individuals possess. In this paper, we propose a novel approach utilizing small parameter models as preference agents to generate natural language rules that guide a larger, pre-trained model, enabling efficient personalization. Our method involves a small, local "steering wheel" model that directs the outputs of a much larger foundation model, producing content tailored to an individual's preferences while leveraging the extensive knowledge and capabilities of the large model. Importantly, this personalization is achieved without the need to fine-tune the large model. Experimental results on email and article datasets, demonstrate that our technique significantly outperforms baseline personalization methods. By allowing foundation models to adapt to individual preferences in a data and compute-efficient manner, our approach paves the way for highly personalized language model applications.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

From Bottom to Top: Extending the Potential of Parameter Efficient Fine-Tuning

Jihao Gu, Zelin Wang, Yibo Zhang, Zijie Zhang, Ping Gong

With the proliferation of large language models, Parameter Efficient Fine-Tuning (PEFT) method, which freeze pre-trained parameters and only fine-tune a few task-specific parameters, are playing an increasingly important role. However, previous work primarily applied uniform operations across all layers of the model, overlooking the fact that different layers in a transformer store different information. In the process of exploration, We find that there is a significant differences in fine-tuning strategies between different layers, and fine-tuning only a subset of layers can even achieve comparable performance. Based on this, we propose the Hybrid LoRA-Prefix Tuning(HLPT) method, which uses enhanced LoRA and Prefix-tuning methods with learnable adaptive mechanism separately for the bottom and top layers, and the Half Hybrid LoRA-Prefix Tuning(H^2 LPT) method, which goes a step further, reducing the parameter count to nearly half by omitting fine-tuning in the middle layers. Extensive experiments with large language models on various downstream tasks provide strong evidence for the potential of PEFT focusing on different layers' interactions and the effectiveness of our methods. Furthermore, we validate the robustness of these methods and their advantages in speeding up training convergence, reducing inference time requirements.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

MPT: Multimodal Prompt Tuning for Zero-shot Instruction Learning

Taowen Wang, Yiyang Liu, James Chenhao Liang, junhan zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, Cheng Han, Lifu Huang, Qifan Wang, Dongfang Liu

Multimodal Large Language Models (MLLMs) demonstrate remarkable performance across a wide range of domains, with increasing emphasis on enhancing their zero-shot generalization capabilities for unseen tasks across various modalities. Instruction tuning has emerged as an effective strategy for achieving zero-shot generalization by finetuning pretrained models on diverse multimodal tasks. As the scale of MLLMs continues to grow, parameter-efficient finetuning becomes increasingly critical. However, most existing parameter-efficient approaches focus only on single modalities and often overlook the multimodal characteristics during finetuning. In this work, we introduce a novel Multimodal Prompt Tuning (M^2 PT) approach for efficient instruction tuning of MLLMs. M^2 PT effectively integrates visual and textual prompts into the vision encoder and language processor respectively during finetuning, facilitating the extraction and alignment of features across modalities. Empirical results on various multimodal evaluation datasets demonstrate the superior performance of our approach compared to several state-of-the-art baselines. A comprehensive set of ablation studies validates the effectiveness of our prompt design and the efficiency of our approach.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

VPTQ: Extreme Low-bit Vector Post-Training Quantization for Large Language Models

Yifei Liu, Jicheng Wen, Yang Wang, Shengyu Ye, Li Lyra Zhang, Ting Cao, Cheng Li, Mao Yang

Scaling model size significantly challenges the deployment and inference of Large Language Models (LLMs). Due to the redundancy in LLM weights, recent research has focused on pushing weight-only quantization to extremely low-bit (even down to 2 bits). It reduces memory requirements, optimizes storage costs, and decreases memory bandwidth needs during inference. However, due to numerical representation limitations, traditional scalar-based weight quantization struggles to achieve such extreme low-bit. Recent research on Vector Quantization (VQ) for LLMs has demonstrated the potential for extremely low-bit model quantization by compressing vectors into indices using lookup tables. In this paper, we introduce **Vector Post-Training Quantization (VPTQ)** for extremely low-bit quantization of LLMs. We use Second-Order Optimization to formulate the LLM VQ problem and guide our quantization algorithm design by solving the optimization. We further refine the weights using Channel-Independent Second-Order Optimization for a granular VQ. In addition, by decomposing the optimization problem, we propose a brief and effective codebook initialization algorithm. We also extend VPTQ to support residual and outlier quantization, which enhances model accuracy and further compresses the model. Our experimental results show that VPTQ reduces model quantization perplexity by 0.01-0.34 on LLaMA-2, 0.38-0.68 on Mistral-7B, 4.41-7.34 on LLaMA-3 over SOTA at 2-bit, with an average accuracy improvement of 0.79-1.5% on LLaMA-2, 1% on Mistral-7B, 11-22% on LLaMA-3 on QA tasks on average. We only utilize 10.4-18.6% of the quantization algorithm execution time, resulting in a 1.6-1.8× increase in inference throughput compared to SOTA.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Can Active Label Correction Improve LLM-based Modular AI Systems?

Karan Taneja, Ashok Goel

Modular AI systems can be developed using LLM-prompts-based modules to minimize deployment time even for complex tasks. However, these systems do not always perform well and improving them using the data traces collected from a deployment remains an open challenge. The data traces contain LLM inputs and outputs, but the annotations from LLMs are noisy. We hypothesize that Active Label Correction

(ALC) can be used on the collected data to train smaller task-specific improved models that can replace LLM-based modules. In this paper, we study the noise in three GPT-3.5-annotated datasets and their denoising with human feedback. We also propose a novel method ALC3 that iteratively applies three updates to the training dataset: auto-correction, correction using human feedback and filtering. Our results show that ALC3 can lead to oracle performance with feedback on 17-24% fewer examples than the number of noisy examples in the dataset across three different NLP tasks.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

ORPO: Monolithic Preference Optimization without Reference Model

Jiwoo Hong, Noah Lee, James Thorne

While recent preference alignment algorithms for language models have demonstrated promising results, supervised fine-tuning (SFT) remains imperative for achieving successful convergence. In this paper, we revisit SFT in the context of preference alignment, emphasizing that a minor penalty for the disfavored style is sufficient for preference alignment. Building on this foundation, we introduce a straightforward reference model-free monolithic odds ratio preference optimization algorithm, ORPO, eliminating the need for an additional preference alignment phase. We demonstrate, both empirically and theoretically, that the odds ratio is a sensible choice for contrasting favored and disfavored styles during SFT across diverse sizes from 125M to 7B. Specifically, fine-tuning Phi-2 (2.7B), Llama-2 (7B), and Mistral (7B) with ORPO on the UltraFeedback alone surpasses the performance of state-of-the-art language models including Llama-2 Chat and Zephyr with more than 7B and 13B parameters: achieving up to 12.20% on AlpacaEval 2.0 (Figure 1), and 7.32 in MT-Bench (Table 2). We release code and model checkpoints for Mistral-ORPO- α (7B) and Mistral-ORPO- β (7B).

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Decoding with Limited Teacher Supervision Requires Understanding When to Trust the Teacher

Hyunjong Ok, Jegwang Ryu, Jaeho Lee

How can small-scale large language models (LLMs) efficiently utilize the supervision of LLMs to improve their generative quality? This question has been well studied in scenarios where there is no restriction on the number of LLM supervisions one can use, giving birth to many decoding algorithms that utilize supervision without further training. However, it is still unclear what is an effective strategy under the *limited supervision* scenario, where we assume that no more than a few tokens can be generated by LLMs. To this end, we develop an algorithm to effectively aggregate the small-scale LLM and LLM predictions on initial tokens so that the generated tokens can more accurately condition the subsequent token generation by small-scale LLM only. Critically, we find that it is essential to adaptively overtrust or disregard the LLM prediction based on the confidence of the small-scale LLM. Through our experiments on a wide range of models and datasets, we demonstrate that our method provides a consistent improvement over conventional decoding strategies. **Code:** <https://github.com/HJ-Ok/DecLimSup>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

ShadowLLM: Predictor-based Contextual Sparsity for Large Language Models

Yash Akhauri, Ahmed F AbouElhamayel, Jordan Dotzel, Zhiru Zhang, Alexander M Rush, Safeen Huda, Mohamed S Abdelfattah

The high power consumption and latency-sensitive deployments of large language models (LLMs) have motivated efficiency techniques like quantization and sparsity. Contextual sparsity, where the sparsity pattern is input-dependent, is crucial in LLMs because the permanent removal of attention heads or neurons from LLMs can significantly degrade accuracy. Prior work has attempted to model contextual sparsity using neural networks trained to predict activation magnitudes, which can be used to dynamically prune structures with low predicted activation magnitude. In this paper, we look beyond magnitude-based pruning criteria to assess attention head and neuron importance in LLMs. We develop a novel predictor called ShadowLLM, which can shadow the LLM behavior and enforce better sparsity patterns, resulting in over 15% improvement in end-to-end accuracy compared to prior methods. In addition, ShadowLLM achieves up to a 20% speed-up over the state-of-the-art DejaVu framework. These enhancements are validated on Llama-2 and OPT models with up to 30 billion parameters. Our code is available at https://github.com/abdefattah-lab/shadow_llm/

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Multi-expert Prompting Improves Reliability, Safety and Usefulness of Large Language Models

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, Nancy F. Chen

We present Multi-expert Prompting, a novel enhancement of ExpertPrompting (Xu et al., 2023), designed to improve the large language model (LLM) generation. Specifically, it guides an LLM to fulfill an input instruction by simulating multiple experts, aggregating their responses, and selecting the best among individual and aggregated responses. This process is performed in a single chain of thoughts through our seven carefully designed subtasks derived from the Nominal Group Technique (Ven and Delbecq, 1974), a well-established decision-making framework. Our evaluations demonstrate that Multi-expert Prompting significantly outperforms ExpertPrompting and comparable baselines in enhancing the truthfulness, factuality, informativeness, and usefulness of responses while reducing toxicity and hurtfulness. It further achieves state-of-the-art truthfulness by outperforming the best baseline by 8.69% with ChatGPT. Multi-expert Prompting is efficient, explainable, and highly adaptable to diverse scenarios, eliminating the need for manual prompt construction.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Memorize Step by Step: Efficient Long-Context Prefilling with Incremental Memory and Decremental Chunk

Zhiyuan Zeng, Qipeng Guo, Xiaoran Liu, Zhangye Yin, Wentao Shu, Mianqiu Huang, Bo Wang, Yunhua Zhou, Linlin Li, Qun Liu, Xipeng Qiu

The evolution of Large Language Models (LLMs) has led to significant advancements, with models like Claude and Gemini capable of processing contexts up to 1 million tokens. However, efficiently handling long sequences remains challenging, particularly during the prefilling stage when input lengths exceed GPU memory capacity. Traditional methods often segment sequence into chunks and compress them iteratively with fixed-size memory. However, our empirical analysis shows that the fixed-size memory results in wasted computational and GPU memory resources. Therefore, we introduce Incremental Memory (IM), a method that starts with a small memory size and gradually increases it, optimizing computational efficiency. Additionally, we propose Decremental Chunk based on Incremental Memory (IMDC), which reduces chunk size while increasing memory size, ensuring stable and lower GPU memory usage. Our experiments demonstrate that IMDC is consistently faster (1.45x) and reduces GPU memory consumption by 23.3% compared to fixed-size memory, achieving comparable performance on the LongBench Benchmark.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Scalable Data Ablation Approximations for Language Models through Modular Training and Merging

Clara Na, Ian Magnusson, Ananya Harsh Jha, Tom Sherborne, Emma Strubell, Jesse Dodge, Pradeep Dasigi

Training data compositions for Large Language Models (LLMs) can significantly affect their downstream performance. However, a thorough data ablation study exploring large sets of candidate data mixtures is typically prohibitively expensive since the full effect is seen only after training the models; this can lead practitioners to settle for sub-optimal data mixtures. We propose an efficient method for approximating data ablations which trains individual models on subsets of a training corpus and reuses them across evaluations of combinations of subsets. In continued pre-training experiments, we find that, given an arbitrary evaluation set, the perplexity score of a single model trained on a candidate

set of data is strongly correlated with perplexity scores of parameter averages of models trained on distinct partitions of that data. From this finding, we posit that researchers and practitioners can conduct inexpensive simulations of data ablations by maintaining a pool of models that were each trained on partitions of a large training corpus, and assessing candidate data mixtures by evaluating parameter averages of combinations of these models. This approach allows for substantial improvements in amortized training efficiency – scaling only linearly with respect to new data – by enabling reuse of previous training computation, opening new avenues for improving model performance through rigorous, incremental data assessment and mixing.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Attention Score is not All You Need for Token Importance Indicator in KV Cache Reduction: Value Also Matters

Zhiyu Guo, Hidetaka Kamigaito, Taro Watanabe

Scaling the context size of large language models (LLMs) enables them to perform various new tasks, e.g., book summarization. However, the memory cost of the Key and Value (KV) cache in attention significantly limits the practical applications of LLMs. Recent works have explored token pruning for KV cache reduction in LLMs, relying solely on attention scores as a token importance indicator. However, our investigation into value vector norms revealed a notably non-uniform pattern questioning their reliance only on attention scores. Inspired by this, we propose a new method: Value-Aware Token Pruning (VATP) which uses both attention scores and the ℓ_1 norm of value vectors to evaluate token importance. Extensive experiments on LLaMA2-7B-chat and Vicuna-v1.5-7B across 16 LongBench tasks demonstrate that VATP outperforms attention-score-only baselines in over 12 tasks, confirming the effectiveness of incorporating value vector norms into token importance evaluation of LLMs.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Scalable Fine-tuning from Multiple Data Sources: A First-Order Approximation Approach

Dongyue Li, Ziniu Zhang, Lu Wang, Hongyang R. Zhang

We study the problem of fine-tuning a language model (LM) for a target task by optimally using the information from n auxiliary tasks. This problem has broad applications in NLP, such as targeted instruction tuning and data selection in chain-of-thought fine-tuning. The key challenge of this problem is that not all auxiliary tasks are useful to improve the performance of the target task. Thus, choosing the right subset of auxiliary tasks is crucial. Conventional subset selection methods, such as forward & backward selection, are unsuitable for LM fine-tuning because they require repeated training on subsets of auxiliary tasks. This paper introduces a new algorithm to estimate model fine-tuning performances without repeated training. Our algorithm first performs multitask training using the data of all the tasks to obtain a meta initialization. Then, we approximate the model fine-tuning loss of a subset using functional values and gradients from the meta initialization. Empirically, we find that this gradient-based approximation holds with remarkable accuracy for twelve transformer-based LMs. Thus, we can now estimate fine-tuning performances on CPUs within a few seconds. We conduct extensive experiments to validate our approach, delivering a speedup of $30 \times$ over conventional subset selection while incurring only 1% error of the true fine-tuning performances. In downstream evaluations of instruction tuning and chain-of-thought fine-tuning, our approach improves over prior methods that utilize gradient or representation similarity for subset selection by up to 3.8%.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

STTATTS: Unified Speech-To-Text And Text-To-Speech Model

Hawau Olamide Toyin, Hao Li, Hanan Aldarmaki

Speech recognition and speech synthesis models are typically trained separately, each with its own set of learning objectives, training data, and model parameters, resulting in two distinct large networks. We propose a parameter-efficient approach to learning ASR and TTS jointly via a multi-task learning objective and shared parameters. Our evaluation demonstrates that the performance of our multi-task model is comparable to that of individually trained models while significantly saving computational and memory costs ($\sim 50\%$ reduction in the total number of parameters required for the two tasks combined). We experiment with English as a resource-rich language, and Arabic as a relatively low-resource language due to shortage of TTS data. Our models are trained with publicly available data, and both the training code and model checkpoints are openly available for further research.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

RoLoRA: Fine-tuning Rotated Outlier-free LLMs for Effective Weight-Activation Quantization

Xijie Huang, Zechun Liu, Shih-Yang Liu, Kwang-Ting Cheng

Low-Rank Adaptation (LoRA), as a representative Parameter-Efficient Fine-Tuning (PEFT) method, significantly enhances the training efficiency by updating only a small portion of the weights in Large Language Models (LLMs). Recently, weight-only quantization techniques have also been applied to LoRA methods to reduce the memory footprint of fine-tuning. However, applying weight-activation quantization to the LoRA pipeline is under-explored, and we observe substantial performance degradation primarily due to the presence of activation outliers. In this work, we propose RoLoRA, the first LoRA-based scheme to apply rotation for outlier elimination, and then fine-tune rotated outlier-free LLMs for effective weight-activation quantization. Different from previous work tackling the outlier challenges from a post-training perspective, we propose rotation-aware fine-tuning to eliminate and preserve the outlier-free characteristics brought by rotation operations. RoLoRA can improve low-bit LoRA convergence and post-training quantization robustness in weight-activation settings. RoLoRA is evaluated across various LLM series (LLaMA2, LLaMA3, LLaVA-1.5), tasks, and quantization settings, achieving up to 29.5% absolute accuracy gain of 4-bit weight-activation quantized LLaMA2-13B on commonsense reasoning tasks compared to LoRA baseline. We further demonstrate its effectiveness on Large Multimodal Models (LMMs) and prove the compatibility with advanced LoRA variants.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

PromptIntern: Saving Inference Costs by Internalizing Recurrent Prompt during Large Language Model Fine-tuning

Jiaru Zou, Mengyu Zhou, Tao Li, Shi Han, Dongmei Zhang

Recent advances in fine-tuning large language models (LLMs) have greatly enhanced their usage in domain-specific tasks. Despite the success, fine-tuning continues to rely on repeated and lengthy prompts, which escalate computational expenses, require more resources, and lead to slower inference. In this paper, we present a novel approach, PromptIntern, which internalizes prompt knowledge during model fine-tuning to achieve efficient inference and save costs. Instead of compressing the prompts for a vanilla model, PromptIntern aims to embed the recurrent prompt directly into the model parameters. We design a fine-tuning pipeline that includes instruction template compression, few-shot example absorption, and a progressive internalization strategy, effectively diminishing the need for intricate prompts during inference. Comprehensive experiments on challenging NL2Code tasks demonstrate that our method reduces input tokens by more than 90%, accelerates inference by 4.2 times, and reduces monetary inference costs by 88.3%.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

MoE-I²: Compressing Mixture of Experts Models through Inter-Expert Pruning and Intra-Expert Low-Rank Decomposition

Cheng Yang, Yang Sui, Jingi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, Bo Yuan

The emergence of Mixture of Experts (MoE) LLMs has significantly advanced the development of language models. Compared to traditional LLMs, MoE LLMs outperform traditional LLMs by achieving higher performance with considerably fewer activated parameters. Despite this efficiency, their enormous parameter size still leads to high deployment costs. In this paper, we introduce a two-stage compression method

tailored for MoE to reduce the model size and decrease the computational cost. First, in the inter-expert pruning stage, we analyze the importance of each layer and propose the Layer-wise Genetic Search and Block-wise KT-Reception Field with the non-uniform pruning ratio to prune the individual expert. Second, in the intra-expert decomposition stage, we apply the low-rank decomposition to further compress the parameters within the remaining experts. Extensive experiments on Qwen1.5-MoE-A2.7B, Deepseek-V2-Lite, and Mixtral-8 \times 7B, demonstrate that our proposed methods can both reduce the model size and enhance inference efficiency while maintaining performance in various zero-shot tasks.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Fighting Randomness with Randomness: Mitigating Optimisation Instability of Fine-Tuning using Delayed Ensemble and Noisy Interpolation

Branislav Pecher, Jan Cegin, Robert Belanec, Jakub Simko, Ivan Srba, Maria Bielikova

While fine-tuning of pre-trained language models generally helps to overcome the lack of labelled training samples, it also displays model performance instability. This instability mainly originates from randomness in initialisation or data shuffling. To address this, researchers either modify the training process or augment the available samples, which typically results in increased computational costs. We propose a new mitigation strategy, called **Delayed Ensemble with Noisy Interpolation (DENI)**, that leverages the strengths of ensembling, noise regularisation and model interpolation, while retaining computational efficiency. We compare DENI with 9 representative mitigation strategies across 3 models, 4 tuning strategies and 7 text classification datasets. We show that: 1) DENI outperforms the best performing mitigation strategy (Ensemble), while using only a fraction of its cost; 2) the mitigation strategies are beneficial for parameter-efficient fine-tuning (PEFT) methods, outperforming full fine-tuning in specific cases; and 3) combining DENI with data augmentation often leads to even more effective instability mitigation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity

Selim Furkan Tekin, Faith Ilhan, Tiansheng Huang, Sihao Hu, Ling Liu

Combining large language models during training or at inference time has shown substantial performance gain over component LLMs. This paper presents LLM-TOPLA, a diversity-optimized LLM ensemble method with three unique properties: (i) We introduce the focal diversity metric to capture the diversity-performance correlation among component LLMs of an ensemble. (ii) We develop a diversity-optimized ensemble pruning algorithm to select the top-k sub-ensembles from a pool of N base LLMs. Our pruning method recommends top-performing LLM subensembles of size S , often much smaller than N . (iii) We generate new output for each prompt query by utilizing a learn-to-ensemble approach, which learns to detect and resolve the output inconsistency among all component LLMs of an ensemble. Extensive evaluation on four different benchmarks shows good performance gain over the best LLM ensemble methods: (i) In constrained solution set problems, LLM-TOPLA outperforms the best-performing ensemble (Mixtral) by 2.2% in accuracy on MMLU and the best-performing LLM ensemble (MoreAgent) on GSM8k by 2.1%. (ii) In generative tasks, LLM-TOPLA outperforms the top-2 performers (Llama70b/Mixtral) on SearchQA by 3.9x in F1, and on XSum by more than 38 in ROUGE-1. Our code and dataset, which contains outputs of 8 modern LLMs on 4 benchmarks is available at <https://github.com/git-disl/llm-topla>

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Beyond Accuracy Optimization: Computer Vision Losses for Large Language Model Fine-Tuning

Daniele Rege Cambrini, Giuseppe Gallipoli, Irene Benedetto, Luca Cagliero, Paolo Garza

Large Language Models (LLMs) have demonstrated impressive performance across various tasks. However, current training approaches combine standard cross-entropy loss with extensive data, human feedback, or ad hoc methods to enhance performance. These solutions are often not scalable or feasible due to their associated costs, complexity, or resource requirements. This study investigates the use of established semantic segmentation loss functions in natural language generation to create a versatile, practical, and scalable solution for fine-tuning different architectures. We evaluate their effectiveness in solving Math Word Problems and question answering across different models of varying sizes. For the analyzed tasks, we found that the traditional Cross-Entropy loss represents a sub-optimal choice, while models trained to minimize alternative (task-dependent) losses, such as Focal or Lovász, achieve a mean improvement of +36% on exact match without requiring additional data or human feedback. These findings suggest a promising pathway for more efficient and accessible training processes.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Fast Matrix Multiplications for Lookup Table-Quantized LLMs

Han Guo, William Brandon, Radoslin Cholakov, Jonathan Ragan-Kelley, Eric P. Xing, Yoon Kim

The deployment of large language models (LLMs) is often constrained by memory bandwidth, where the primary bottleneck is the cost of transferring model parameters from the GPU's global memory to its registers. When coupled with custom kernels that fuse the dequantization and matmul operations, weight-only quantization can thus enable faster inference by reducing the amount of memory movement. However, developing high-performance kernels for weight-quantized LLMs presents substantial challenges, especially when the weights are compressed to non-evenly-divisible bit widths (e.g., 3 bits) with non-uniform, lookup table (LUT) quantization. This paper describes FLUTE, a flexible lookup-table engine for LUT-quantized LLMs, which uses offline restructuring of the quantized weight matrix to minimize bit manipulations associated with unpacking, and vectorization and duplication of the lookup table to mitigate shared memory bandwidth constraints. At batch sizes < 32 and quantization group size of 128 (typical in LLM inference), the FLUTE kernel can be 2-4x faster than existing GEMM kernels. As an application of FLUTE, we explore a simple extension to lookup table-based NormalFloat quantization and apply it to quantize LLaMA3 to various configurations, obtaining competitive quantization performance against strong baselines while obtaining an end-to-end throughput increase of 1.5 to 2 times.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Temperature-Centric Investigation of Speculative Decoding with Knowledge Distillation

Siru Ouyang, Shuohang Wang, Minhao Jiang, Ming Zhong, Donghan Yu, Jiawei Han, yelong shen

Speculative decoding stands as a pivotal technique to expedite inference in autoregressive (large) language models. This method employs a smaller *draft* model to speculate a block of tokens, which the *target* model then evaluates for acceptance. Despite a wealth of studies aimed at increasing the efficiency of speculative decoding, the influence of generation configurations on the decoding process remains poorly understood, especially concerning decoding temperatures. This paper delves into the effects of decoding temperatures on speculative decodings efficacy. Beginning with knowledge distillation (KD), we first highlight the challenge of decoding at higher temperatures, and demonstrate KD in a consistent temperature setting could be a remedy. We also investigate the effects of out-of-domain testing sets with out-of-range temperatures. Building upon these findings, we take an initial step to further the speedup for speculative decoding, particularly in a high-temperature generation setting. Our work offers new insights into how generation configurations drastically affect the performance of speculative decoding, and underscores the need for developing methods that focus on diverse decoding configurations.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Style-Compress: An LLM-Based Prompt Compression Framework Considering Task-Specific Styles

Xiao Pu, Tianxing He, Xiaojun Wan

Prompt compression condenses contexts while maintaining their informativeness for different usage scenarios. It not only shortens the inference time and reduces computational costs during the usage of large language models, but also lowers expenses when using closed-source models. In a preliminary study, we discover that when instructing language models to compress prompts, different compression styles (e.g., extractive or abstractive) impact performance of compressed prompts on downstream tasks. Building on this insight, we propose Style-Compress, a lightweight framework that adapts a smaller language model to compress prompts for a larger model on a new task without additional training. Our approach iteratively generates and selects effective compressed prompts as task-specific demonstrations through style variation and in-context learning, enabling smaller models to act as efficient compressors with task-specific examples. Style-Compress outperforms two baseline compression models in four tasks: original prompt reconstruction, text summarization, multi-hop QA, and CoT reasoning. In addition, with only 10 samples and 100 queries for adaptation, prompts compressed by Style-Compress achieve performance on par with or better than original prompts at a compression ratio of 0.25 or 0.5.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Eigen Attention: Attention in Low-Rank Space for KV Cache Compression

Utkarsh Saxena, Gobinda Saha, Sakshi Choudhary, Kaushik Roy

Large language models (LLMs) represent a groundbreaking advancement in the domain of natural language processing due to their impressive reasoning abilities. Recently, there has been considerable interest in increasing the context lengths for these models to enhance their applicability to complex tasks. However, at long context lengths and large batch sizes, the key-value (KV) cache, which stores the attention keys and values, emerges as the new bottleneck in memory usage during inference. To address this, we propose Eigen Attention, which performs the attention operation in a low-rank space, thereby reducing the KV cache memory overhead. Our proposed approach is orthogonal to existing KV cache compression techniques and can be used synergistically with them. Through extensive experiments over OPT, MPT, and Llama model families, we demonstrate that Eigen Attention results in up to 40% reduction in KV cache sizes and up to 60% reduction in attention operation latency, with minimal drop in performance.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

How Does Quantization Affect Multilingual LLMs?

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, Sebastian Ruder

Quantization techniques are widely used to improve inference speed and deployment of large language models. While a wide body of work examines the impact of quantization on LLMs in English, none have evaluated across languages. We conduct a thorough analysis of quantized multilingual LLMs, focusing on performance across languages and at varying scales. We use automatic benchmarks, LLM-as-a-Judge, and human evaluation, finding that (1) harmful effects of quantization are apparent in human evaluation, which automatic metrics severely underestimate; a 1.7% average drop in Japanese across automatic tasks corresponds to a 16.0% drop reported by human evaluators on realistic prompts; (2) languages are disparately affected by quantization, with non-Latin script languages impacted worst; and (3) challenging tasks like mathematical reasoning degrade fastest. As the ability to serve low-compute models is critical for wide global adoption of NLP technologies, our results urge consideration of multilingual performance as a key evaluation criterion for efficient models.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

When Compression Meets Model Compression: Memory-Efficient Double Compression for Large Language Models

Weilan Wang, Yu Mao, TANG DONGDONG, Du Hongchao, Nan Guan, Chun Jason Xue

Large language models (LLMs) exhibit excellent performance in various tasks. However, the memory requirements of LLMs present a great challenge when deploying on memory-limited devices, even for quantized LLMs. This paper introduces a framework to compress LLM after quantization further, achieving about 2.2x compression ratio. A compression-aware quantization is first proposed to enhance model weight compressibility by re-scaling the model parameters before quantization, followed by a pruning method to improve further. Upon this, we notice that decompression can be a bottleneck during practical scenarios. We then give a detailed analysis of the trade-off between memory usage and latency brought by the proposed method. A speed-adaptive method is proposed to overcome it. The experimental results show inference with the compressed model can achieve a 40% reduction in memory size with negligible loss in accuracy and inference speed.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Divide-or-Conquer? Which Part Should You Distill Your LLM?

Zhuofeng Wu, Richard He Bai, Aonan Zhang, Jiatao Gu, V.G. Vinod Vydiswaran, Navdeep Jaitly, Yizhe Zhang

Recent methods have demonstrated that Large Language Models (LLMs) can solve reasoning tasks better when they are encouraged to solve subtasks of the main task first. In this paper we devise a similar strategy that breaks down reasoning tasks into a problem decomposition phase and a problem solving phase and show that the strategy is able to outperform a single stage solution. Further, we hypothesize that the decomposition should be easier to distill into a smaller model compared to the problem solving because the latter requires large amounts of domain knowledge while the former only requires learning general problem solving strategies. We propose methods to distill these two capabilities and evaluate their impact on reasoning outcomes and inference cost. We find that we can distill the problem decomposition phase and at the same time achieve good generalization across tasks, datasets, and models. However, it is harder to distill the problem solving capability without losing performance and the resulting distilled model struggles with generalization. These results indicate that by using smaller, distilled problem decomposition models in combination with problem solving LLMs we can achieve reasoning with cost-efficient inference and local adaptation.

Nov 14 (Thu) 10:30-12:00 - Riverfront Hall

Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models

Songtao Jiang, Tao Zheng, Yan Zhang, YEYING JIN, Li Yuan, Zuozhu Liu

Recent advancements in general-purpose or domain-specific multimodal large language models (LLMs) have witnessed remarkable progress for medical decision-making. However, they are designed for specific classification or generative tasks, and require model training or finetuning on large-scale datasets with sizeable parameters and tremendous computing, hindering their clinical utility across diverse resource-constrained scenarios in practice. In this paper, we propose a novel and lightweight framework Med-MoE (Mixture-of-Experts) that tackles both discriminative and generative multimodal medical tasks. The learning of Med-MoE consists of three steps: multimodal medical alignment, Instruction tuning and routing, and domain-specific MoE tuning. After aligning multimodal medical images with LLM tokens, we then enable the model for different multimodal medical tasks with instruction tuning, together with a trainable router tailored for expert selection across input modalities. Finally, the model is tuned by integrating the router with multiple domain-specific experts, which are selectively activated and further empowered by meta experts. Comprehensive experiments on both open- and close-end medical question answering (Med-VQA) and image classification tasks across datasets such as VQA-RAD, SLAKE and Path-VQA demonstrate that our model can achieve performance superior to or on par with state-of-the-art baselines, while only requiring approximately 30%-50% of activated model parameters. Extensive analysis and ablations corroborate the effectiveness and practical utility of our method.

Session 12 - Nov 14 (Thu) 14:00-15:30

Computational Social Science and Cultural Analytics 4

Nov 14 (Thu) 14:00-15:30 - Room: Jasmine

Nov 14 (Thu) 14:00-15:30 - Jasmine

Diving LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models

Flor Miriam Plaza-del-Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, Dirk Hovy

Emotions play important epistemological and cognitive roles in our lives, revealing our values and guiding our actions. Previous work has shown that LLMs display biases in emotion attribution along gender lines. However, unlike gender, which says little about our values, religion, as a socio-cultural system, prescribes a set of beliefs and values for its followers. Religions, therefore, cultivate certain emotions. Moreover, these rules are explicitly laid out and interpreted by religious leaders. Using emotion attribution, we explore how different religions are represented in LLMs. We find that major religions in the US and European countries are represented with more nuance, displaying a more shaded model of their beliefs. Eastern religions like Hinduism and Buddhism are strongly stereotyped. Judaism and Islam are stigmatized – the models' refusal skyrocket. We ascribe these to cultural bias in LLMs and the scarcity of NLP literature on religion. In the rare instances where religion is discussed, it is often in the context of toxic language, perpetuating the perception of these religions as inherently toxic. This finding underscores the urgent need to address and rectify these biases. Our research emphasizes the crucial role emotions play in shaping our lives and how our values influence them.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Recent Advances in Online Hate Speech Moderation: Multimodality and the Role of Large Models

Ming Shan Hee, Shivam Sharma, RUI CAO, Palash Nandi, Preslav Nakov, Tammo Chakraborty, Roy Ka-Wei Lee

Moderating hate speech (HS) in the evolving online landscape is a complex challenge, compounded by the multimodal nature of digital content. This survey examines recent advancements in HS moderation, focusing on the burgeoning role of large language models (LLMs) and large multimodal models (LMMs) in detecting, explaining, debiasing, and countering HS. We begin with a comprehensive analysis of current literature, uncovering how text, images, and audio interact to spread HS. The combination of these modalities adds complexity and subtlety to HS dissemination. We also identified research gaps, particularly in underrepresented languages and cultures, and highlight the need for solutions in low-resource settings. The survey concludes with future research directions, including novel AI methodologies, ethical AI governance, and the development of context-aware systems. This overview aims to inspire further research and foster collaboration towards responsible and human-centric approaches to HS moderation in the digital age.

Nov 14 (Thu) 14:00-15:30 - Jasmine

An Experimental Analysis on Evaluating Patent Citations

Rabindra Nath Nandi, Suman Maity, Brian Uzzi, Sourav Medya

The patent citation count is a good indicator of patent quality. This often generates monetary value for the inventors and organizations. However, the factors that influence a patent receiving high citations over the year are still not well understood. With the patents over the past two decades, we study the problem of patent citation prediction and formulate this as a binary classification problem. We create a semantic graph of patents based on their semantic similarities, enabling the use of Graph Neural Network (GNN)-based approaches for predicting citations. Our experimental results demonstrate the effectiveness of our GNN-based methods when applied to the semantic graph, showing that they can accurately predict patent citations using only patent text. More specifically, these methods produce up to 94% recall for patents with high citations and outperform existing baselines. Furthermore, we leverage this constructed graph to gain insights and explanations for the predictions made by the GNNs.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Surveying the Dead Minds: Historical-Psychological Text Analysis with Contextualized Construct Representation (CCR) for Classical Chinese

Yugu Chen, Sixuan Li, Ying Li, Mohammad Atari

In this work, we develop a pipeline for historical-psychological text analysis in classical Chinese. Humans have produced texts in various languages for thousands of years; however, most of the computational literature is focused on contemporary languages and corpora. The emerging field of historical psychology relies on computational techniques to extract aspects of psychology from historical corpora using new methods developed in natural language processing (NLP). The present pipeline, called Contextualized Construct Representations (CCR), combines expert knowledge in psychometrics (i.e., psychological surveys) with text representations generated via Transformer-based language models to measure psychological constructs such as traditionalism, norm strength, and collectivism in classical Chinese corpora. Considering the scarcity of available data, we propose an indirect supervised contrastive learning approach and build the first Chinese historical psychology corpus (C-HI-PSY) to fine-tune pre-trained models. We evaluate the pipeline to demonstrate its superior performance compared with other approaches. The CCR method outperforms word-embedding-based approaches across all of our tasks and exceeds prompting with GPT-4 in most tasks. Finally, we benchmark the pipeline against objective, external data to further verify its validity.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Noise, Novels, Numbers. A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature

ALI ALLAITH, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, Timothy R Tangherlini

We present a framework for detecting and categorizing noise in literary texts, demonstrated through its application to Danish and Norwegian literature from the late 19th century. Noise, understood as "aberrant sonic behaviour," is not only an auditory phenomenon but also a cultural construct tied to the processes of civilization and urbanization. We begin by utilizing topic modeling techniques to identify noise-related documents, followed by fine-tuning BERT-based language models trained on Danish and Norwegian texts to analyze a corpus of over 800 novels. We identify and track the prevalence of noise in these texts, offering insights into the literary perceptions of noise during the Scandinavian "Modern Breakthrough" period (1870-1899). Our contributions include the development of a comprehensive dataset annotated for noise-related segments and their categorization into human-made, non-human-made, and musical noises. This study illustrates the framework's potential for enhancing the understanding of the relationship between noise and its literary representations, providing a deeper appreciation of the auditory elements in literary works, including as sources for cultural history.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Contrastive Entity Coreference and Disambiguation for Historical Texts

Abhishek Arora, Emily Silcock, Melissa Dell, Leander Heldring

Massive-scale historical document collections are crucial for social science research. Despite increasing digitization, these documents typically lack unique cross-document identifiers for individuals mentioned within the texts, as well as individual identifiers from external knowl-

edge bases like Wikipedia/Wikidata. Existing entity disambiguation methods often fall short in accuracy for historical documents, which are replete with individuals not remembered in contemporary knowledge bases. This study makes three key contributions to improve cross-document coreference resolution and disambiguation in historical texts: a massive-scale training dataset replete with hard negatives - that sources over 190 million entity pairs from Wikipedia contexts and disambiguation pages - high-quality evaluation data from hand-labeled historical newswire articles, and trained models evaluated on this historical benchmark. We contrastively train bi-encoder models for coreferencing and disambiguating individuals in historical texts, achieving accurate, scalable performance that identifies out-of-knowledge base individuals. Our approach significantly surpasses other entity disambiguation models on our historical newswire benchmark. Our models also demonstrate competitive performance on modern entity disambiguation benchmarks, particularly on certain news disambiguation datasets.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Susu Box or Piggy Bank: Assessing Cultural Commonsense Knowledge between Ghana and the US

Christabel Acquaye, Haozhe An, Rachel Rudinger

Recent work has highlighted the culturally-contingent nature of commonsense knowledge. We introduce AMAMMER, a test set of 525 multiple-choice questions designed to evaluate the commonsense knowledge of English LLMs, relative to the cultural contexts of Ghana and the United States. To create AMAMMER, we select a set of multiple-choice questions (MCQs) from existing commonsense datasets and rewrite them in a multi-stage process involving surveys of Ghanaian and U.S. participants. In three rounds of surveys, participants from both pools are solicited to (1) write correct and incorrect answer choices, (2) rate individual answer choices on a 5-point Likert scale, and (3) select the best answer choice from the newly-constructed MCQ items, in a final validation step. By engaging participants at multiple stages, our procedure ensures that participant perspectives are incorporated both in the creation and validation of test items, resulting in high levels of agreement within each pool. We evaluate several off-the-shelf English LLMs on AMAMMER. Uniformly, models prefer answers choices that align with the preferences of U.S. annotators over Ghanaian annotators. Additionally, when test items specify a cultural context (Ghana or the U.S.), models exhibit some ability to adapt, but performance is consistently better in U.S. contexts than Ghanaian. As large resources are devoted to the advancement of English LLMs, our findings underscore the need for culturally adaptable models and evaluations to meet the needs of diverse English-speaking populations around the world.

Nov 14 (Thu) 14:00-15:30 - Jasmine

The Lou Dataset - Exploring the Impact of Gender-Fair Language in German Text Classification

Andreas Waldti, Joel Birrer, Anne Lauscher, Iryna Gurevych

Gender-fair language, an evolving linguistic variation in German, fosters inclusion by addressing all genders or using neutral forms. However, there is a notable lack of resources to assess the impact of this language shift on language models (LMS) might not have been trained on examples of this variation. Addressing this gap, we present Lou, the first dataset providing high-quality reformulations for German text classification covering seven tasks, like stance detection and toxicity classification. We evaluate 16 mono- and multi-lingual LMs and find substantial label flips, reduced prediction certainty, and significantly altered attention patterns. However, existing evaluations remain valid, as LM rankings are consistent across original and reformulated instances. Our study provides initial insights into the impact of gender-fair language on classification for German. However, these findings are likely transferable to other languages, as we found consistent patterns in multi-lingual and English LMs.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Understanding Slang with LLMs: Modelling Cross-Cultural Nuances through Paraphrasing

Ifeoluwa Waraola, Nina Dethlefs, Daniel Marciniak

In the realm of social media discourse, the integration of slang enriches communication, reflecting the sociocultural identities of users. This study investigates the capability of large language models (LLMs) to paraphrase slang within climate-related tweets from Nigeria and the UK, with a focus on identifying emotional nuances. Using DistilRoBERTa as the base-line model, we observe its limited comprehension of slang. To improve cross-cultural understanding, we gauge the effectiveness of leading LLMs ChatGPT 4, Gemini, and LLaMA3 in slang paraphrasing. While ChatGPT 4 and Gemini demonstrate comparable effectiveness in slang paraphrasing, LLaMA3 shows less coverage, with all LLMs exhibiting limitations in coverage, especially of Nigerian slang. Our findings underscore the necessity for culturally sensitive LLM development in emotion classification, particularly in non-anglocentric regions.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Hate Personified: Investigating the role of LLMs in content moderation pipeline for hate speech

Sarah Masud, Sahajpreet Singh, Viktor Hanya, Alexander Fraser, Tanmoy Chakraborty

For subjective tasks such as hate detection, where people perceive hate differently, the Large Language Model's (LLM) ability to represent diverse groups is unclear. By including additional context in prompts, we comprehensively analyze LLM's sensitivity to geographical priming, persona attributes, and numerical information to assess how well the needs of various groups are reflected. Our findings on two LLMs, five languages, and six datasets reveal that mimicking persona-based attributes leads to annotation variability. Meanwhile, incorporating geographical signals leads to better regional alignment. We also find that the LLMs are sensitive to numerical anchors, indicating the ability to leverage community-based flagging efforts and exposure to adversaries. Our work provides preliminary guidelines and highlights the nuances of applying LLMs in culturally sensitive cases.

Nov 14 (Thu) 14:00-15:30 - Jasmine

AutoPersuade: A Framework for Evaluating and Explaining Persuasive Arguments

Till Raphael Saenger, Musashi Hinck, Justin Grimmer, Brandon M. Stewart

We introduce a three-part framework for constructing persuasive messages, AutoPersuade. First, we curate a large collection of arguments and gather human evaluations of their persuasiveness. Next, we introduce a novel topic model to identify the features of these arguments that influence persuasion. Finally, we use the model to predict the persuasiveness of new arguments and to assess the causal effects of argument components, offering an explanation of the results. We demonstrate the effectiveness of AutoPersuade in an experimental study on arguments for veganism, validating our findings through human studies and out-of-sample predictions.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Community-Cross-Instruct: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities

Zihao He, Rebecca Dorn, Minh Duc Chu, Siyi Guo, Kristina Lerman

Social scientists use surveys to probe the opinions and beliefs of populations, but these methods are slow, costly, and prone to biases. Recent advances in large language models (LLMs) enable the creating of computational representations or "digital twins" of populations that generate human-like responses mimicking the population's language, styles, and attitudes. We introduce Community-Cross-Instruct, an unsupervised framework for aligning LLMs to online communities to elicit their beliefs. Given a corpus of a community's online discussions, Community-Cross-Instruct automatically generates instruction-output pairs by an advanced LLM to (1) finetune a foundational LLM to faithfully represent that community, and (2) evaluate the alignment of the finetuned model to the community. We demonstrate the method's utility in accurately representing political and diet communities on Reddit. Unlike prior methods requiring human-authored instructions, Community-Cross-Instruct generates instructions in a fully unsupervised manner, enhancing scalability and generalization across domains. This work enables

cost-effective and automated surveying of diverse online communities.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, Haohan Wang

The complexity of text-embedded images presents a formidable challenge in machine learning given the need for multimodal understanding of multiple aspects of expression conveyed by them. While previous research in multimodal analysis has primarily focused on singular aspects such as hate speech and its subclasses, this study expands this focus to encompass multiple aspects of linguistics: hate, targets of hate, stance, and humor. We introduce a novel dataset PrideMM comprising 5,063 text-embedded images associated with the LGBTQ+ Pride movement, thereby addressing a serious gap in existing resources. We conduct extensive experimentation on PrideMM by using unimodal and multimodal baseline methods to establish benchmarks for each task. Additionally, we propose a novel framework MemeCLIP for efficient downstream learning while preserving the knowledge of the pre-trained CLIP model. The results of our experiments show that MemeCLIP achieves superior performance compared to previously proposed frameworks on two real-world datasets. We further compare the performance of MemeCLIP and zero-shot GPT-4 on the hate classification task. Finally, we discuss the shortcomings of our model by qualitatively analyzing misclassified samples. Our code and dataset are publicly available at: <https://github.com/SiddhantBikram/MemeCLIP>.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Building a Multi-Platform, BERT Classifier for Detecting Connective Language

Josephine Lukito, Bin Chen, Gina M. Masullo, Natalie Jomini Stroud

This study presents an approach for detecting connective language—defined as language that facilitates engagement, understanding, and conversation—from social media discussions. We developed and evaluated two types of classifiers: BERT and GPT-3.5 turbo. Our results demonstrate that the BERT classifier significantly outperforms GPT-3.5 turbo in detecting connective language. Furthermore, our analysis confirms that connective language is distinct from related concepts measuring discourse qualities, such as politeness and toxicity. We also explore the potential of BERT-based classifiers for platform-agnostic tools. This research advances our understanding of the linguistic dimensions of online communication and proposes practical tools for detecting connective language across diverse digital environments.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Language is Scary When Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, Tommaso Caselli

We propose misogyny detection as an Argumentative Reasoning task and we investigate the capacity of large language models (LLMs) to understand the implicit reasoning used to convey misogyny in both Italian and English. The central aim is to generate the missing reasoning link between a message and the implied meanings encoding the misogyny. Our study uses argumentation theory as a foundation to form a collection of prompts in both zero-shot and few-shot settings. These prompts integrate different techniques, including chain-of-thought reasoning and augmented knowledge. Our findings show that LLMs fall short on reasoning capabilities about misogynistic comments and that they mostly rely on their implicit knowledge derived from internalized common stereotypes about women to generate implied assumptions, rather than on inductive reasoning.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Style-Shifting Behaviour of the Manosphere on Reddit

Iai Aggarwal, Suzanne Stevenson

Hate speech groups (HSGs) may negatively influence online platforms through their distinctive language, which may affect the tone and topics of other spaces if spread beyond the HSGs. We explore the linguistic style of the Manosphere, a misogynistic HSG, on Reddit. We find that Manospheric authors have a distinct linguistic style using not only uncivil language, but a greater focus on gendered topics, which are retained when posting in other communities. Thus, potentially harmful aspects of Manospheric style carry over into posts on non-Manospheric subreddits, motivating future work to explore how this stylistic spillover may negatively influence community health.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Rater Cohesion and Quality from a Vicarious Perspective

Deepak Pandita, Tharindu Cyril Weerasooriya, Sujan Dutta, Sarah K. K. Luger, Tharindu Ranasinghe, Ashiqur R. KhudaBukhsh, Marcos Zampieri, Christopher M Hanani

Human feedback is essential for building human-centered AI systems across domains where disagreement is prevalent, such as AI safety, content moderation, or sentiment analysis. Many disagreements, particularly in politically charged settings, arise because raters have opposing values or beliefs. Vicarious annotation is a method for breaking down disagreement by asking raters how they think others would annotate the data. In this paper, we explore the use of vicarious annotation with analytical methods for moderating rater disagreement. We employ rater cohesion metrics to study the potential influence of political affiliations and demographic backgrounds on raters' perceptions of offense. Additionally, we utilize CrowdTruth's rater quality metrics, which consider the demographics of the raters, to score the raters and their annotations. We study how the rater quality metrics influence the in-group and cross-group rater cohesion across the personal and vicarious levels.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents

Zengqiang Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, Chuan Xiao

Large Language Models (LLMs) have increasingly been utilized in social simulations, where they are often guided by carefully crafted instructions to stably exhibit human-like behaviors during simulations. Nevertheless, we doubt the necessity of shaping agents' behaviors for accurate social simulations. Instead, this paper emphasizes the importance of spontaneous phenomena, wherein agents deeply engage in contexts and make adaptive decisions without explicit directions. We explored spontaneous cooperation across three competitive scenarios and successfully simulated the gradual emergence of cooperation, findings that align closely with human behavioral data. This approach not only aids the computational social science community in bridging the gap between simulations and real-world dynamics but also offers the AI community a novel method to assess LLMs' capability of deliberate reasoning. Our source code is available at https://github.com/wuzengqing0122/S-ABM_ShallWeTeamUp

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Toeing the party line: election manifestos as a key to understand political discourse on Twitter

Maximilian Maurer, Tanise Ceron, Sebastian Padó, Gabriella Lopesa

Political discourse on Twitter is a moving target: politicians continuously make statements about their positions. It is therefore crucial to track their discourse on social media to understand their ideological positions and goals. However, Twitter data is also challenging to work with since it is ambiguous and often dependent on social context, and consequently, recent work on political positioning has tended to focus strongly on manifestos (parties' electoral programs) rather than social media. In this paper, we extend recently proposed methods to predict pairwise positional similarities between parties from the manifesto case to the Twitter case, using hashtags as a signal to fine-tune text representations.

sentations, without the need for manual annotation. We verify the efficacy of fine-tuning and conduct a series of experiments that assess the robustness of our method for low-resource scenarios. We find that our method yields stable positionings reflective of manifesto positionings, both in scenarios with all tweets of candidates across years available and when only smaller subsets from shorter time periods are available. This indicates that it is possible to reliably analyze the relative positioning of actors without the need for manual annotation, even in the noisier context of social media.

Nov 14 (Thu) 14:00-15:30 - Jasmine

On the Rigour of Scientific Writing: Criteria, Analysis, and Insights

Joseph James, Chenghao Xiao, YUCHENG LI, Chenghua Lin

Rigour is crucial for scientific research as it ensures the reproducibility and validity of results and findings. Despite its importance, little work exists on modelling rigour computationally, and there is a lack of analysis on whether these criteria can effectively signal or measure the rigour of scientific papers in practice. In this paper, we introduce a bottom-up, data-driven framework to automatically identify and define rigour criteria and assess their relevance in scientific writing. Our framework includes rigour keyword extraction, detailed rigour definition generation, and salient criteria identification. Furthermore, our framework is domain-agnostic and can be tailored to the evaluation of scientific rigour for different areas accommodating the distinct salient criteria across fields. We conducted comprehensive experiments based on datasets collected from different domains (e.g. ICLR, ACL) to demonstrate the effectiveness of our framework in modelling rigour. In addition, we analyse linguist patterns of rigour, revealing that framing certainty is crucial for enhancing the perception of scientific rigour, while suggestion certainty and probability uncertainty diminish it.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Conversation Redirection in Mental Health Therapy

Vivian Nguyen, Sang Min Jung, Lillian Lee, Thomas D. Hull, Cristian Danescu-Niculescu-Mizil

Mental-health therapy involves a complex conversation flow in which patients and therapists continuously negotiate what should be talked about next. For example, therapists might try to shift the conversations direction to keep the therapeutic process on track and avoid stagnation, or patients might push the discussion towards issues they want to focus on. How do such patient and therapist redirections relate to the development and quality of their relationship? To answer this question, we introduce a probabilistic measure of the extent to which a certain utterance immediately redirects the flow of the conversation, accounting for both the intention and the actual realization of such a change. We apply this new measure to characterize the development of patient-therapist relationships over multiple sessions in a very large, widely-used online therapy platform. Our analysis reveals that (1) patient control of the conversations direction generally increases relative to that of the therapist as their relationship progresses; and (2) patients who have less control in the first few sessions are significantly more likely to eventually express dissatisfaction with their therapist and terminate the relationship.

Nov 14 (Thu) 14:00-15:30 - Jasmine

How Entangled is Factuality and Deception in German?

Aswathy Velutharambath, Amelie Wuehrl, Roman Klinger

The statement "The earth is flat" is factually inaccurate, but if someone truly believes and argues in its favor, it is not deceptive. Research on deception detection and fact checking often conflates factual accuracy with the truthfulness of statements. This assumption makes it difficult to (a) study subtle distinctions and interactions between the two and (b) gauge their effects on downstream tasks. The belief-based deception framework disentangles these properties by defining texts as deceptive when there is a mismatch between what people say and what they truly believe. In this study, we assess if presumed patterns of deception generalize to German language texts. We test the effectiveness of computational models in detecting deception using an established corpus of belief-based argumentation. Finally, we gauge the impact of deception on the downstream task of fact checking and explore if this property confounds verification models. Surprisingly, our analysis finds no correlation with established cues of deception. Previous work claimed that computational models can outperform humans in deception detection accuracy; however, our experiments show that both traditional and state-of-the-art models struggle with the task, performing no better than random guessing. For fact checking, we find that natural language inference-based verification performs worse on non-factual and deceptive content, while prompting large language models for the same task is less sensitive to these properties.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Do *they* mean 'us'? Interpreting Referring Expressions in Intergroup Bias

Venkata Subrahmanyam Govindarajan, Matianyu Zang, Kyle Mahowald, David Beaver, Junyi Jessy Li

The variations between in-group and out-group speech (intergroup bias) are subtle and could underlie many social phenomena like stereotype perpetuation and implicit bias. In this paper, we model intergroup bias as a tagging task on English sports comments from forums dedicated to fandom for NFL teams. We curate a dataset of over 6 million game-time comments from opposing perspectives (the teams in the game), each comment grounded in a non-linguistic description of the events that precipitated these comments (live win probabilities for each team). Expert and crowd annotations justify modeling the bias through tagging of implicit and explicit referring expressions and reveal the rich, contextual understanding of language and the world required for this task. For large-scale analysis of intergroup variation, we use LLMs for automated tagging, and discover that LLMs occasionally perform better when prompted with linguistic descriptions of the win probability at the time of the comment, rather than numerical probability. Further, large-scale tagging of comments using LLMs uncovers linear variations in the form of referent across win probabilities that distinguish in-group and out-group utterances.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Towards Effective Counter-Responses: Aligning Human Preferences with Strategies to Combat Online Trolling

Huije Lee, Hoyun Song, Jisu Shin, Sukmin Cho, SeungYoon Han, Jong C. Park

Trolling in online communities typically involves disruptive behaviors such as provoking anger and manipulating discussions, leading to a polarized atmosphere and emotional distress. Robust moderation is essential for mitigating these negative impacts and maintaining a healthy and constructive community atmosphere. However, effectively addressing trolls is difficult because their behaviors vary widely and require different response strategies (RSs) to counter them. This diversity makes it challenging to choose an appropriate RS for each specific situation. To address this challenge, our research investigates whether humans have preferred strategies tailored to different types of trolling behaviors. Our findings reveal a correlation between the types of trolling encountered and the preferred RS. In this paper, we introduce a methodology for generating counter-responses to trolls by recommending appropriate RSs, supported by a dataset aligning these strategies with human preferences across various troll contexts. The experimental results demonstrate that our proposed approach guides constructive discussion and reduces the negative effects of trolls, thereby enhancing the online community environment.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Communicate to Play: Pragmatic Reasoning for Efficient Cross-Cultural Communication

Isadora White, Sashrika Pandey, Michelle Pan

In this paper, we study how culture leads to differences in common ground and how this influences communication. During communication, cultural differences in common ground during communication may result in pragmatic failure and misunderstandings. We develop our method Rational Speech Acts for Cross-Cultural Communication (RSA+C3) to resolve cross-cultural differences in common ground. To measure the

success of our method, we study RSA+C3 in the collaborative referential game of Codenames Duet and show that our method successfully improves collaboration between simulated players of different cultures. Our contributions are threefold: (1) creating Codenames players using contrastive learning of an embedding space and LLM prompting that are aligned with human patterns of play, (2) studying culturally induced differences in common ground reflected in our trained models, and (3) demonstrating that our method RSA+C3 can ease cross-cultural communication in gameplay by inferring socio-cultural context from interaction.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Improving Quotation Attribution with Fictional Character Embeddings

Gaspard Michel, Elena V. Epure, Romain Hennequin, Christophe Cerisara

Humans naturally attribute utterances of direct speech to their speaker in literary works. When attributing quotes, we process contextual information but also access mental representations of characters that we build and revise throughout the narrative. Recent methods to automatically attribute such utterances have explored simulating human logic with deterministic rules or learning new implicit rules with neural networks when processing contextual information. However, these systems inherently lack character representations, which often leads to errors in more challenging examples of attribution: anaphoric and implicit quotes. In this work, we propose to augment a popular quotation attribution system, BookNLP, with character embeddings that encode global stylistic information of characters derived from an off-the-shelf stylometric model, Universal Authorship Representation (UAR). We create DramaCV, a corpus of English drama plays from the 15th to 20th century that we automatically annotate for Authorship Verification of fictional characters' utterances, and release two versions of UAR trained on DramaCV, that are tailored for literary characters analysis. Then, through an extensive evaluation on 28 novels, we show that combining BookNLP's contextual information with our proposed global character embeddings improves the identification of speakers for anaphoric and implicit quotes, reaching state-of-the-art performance. Code and data can be found at https://github.com/deezer/character_embeddings_qa.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

The Language of Trauma: Modeling Traumatic Event Descriptions Across Domains with Explainable AI

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, David Jurgens

Psychological trauma can manifest following various distressing events and is captured in diverse online contexts. However, studies traditionally focus on a single aspect of trauma, often neglecting the transferability of findings across different scenarios. We address this gap by training various language models with progressing complexity on trauma-related datasets, including genocide-related court data, a Reddit dataset on post-traumatic stress disorder (PTSD), counseling conversations, and Incel forum posts. Our results show that the fine-tuned RoBERTa model excels in predicting traumatic events across domains, slightly outperforming large language models like GPT-4. Additionally, SLALOM-feature scores and conceptual explanations effectively differentiate and cluster trauma-related language, highlighting different trauma aspects and identifying sexual abuse and experiences related to death as a common traumatic event across all datasets. This transferability is crucial as it allows for the development of tools to enhance trauma detection and intervention in diverse populations and settings.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks

Yun-Shiuin Chuang, Zach Studdiford, Krirk Nirunviroj, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, Timothy T. Rogers

Creating human-like large language model (LLM) agents is crucial for faithful social simulation. Having LLMs role-play based on demographic information sometimes improves human likeness but often does not. This study assessed whether LLM alignment with human behavior can be improved by integrating information from empirically-derived human belief networks. Using data from a human survey, we estimated a belief network encompassing 64 topics loading on nine non-overlapping latent factors. We then seeded LLM-based agents with an opinion on one topic, and assessed the alignment of its expressed opinions on remaining test topics with corresponding human data. Role-playing based on demographic information alone did not align LLM and human opinions, but seeding the agent with a single belief greatly improved alignment for topics related in the belief network, and not for topics outside the network. These results suggest a novel path for human-LLM belief alignment in work seeking to simulate and understand patterns of belief distributions in society.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Large Language Models for Propaganda Span Annotation

Maram Hasanain, Fatema Ahmad, Firoz Alam

The use of propagandistic techniques in online content has increased in recent years aiming to manipulate online audiences. Fine-grained propaganda detection and extraction of textual spans where propaganda techniques are used, are essential for more informed content consumption. Automatic systems targeting the task over lower resourced languages are limited, usually obstructed by lack of large scale training datasets. Our study investigates whether Large Language Models (LLMs), such as GPT-4, can effectively extract propagandistic spans. We further study the potential of employing the model to collect more cost-effective annotations. Finally, we examine the effectiveness of labels provided by GPT-4 in training smaller language models for the task. The experiments are performed over a large-scale in-house manually annotated dataset. The results suggest that providing more annotation context to GPT-4 within prompts improves its performance compared to human annotators. Moreover, when serving as an expert annotator (consolidator), the model provides labels that have higher agreement with expert annotators, and lead to specialized models that achieve state-of-the-art over an unseen Arabic testing set. Finally, our work is the first to show the potential of utilizing LLMs to develop annotated datasets for propagandistic spans detection task prompting it with annotations from human annotators with limited expertise. All scripts and annotations will be shared with the community.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Can Machines Resonate with Humans? Evaluating the Emotional and Empathetic Comprehension of LMs

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, Preslav Nakov

Empathy plays a pivotal role in fostering prosocial behavior, often triggered by the sharing of personal experiences through narratives. However, modeling empathy using NLP approaches remains challenging due to its deep interconnection with human interaction dynamics. Previous approaches, which involve fine-tuning language models (LMs) on human-annotated empathic datasets, have had limited success. In our pursuit of improving empathy understanding in LMs, we propose several strategies, including contrastive learning with masked LMs and supervised fine-tuning with large language models. While these methods show improvements over previous methods, the overall results remain unsatisfactory. To better understand this trend, we performed an analysis which reveals a low agreement among annotators. This lack of consensus hinders training and highlights the subjective nature of the task. We also explore the cultural impact on annotations. To study this, we meticulously collected story pairs in Urdu language and find that subjectivity in interpreting empathy among annotators appears to be independent of cultural background. Our systematic exploration of LMs' understanding of empathy reveals substantial opportunities for further investigation in both task formulation and modeling.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Are Large Language Models Consistent over Value-laden Questions?

Jared Moore, Tanvi Deshpande, Diyi Yang

Large language models (LLMs) appear to bias their survey answers toward certain values. Nonetheless, some argue that LLMs are too incon-

sistent to simulate particular values. Are they? To answer, we first define value consistency as the similarity of answers across 1) *paraphrases* of one question, 2) related questions under one *topic*, 3) multiple-choice and open-ended *use-cases* of one question, and 4) *multilingual* translations of a question to English, Chinese, German, and Japanese. We apply these measures to a few large, open LLMs including *llama-3*, as well as *gpt-4o*, using eight thousand questions spanning more than 300 topics. Unlike prior work, we find that *models are relatively consistent* across paraphrases, use-cases, translations, and within a topic. Still, some inconsistencies remain. Models are more consistent on uncontroversial topics (e.g., in the U.S., "Thanksgiving") than on controversial ones (e.g., "euthanasia"). Base models are both more consistent compared to fine-tuned models and are uniform in their consistency across topics, while fine-tuned models are more inconsistent about some topics (e.g., "euthanasia") than others (e.g., "Women's rights") like our human participants.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Extrinsic Evaluation of Cultural Competence in Large Language Models

Shaify Bhatt, Fernando Diaz

Productive interactions between diverse users and language technologies require outputs from the latter to be culturally relevant and sensitive. Prior works have evaluated models' knowledge of cultural norms, values, and artefacts, without considering how this knowledge manifests in downstream applications. In this work, we focus on extrinsic evaluation of cultural competence in two text generation tasks, open-ended question answering and story generation. We quantitatively and qualitatively evaluate model outputs when an explicit cue of culture, specifically nationality, is perturbed in the prompts. Although we find that model outputs do vary when varying nationalities and feature culturally relevant words, we also find weak correlations between text similarity of outputs for different countries and the cultural values of these countries. Finally, we discuss important considerations in designing comprehensive evaluation of cultural competence in user-facing tasks.

Nov 14 (Thu) 14:00-15:30 - Jasmine

SocialGaze: Improving the Integration of Human Social Norms in Large Language Models

Anvesh Rao Vijini, Rakesh R Menon, Shashank Srivastava, Snigdha Chaturvedi

While much research has explored enhancing the reasoning capabilities of large language models (LLMs) in the last few years, there is a gap in understanding the alignment of these models with social values and norms. We introduce the task of judging social acceptance. Social acceptance requires models to judge and rationalize the acceptability of people's actions in social situations. For example, is it socially acceptable for a neighbor to ask others in the community to keep their pets indoors at night? We find that LLMs' understanding of social acceptance is often misaligned with human consensus. To alleviate this, we introduce SocialGaze, a multi-step prompting framework, in which a language model verbalizes a social situation from multiple perspectives before forming a judgment. Our experiments demonstrate that the SocialGaze approach improves the alignment with human judgments by up to 11 F1 points with the GPT-3.5 model. We also identify biases and correlations in LLMs in assigning blame that is related to features such as the gender (males are significantly more likely to be judged unfairly) and age (LLMs are more aligned with humans for older narrators).

Nov 14 (Thu) 14:00-15:30 - Jasmine

ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanu Mitra, David Jurgens, Yulia Tsvetkov

This study introduces ValueScope, a framework leveraging language models to quantify social norms and values within online communities, grounded in social science perspectives on normative structures. We employ ValueScope to dissect and analyze linguistic and stylistic expressions across 13 Reddit communities categorized under gender, politics, science, and finance. Our analysis provides a quantitative foundation confirming that even closely related communities exhibit remarkably diverse norms. This diversity supports existing theories and adds a new dimension to understanding community interactions. ValueScope not only delineates differences in social norms but also effectively tracks their evolution and the influence of significant external events like the U.S. presidential elections and the emergence of new sub-communities. The framework thus highlights the pivotal role of social norms in shaping online interactions, presenting a substantial advance in both the theory and application of social norm studies in digital spaces.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Revealing Fine-Grained Values and Opinions in Large Language Models

Dustin Wright, Arnav Arora, Nadav Borenstein, Sriishi Yadav, Serge Belongie, Isabelle Augenstein

Uncovering latent values and opinions embedded in large language models (LLMs) can help identify biases and mitigate potential harm. Recently, this has been approached by prompting LLMs with survey questions and quantifying the stances in the outputs towards morally and politically charged statements. However, the stances generated by LLMs can vary greatly depending on how they are prompted, and there are many ways to argue for or against a given position. In this work, we propose to address this by analysing a large and robust dataset of 156k LLM responses to the 62 propositions of the Political Compass Test (PCT) generated by 6 LLMs using 420 prompt variations. We perform coarse-grained analysis of their generated stances and fine-grained analysis of the plain text justifications for those stances. For fine-grained analysis, we propose to identify tropes in the responses: semantically similar phrases that are recurrent and consistent across different prompts, revealing natural patterns in the text that a given LLM is prone to produce. We find that demographic features added to prompts significantly affect outcomes on the PCT, reflecting bias, as well as disparities between the results of tests when eliciting closed-form vs. open domain responses. Additionally, patterns in the plain text rationales via tropes show that similar justifications are repeatedly generated across models and prompts even with disparate stances.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Automated Tone Transcription and Clustering with Tone2Vec

Yi Yang, Yiming Wang, ZhiQiang Tang, Jiahong Yuan

Lexical tones play a crucial role in Sino-Tibetan languages. However, current phonetic fieldwork relies on manual effort, resulting in substantial time and financial costs. This is especially challenging for the numerous endangered languages that are rapidly disappearing, often compounded by limited funding. In this paper, we introduce pitch-based similarity representations for tone transcription, named Tone2Vec. Experiments on dialect clustering and variance show that Tone2Vec effectively captures fine-grained tone variation. Utilizing Tone2Vec, we develop the first automatic approach for tone transcription and clustering by presenting a novel representation transformation for transcriptions. Additionally, these algorithms are systematically integrated into an open-sourced and easy-to-use package, ToneLab, which facilitates automated fieldwork and cross-regional, cross-lexical analysis for tonal languages. Extensive experiments were conducted to demonstrate the effectiveness of our methods.

Nov 14 (Thu) 14:00-15:30 - Jasmine

SRAP-Agent: Simulating and Optimizing Scarce Resource Allocation Policy with LLM-based Agent

Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Qi Qi, Weiran Shen, Yankai Lin

Public scarce resource allocation plays a crucial role in economics as it directly influences the efficiency and equity in society. Traditional studies including theoretical model-based, empirical study-based and simulation-based methods encounter limitations due to the idealized assumption of complete information and individual rationality, as well as constraints posed by limited available data. In this work, we propose an innovative framework, SRAP-Agent, which integrates Large Language Models (LLMs) into economic simulations, aiming to bridge the gap be-

tween theoretical models and real-world dynamics. Using public housing allocation scenarios as a case study, we conduct extensive policy simulation experiments to verify the feasibility and effectiveness of the SRAP-Agent and employ the Policy Optimization Algorithm with certain optimization objectives. The source code can be found in https://github.com/jijiarui-cather/SRAPAgent_Framework.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Are Large Language Models (LLMs) Good Social Predictors?

Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, Hui Liu

With the recent advancement of Large Language Models (LLMs), efforts have been made to leverage LLMs in crucial social science study methods, including predicting human features of social life such as presidential voting. Existing works suggest that LLMs are capable of generating human-like responses. Nevertheless, it is unclear how well LLMs work and where the plausible predictions derive from. This paper critically examines the performance of LLMs as social predictors, pointing out the source of correct predictions and limitations. Based on the notion of mutability that classifies social features, we design three realistic settings and a novel social prediction task, where the LLMs make predictions with input features of the same mutability and accessibility with the response feature. We find that the promising performance achieved by previous studies is because of input shortcut features to the response, which are hard to capture in reality; the performance degrades dramatically to near-random after removing the shortcuts. With the comprehensive investigations on various LLMs, we reveal that LLMs struggle to work as expected on social prediction when given ordinarily available input features without shortcuts. We further investigate possible reasons for this phenomenon and suggest potential ways to enhance LLMs for social prediction.

Nov 14 (Thu) 14:00-15:30 - *Jasmine*

Cost-Efficient Subjective Task Annotation and Modeling through Few-Shot Annotator Adaptation

Preni Golazian, Alireza Salkhordeh Ziabari, Ali Omrani, Morteza Dehghani

In subjective NLP tasks, where a single ground truth does not exist, the inclusion of diverse annotators becomes crucial as their unique perspectives significantly influence the annotations. In realistic scenarios, the annotation budget often becomes the main determinant of the number of perspectives (i.e., annotators) included in the data and subsequent modeling. We introduce a novel framework for annotation collection and modeling in subjective tasks that aims to minimize the annotation budget while maximizing the predictive performance for each annotator. Our framework has a two-stage design: first, we rely on a small set of annotators to build a multitask model, and second, we augment the model for a new perspective by strategically annotating a few samples per annotator. To test our framework at scale, we introduce and release a unique dataset, Moral Foundations Subjective Corpus, of 2000 Reddit posts annotated by 24 annotators for moral sentiment. We demonstrate that our framework surpasses the previous SOTA in capturing the annotators' individual perspectives with as little as 25% of the original annotation budget on two datasets. Furthermore, our framework results in more equitable models, reducing the performance disparity among annotators.

Dialogue and Interactive Systems 3

Nov 14 (Thu) 14:00-15:30 - Room: Riverfront Hall

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Dialog2Flow: Pre-training Action-Driven Sentence Embeddings for Automatic Dialog Flow Extraction

Sergio Burdisso, Srikanth Madikeri, Petr Motlicek

Efficiently deriving structured workflows from unannotated dialogs remains an underexplored and formidable challenge in computational linguistics. Automating this process could significantly accelerate the manual design of workflows in new domains and enable the grounding of large language models in domain-specific flowcharts, enhancing transparency and controllability. In this paper, we introduce Dialog2Flow (D2F) embeddings, which differ from conventional sentence embeddings by mapping utterances to a latent space where they are grouped according to their communicative and informative functions (i.e., the actions they represent). D2F allows for modeling dialogs as continuous trajectories in a latent space with distinct action-related regions. By clustering D2F embeddings, the latent space is quantized, and dialogs can be converted into sequences of region/action IDs, facilitating the extraction of the underlying workflow. To pre-train D2F, we build a comprehensive dataset by unifying twenty task-oriented dialog datasets with normalized per-turn action annotations. We also introduce a novel soft contrastive loss that leverages the semantic information of these actions to guide the representation learning process, showing superior performance compared to standard supervised contrastive loss. Evaluation against various sentence embeddings, including dialog-specific ones, demonstrates that D2F yields superior qualitative and quantitative results across diverse domains.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

What are the Generator Preferences for End-to-end Task-Oriented Dialog System?

Wanshi Xu, Xianwei Zhuang, Zhanpeng Chen, Zhihong Zhu, Xinxin Cheng, Yuxian Zou

Fully end-to-end task-oriented dialogue (EToD) systems have shown excellent performance, which requires the ability to retrieve entities accurately for generation. Existing methods improve the accuracy of entity retrieval and construct data flows between retrieval results and response generator, achieving promising results. However, most of them suffer from the following issues: (1) The entity is retrieved by directly interacting with the context at a coarse-grained level, so the similarity score may be disturbed by irrelevant attributes; (2) The generator pays equal attention to retrieved entities and the context and does not learn the generation preferences for the current turn. In this paper, we propose a framework called Regulating Preferences of Generator (RPG) based on retrieval results, which includes a generator preference extractor, an entity retriever, and a generator with the gate-controlled preference regulator. The generator preference extractor not only improves the entity retriever by filtering the interference of irrelevant attributes but also provides more focused guidance to the generator by performing inter-turn attribute prediction. Experiments and analyses on three standard benchmarks show that our framework outperforms existing methods and improves the quality of the dialogue.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Neeko: Leveraging Dynamic LoRA for Efficient Multi-Character Role-Playing Agent

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, Liehuang Zhu

Large Language Models (LLMs) have revolutionized open-domain dialogue agents but encounter challenges in multi-character role-playing (MCRP) scenarios. To address the issue, we present Neeko, an innovative framework designed for efficient multiple characters imitation. Neeko employs a dynamic low-rank adapter (LoRA) strategy, enabling it to adapt seamlessly to diverse characters. Our framework breaks down the role-playing process into agent pre-training, multiple characters playing, and character incremental learning, effectively handling both seen and unseen roles. This dynamic approach, coupled with distinct LoRA blocks for each character, enhances Neeko's adaptability to unique attributes, personalities, and speaking patterns. As a result, Neeko demonstrates superior performance in MCRP over most existing methods, offering more engaging and versatile user interaction experiences.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, Mausam

Medical task-oriented dialogue systems can assist doctors by collecting patient medical history, aiding in diagnosis, or guiding treatment selection, thereby reducing doctor burnout and expanding access to medical services. However, doctor-patient dialogue datasets are not readily available, primarily due to privacy regulations. Moreover, existing datasets lack comprehensive annotations involving medical slots and their different attributes, such as symptoms and their onset, progression, and severity. These comprehensive annotations are crucial for accurate diagnosis. Finally, most existing datasets are non-English, limiting their utility for the larger research community. In response, we introduce MediTOD, a new dataset of doctor-patient dialogues in English for the medical history-taking task. Collaborating with doctors, we devise a questionnaire-based labeling scheme tailored to the medical domain. Then, medical professionals create the dataset with high-quality comprehensive annotations, capturing medical slots and their attributes. We establish benchmarks in supervised and few-shot settings on MediTOD for natural language understanding, policy learning, and natural language generation subtasks, evaluating models from both TOD and biomedical domains. We make MediTOD publicly available for future research.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MP2D: An Automated Topic Shift Dialogue Generation Framework Leveraging Knowledge Graphs

Yerin Hwang, Yongil Kim, Yunah Jang, Jeesoo Bang, Hyunkyoung Bae, Kyomin Jung

Despite advancements in on-topic dialogue systems, effectively managing topic shifts within dialogues remains a persistent challenge, largely attributed to the limited availability of training datasets. To address this issue, we propose Multi-Passage to Dialogue (MP2D), a data generation framework that automatically creates conversational question-answering datasets with natural topic transitions. By leveraging the relationships between entities in a knowledge graph, MP2D maps the flow of topics within a dialogue, effectively mirroring the dynamics of human conversation. It retrieves relevant passages corresponding to the topics and transforms them into dialogues through the passage-to-dialogue method. Through quantitative and qualitative experiments, we demonstrate MP2D's efficacy in generating dialogue with natural topic shifts. Furthermore, this study introduces a novel benchmark for topic shift dialogues, TS-WikiDialog. Utilizing the dataset, we demonstrate that even Large Language Models (LLMs) struggle to handle topic shifts in dialogue effectively, and we showcase the performance improvements of models trained on datasets generated by MP2D across diverse topic shift dialogue tasks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

A Survey of Ontology Expansion for Conversational Understanding

Jinggui Liang, Yuxia Wu, Yuan Fang, Hao Fei, Lizi Liao

In the rapidly evolving field of conversational AI, Ontology Expansion (OnExp) is crucial for enhancing the adaptability and robustness of conversational agents. Traditional models rely on static, predefined ontologies, limiting their ability to handle new and unforeseen user needs. This survey paper provides a comprehensive review of the state-of-the-art techniques in OnExp for conversational understanding. It categorizes the existing literature into three main areas: (1) New Intent Discovery, (2) New Slot-Value Discovery, and (3) Joint OnExp. By examining the methodologies, benchmarks, and challenges associated with these areas, we highlight several emerging frontiers in OnExp to improve agent performance in real-world scenarios and discuss their corresponding challenges. This survey aspires to be a foundational reference for researchers and practitioners, promoting further exploration and innovation in this crucial domain.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models

Chani Jung, Dongkwan Kim, Jihoo Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, Hyunwoo Kim

While humans naturally develop theory of mind (ToM), the capability to understand other people's mental states and beliefs, state-of-the-art large language models (LLMs) underperform on simple ToM benchmarks. We posit that we can extend our understanding of LLMs' ToM abilities by evaluating key human ToM precursors—perception inference and perception-to-belief inference—in LLMs. We introduce two datasets, Percept-ToMi and Percept-FANToM, to evaluate these precursory inferences for ToM in LLMs by annotating characters' perceptions on ToMi and FANToM, respectively. Our evaluation of eight state-of-the-art LLMs reveals that the models generally perform well in perception inference while exhibiting limited capability in perception-to-belief inference (e.g., lack of inhibitory control). Based on these results, we present PercepToM, a novel ToM method leveraging LLMs' strong perception inference capability while supplementing their limited perception-to-belief inference. Experimental results demonstrate that PercepToM significantly enhances LLM's performance, especially in false belief scenarios.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Learning from Relevant Subgoals in Successful Dialogs using Iterative Training for Task-oriented Dialog Systems

Magdalena Kaiser, Patrick Ernst, György Szávics

Task-oriented Dialog (ToD) systems have to solve multiple subgoals to accomplish user goals, whereas feedback is often obtained only at the end of the dialog. In this work, we propose SUIT (Subgoal-aware ITerative Training), an iterative training approach for improving ToD systems. We sample dialogs from the model we aim to improve and determine subgoals that contribute to dialog success using distant supervision to obtain high quality training samples. We show how this data improves supervised fine-tuning or, alternatively, preference learning results. Performance improves when applying these steps over several iterations: SUIT reaches new state-of-the-art performance on a popular ToD benchmark.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

NeBuLa: A discourse aware Minecraft Builder

Akhshay Chaturvedi, Kate Thompson, Nicholas Asher

When engaging in collaborative tasks, humans efficiently exploit the semantic structure of a conversation to optimize verbal and nonverbal interactions. But in recent "language to code" or "language to action" models, this information is lacking. We show how incorporating the prior discourse and nonlinguistic context of a conversation situated in a nonlinguistic environment can improve the "language to action" component of such interactions. We finetune an LLM to predict actions based on prior context; our model, Nebula, doubles the net-action F1 score over the baseline on this task of Jayannavar et al. (2020). We also investigate our model's ability to construct shapes and understand location descriptions using a synthetic dataset.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Contextualized Graph Representations for Generating Counter-Narrative against Hate Speech

Selene Baez Santamaría, Helena Gomez Adorno, Ilia Markov

Hate speech (HS) is a widely acknowledged societal problem with potentially grave effects on vulnerable individuals and minority groups. Developing counter-narratives (CNs) that confront biases and stereotypes driving hateful narratives is considered an impactful strategy. Current automatic methods focus on isolated utterances to detect and react to hateful content online, often omitting the conversational context where HS naturally occurs. In this work, we explore strategies for the incorporation of conversational history for CN generation, comparing text and graphical representations with varying degrees of context. Overall, automatic and human evaluations show that 1) contextualized

representations are comparable to those of isolated utterances, and 2) models based on graph representations outperform text representations, thus opening new research directions for future work.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Rewarding What Matters: Step-by-Step Reinforcement Learning for Task-Oriented Dialogue

Huifang Du, Shuqin Li, Minghao Wu, Xuejing Feng, Yuan-Fang Li, Haofen Wang

Reinforcement learning (RL) is a powerful approach to enhance task-oriented dialogue (TOD) systems. However, existing RL methods tend to mainly focus on generation tasks, such as dialogue policy learning (DPL) or response generation (RG), while neglecting dialogue state tracking (DST) for understanding. This narrow focus limits the systems to achieve globally optimal performance by overlooking the interdependence between understanding and generation. Additionally, RL methods face challenges with sparse and delayed rewards, which complicates training and optimization. To address these issues, we extend RL into both understanding and generation tasks by introducing step-by-step rewards throughout the token generation. The understanding reward increases as more slots are correctly filled in DST, while the generation reward grows with the accurate inclusion of user requests. Our approach provides a balanced optimization aligned with task completion. Experimental results demonstrate that our approach effectively enhances the performance of TOD systems and achieves new state-of-the-art results on three widely used datasets, including MultiWOZ2.0, MultiWOZ2.1, and In-Car. Our approach also shows superior few-shot ability in low-resource settings compared to current models.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Devil's Advocate: Anticipatory Reflection for LLM Agents

Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, Yang Li

In this work, we introduce a novel approach that equips LLM agents with introspection, enhancing consistency and adaptability in solving complex tasks. Our approach prompts LLM agents to decompose a given task into manageable subtasks (i.e., to make a plan), and to continuously introspect upon the suitability and results of their actions. We implement a three-fold introspective intervention: 1) anticipatory reflection on potential failures and alternative remedy before action execution, 2) post-action alignment with subtask objectives and backtracking with remedy to ensure utmost effort in plan execution, and 3) comprehensive review upon plan completion for future strategy refinement. By deploying and experimenting with this methodology—a zero-shot approach—within WebArena for practical tasks in web environments, our agent demonstrates superior performance with a success rate of 23.5% over existing zero-shot methods by 3.5%. The experimental results suggest that our introspection-driven approach not only enhances the agent's ability to navigate unanticipated challenges through a robust mechanism of plan execution, but also improves efficiency by reducing the number of trials and plan revisions by 45% needed to achieve a task.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Instruct, Not Assist: LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging

Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, Jiawei Han

Socratic questioning is an effective teaching strategy, encouraging critical thinking and problem-solving. The conversational capabilities of large language models (LLMs) show great potential for providing scalable, real-time student guidance. However, current LLMs often give away solutions directly, making them ineffective instructors. We tackle this issue in the code debugging domain with TreeInstruct, an Instructor agent guided by a novel state-space-based planning algorithm. TreeInstruct asks probing questions to help students independently identify and resolve errors. It estimates a student's conceptual and syntactical knowledge to dynamically construct a question tree based on their responses and current knowledge state, effectively addressing both independent and dependent mistakes concurrently in a multi-turn interaction setting. In addition to using an existing single-bug debugging benchmark, we construct a more challenging multi-bug dataset of 150 coding problems, incorrect solutions, and bug fixes—all carefully constructed and annotated by experts. Extensive evaluation shows TreeInstruct's state-of-the-art performance on both datasets, proving it to be a more effective instructor than baselines. Furthermore, a real-world case study with five students of varying skill levels further demonstrates TreeInstruct's ability to guide students to debug their code efficiently with minimal turns and highly Socratic questioning.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Ask-before-Plan: Proactive Language Agents for Real-World Planning

Xuan Zhang, Yang Deng, Zijeng Ren, See-Kiong Ng, Tat-Seng Chua

The evolution of large language models (LLMs) has enhanced the planning capabilities of language agents in diverse real-world scenarios. Despite these advancements, the potential of LLM-powered agents to comprehend ambiguous user instructions for reasoning and decision-making is still under exploration. In this work, we introduce a new task, Proactive Agent Planning, which requires language agents to predict clarification needs based on user-agent conversation and agent-environment interaction, invoke external tools to collect valid information, and generate a plan to fulfill the user's demands. To study this practical problem, we establish a new benchmark dataset, Ask-before-Plan. To tackle the deficiency of LLMs in proactive planning, we propose a novel multi-agent framework, Clarification-Execution-Planning (CEP), which consists of three agents specialized in clarification, execution, and planning. We introduce the trajectory tuning scheme for the clarification agent and static execution agent, as well as the memory recollection mechanism for the dynamic execution agent. Extensive evaluations and comprehensive analyses conducted on the Ask-before-Plan dataset validate the effectiveness of our proposed framework.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval

Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, Monica Lam

Persuasion plays a pivotal role in a wide range of applications from health intervention to the promotion of social good. Persuasive chatbots employed responsibly for social good can be an enabler of positive individual and social change. Existing methods rely on fine-tuning persuasive chatbots with task-specific training data which is costly, if not infeasible, to collect. Furthermore, they employ only a handful of pre-defined persuasion strategies. We propose PersuBot, a zero-shot chatbot based on Large Language Models (LLMs) that is factual and more persuasive by leveraging many more nuanced strategies. PersuBot uses an LLM to first generate a natural responses, from which the strategies used are extracted. To combat hallucination of LLMs, PersuBot replace any unsubstantiated claims in the response with retrieved facts supporting the extracted strategies. We applied our chatbot, PersuBot, to three significantly different domains needing persuasion skills: donation solicitation, recommendations, and health intervention. Our experiments on simulated and human conversations show that our zero-shot approach is more persuasive than prior work, while achieving factual accuracy surpassing state-of-the-art knowledge-oriented chatbots.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Soda-Eval: Open-Domain Dialogue Evaluation in the age of LLMs

John Mendonça, Isabel Trancoso, Alon Lavie

Although human evaluation remains the gold standard for open-domain dialogue evaluation, the growing popularity of automated evaluation using Large Language Models (LLMs) has also extended to dialogue. However, most frameworks leverage benchmarks that assess older chatbots on aspects such as fluency and relevance, which are not reflective of the challenges associated with contemporary models. In fact,

a qualitative analysis on Soda. (Kim et al., 2023), a GPT-3.5 generated dialogue dataset, suggests that current chatbots may exhibit several recurring issues related to coherence and commonsense knowledge, but generally produce highly fluent and relevant responses. Noting the aforementioned limitations, this paper introduces Soda-Eval, an annotated dataset based on Soda that covers over 120K turn-level assessments across 10K dialogues, where the annotations were generated by GPT-4. Using Soda-Eval as a benchmark, we then study the performance of several open-access instruction-tuned LLMs, finding that dialogue evaluation remains challenging. Fine-tuning these models improves performance over few-shot inferences, both in terms of correlation and explanation.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Mixed-Session Conversation with Egocentric Memory

Jihyoung Jang, Taeyoung Kim, Hyounghun Kim

Recently introduced dialogue systems have demonstrated high usability. However, they still fall short of reflecting real-world conversation scenarios. Current dialogue systems exhibit an inability to replicate the dynamic, continuous, long-term interactions involving multiple partners. This shortfall arises because there have been limited efforts to account for both aspects of real-world dialogues: deeply layered interactions over the long-term dialogue and widely expanded conversation networks involving multiple participants. As the effort to incorporate these aspects combined, we introduce Mixed-Session Conversation, a dialogue system designed to construct conversations with various partners in a multi-session dialogue setup. We propose a new dataset called MiSC to implement this system. The dialogue episodes of MiSC consist of 6 consecutive sessions, with four speakers (one main speaker and three partners) appearing in each episode. Also, we propose a new dialogue model with a novel memory management mechanism, called Egocentric Memory Enhanced Mixed-Session Conversation Agent (EMMA). EMMA collects and retains memories from the main speaker's perspective during conversations with partners, enabling seamless continuity in subsequent interactions. Extensive human evaluations validate that the dialogues in MiSC demonstrate a seamless conversational flow, even when conversation partners change in each session. EMMA trained with MiSC is also evaluated to maintain high memorability without contradiction throughout the entire conversation.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Stark: Social Long-Term Multi-Modal Conversation with Persona Commonsense Knowledge

Young-Jun Lee, Dokyong Lee, Jihyoung Yoon, Kyeong-Jin Oh, Byungssoo Ko, Jonghwan Hyeon, Ho-Jin Choi

Humans share a wide variety of images related to their personal experiences within conversations via instant messaging tools. However, existing works focus on (1) image-sharing behavior in singular sessions, leading to limited long-term social interaction, and (2) a lack of personalized image-sharing behavior. In this work, we introduce dataset, a large-scale long-term multi-modal dialogue dataset that covers a wide range of social personas in a multi-modality format, time intervals, and images. To construct datasetName automatically, we propose a novel multi-modal contextualization framework, frameworkName, that generates long-term multi-modal dialogues distilled from ChatGPT and our proposed planExecute image aligner. Using our dataset, we train a multi-modal conversation model, model 7B, which demonstrates impressive visual imagination ability. Furthermore, we demonstrate the effectiveness of our dataset in human evaluation. The code, dataset, and model will be publicly released after publication.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Problem-Oriented Segmentation and Retrieval: Case Study on Tutoring Conversations

Rose E Wang, Pawan Wirawarn, Kenny Lam, Omar Khattab, Dorothy Demsky

Many open-ended conversations (e.g., tutoring lessons or business meetings) revolve around pre-defined reference materials, like worksheets or meeting bullets. To provide a framework for studying such conversation structure, we introduce *Problem-Oriented Segmentation & Retrieval (POSR), the task of jointly breaking down conversations into segments and linking each segment to the relevant reference item. As a case study, we apply POSR to education where effectively structuring lessons around problems is critical yet difficult. We present *Lesson-Link*, the first dataset of real-world tutoring lessons, featuring 3,500 segments, spanning 24,300 minutes of instruction and linked to 116 SAT Math problems. We define and evaluate several joint and independent approaches for POSR, including segmentation (e.g., TextTiling), retrieval (e.g., ColBERT), and large language models (LLMs) methods. Our results highlight that modeling POSR as one joint task is essential: POSR methods outperform independent segmentation and retrieval pipelines by up to +76% on joint metrics and surpass traditional segmentation methods by up to +78% on segmentation metrics. We demonstrate POSR's practical impact on downstream education applications, deriving new insights on the language and time use in real-world lesson structures.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Prospector: Improving LLM Agents with Self-Asking and Trajectory Ranking

Byoungjin Kim, Youngsoo Jang, Lajanugen Logeswaran, Geon-Hyeong Kim, Yu Jin Kim, Honglak Lee, Moontae Lee

Large language models (LLMs) have shown the ability to solve complex decision-making tasks beyond natural language processing tasks. LLM agents based on few-shot in-context learning (ICL) achieve surprisingly high performance without training. Despite their simplicity and generalizability, ICL-based agents are limited in their ability to incorporate feedback from an environment. In this paper, we introduce Prospector, an LLM agent that consists of two complementary LLMs, an Actor and a Critic. To elicit better instruction-aligned actions from the LLM agent, we propose AskAct prompting that performs an additional self-asking step such as goal and progress checking before generating an action. Furthermore, to implicitly incorporate the environment feedback, we propose Trajectory Ranking that orders generated trajectories by predicting the expected total reward. Prospector encourages the LLM Actor to generate diverse (creative) trajectories, and harnesses the LLM Critic to select the most rewarding trajectory. On representative decision-making benchmark environments such as ALFWORLD and WebShop, we empirically demonstrate that Prospector can considerably increase the success rate of given tasks, while outperforming recent advancements such as ReAct and Reflexion.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Large Language Models Know What To Say But Not When To Speak

Muhammad Umar, Vasanth Sarathy, Jan Ruiter

Turn-taking is a fundamental mechanism in human communication that ensures smooth and coherent verbal interactions. Recent advances in Large Language Models (LLMs) have motivated their use in improving the turn-taking capabilities of Spoken Dialogue Systems (SDS), such as their ability to respond at appropriate times. However, existing models often struggle to predict opportunities for speaking, called Transition Relevance Places (TRPs) in natural, unscripted conversations, focusing only on turn-final TRPs and not within-turn TRPs. To address these limitations, we introduce a novel dataset of participant-labeled within-turn TRPs and use it to evaluate the performance of state-of-the-art LLMs in predicting opportunities for speaking. Our experiments reveal the current limitations of LLMs in modeling unscripted spoken interactions, highlighting areas for improvement and paving the way for more naturalistic dialogue systems.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

LLM generated responses to mitigate the impact of hate speech

Jakub Podolak, Szymon ukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bkowicz, Piotr Sankowski

In this study, we explore the use of Large Language Models (LLMs) to counteract hate speech. We conducted the first real-life A/B test assessing the effectiveness of LLM-generated counter-speech. During the experiment, we posted 753 automatically generated responses aimed

at reducing user engagement under tweets that contained hate speech toward Ukrainian refugees in Poland. Our work shows that interventions with LLM-generated responses significantly decrease user engagement, particularly for original tweets with at least ten views, reducing it by over 20%. This paper outlines the design of our automatic moderation system, proposes a simple metric for measuring user engagement and details the methodology of conducting such an experiment. We discuss the ethical considerations and challenges in deploying generative AI for discourse moderation.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Multi-trait User Simulation with Adaptive Decoding for Conversational Task Assistants

Rafael Ferreira, David Semedo, Joao Magalhaes

Conversational systems must be robust to user interactions that naturally exhibit diverse conversational traits. Capturing and simulating these diverse traits coherently and efficiently presents a complex challenge. This paper introduces Multi-Trait Adaptive Decoding (mTAD), a method that generates diverse user profiles at decoding-time by sampling from various trait-specific Language Models (LMS). mTAD provides an adaptive and scalable approach to user simulation, enabling the creation of multiple user profiles without the need for additional fine-tuning. By analyzing real-world dialogues from the Conversational Task Assistant (CTA) domain, we identify key conversational traits and developed a framework to generate profile-aware dialogues that enhance conversational diversity. Experimental results validate the effectiveness of our approach in modeling single-trait using specialized LMs, which can capture less common patterns, even in out-of-domain tasks. Furthermore, the results demonstrate that mTAD is a robust and flexible framework for combining diverse user simulators.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Auto-Intent: Automated Intent Discovery and Self-Exploration for Large Language Model Agents

Jaekeyeon Kim, Dong-Ki Kim, Lajanugen Logeswaran, Sungryull Sohn, Honglak Lee

In this paper, we introduce Auto-Intent, a method to adapt a pre-trained large language model (LLM) as an agent for a target domain without direct fine-tuning, where we empirically focus on web navigation tasks. Our approach first discovers the underlying intents from target domain demonstrations unsupervisedly, in a highly compact form (up to three words). With the extracted intents, we train our intent predictor to predict the next intent given the agents past observations and actions. In particular, we propose a self-exploration approach where top-k probable intent predictions are provided as a hint to the pre-trained LLM agent, which leads to enhanced decision-making capabilities. Auto-Intent substantially improves the performance of GPT-3.5, 4 and Llama-3.1-70B, 405B agents on the large-scale real-website navigation benchmarks from Mind2Web and online navigation tasks from WebArena with its cross-benchmark generalization from Mind2Web.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

EDEN: Empathetic Dialogues for English learning

Siyuan Li, Teresa Yu, Zhou Yu, Julia Hirschberg

Dialogue systems have been used as conversation partners in English learning, but few have studied whether these systems improve learning outcomes. Student passion and perseverance, or grit, has been associated with language learning success. Recent work establishes that as students perceive their English teachers to be more supportive, their grit improves. Hypothesizing that the same pattern applies to English-teaching chatbots, we create EDEN, a robust open-domain chatbot for spoken conversation practice that provides empathetic feedback. To construct EDEN, we first train a specialized spoken utterance grammar correction model and a high-quality social chit-chat conversation model. We then conduct a preliminary user study with a variety of strategies for empathetic feedback. Our experiment suggests that using adaptive empathetic feedback leads to higher *perceived affective support*. Furthermore, elements of perceived affective support positively correlate with student grit.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

TOOLVERIFIER: Generalization to New Tools via Self-Verification

Dheeraj Mekala, Jason E Weston, Jack Lanchantin, Roberta Raileanu, Maria Lomeli, Jingbo Shang, Jane Dwivedi-Yu

Teaching language models to use tools is an important milestone towards building general assistants, but remains an open problem. While there has been significant progress on learning to use specific tools via fine-tuning, language models still struggle with learning how to robustly use new tools from only a few demonstrations. In this work we introduce a self-verification method which distinguishes between close candidates by self-asking contrastive questions during (1) tool selection; and parameter generation. We construct synthetic, high-quality, self-generated data for this goal using Llama-2 70B, which we intend to release publicly. Extensive experiments on 4 tasks from the ToolBench benchmark, consisting of 17 unseen tools, demonstrate an average improvement of 22% over few-shot baselines, even in scenarios where the distinctions between candidate tools are finely nuanced.

Industry

Nov 14 (Thu) 14:00-15:30 - Room: Jasmine

Nov 14 (Thu) 14:00-15:30 - Jasmine

PRISM: A New Lens for Improved Color Understanding

Arjun Reddy Akula, Garima Pruthi, Inderjeet S Dhillon, Pradyumna Narayana, S Basu, Varun Jampani

While image-text pre-trained models, such as CLIP, have demonstrated impressive capabilities in learning robust text and image representations, a critical area for substantial improvement remains precise color understanding. In this paper, we address this limitation by introducing PRISM, a simple yet highly effective method that extends CLIP's capability to grasp the nuances of precise colors. PRISM seamlessly adapts to both recognized HTML colors and out-of-vocabulary RGB inputs through the utilization of our curated dataset of 100 image-text pairs, which can be effortlessly repurposed for fine-tuning with any desired color. Importantly, PRISM achieves these enhancements without compromising CLIP's performance on established benchmarks. Furthermore, we introduce a novel evaluation framework, ColorLens, featuring both seen and unseen test sets that can be readily repurposed to assess a model's precision in understanding precise colors. Our comprehensive evaluation and results demonstrate significant improvements over baseline models.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

SEED: Semantic Knowledge Transfer for Language Model Adaptation to Materials Science

Yeachan Kim, Jun-Hyung Park, Sungho Kim, Juhyeong Park, Sangyun Kim, Sangkeun Lee

Materials science is an interdisciplinary field focused on studying and discovering materials around us. However, due to the vast space of materials, datasets in this field are typically scarce and have limited coverage. This inherent limitation makes current adaptation methods less effective when adapting pre-trained language models (PLMs) to materials science, as these methods rely heavily on the frequency information from limited downstream datasets. In this paper, we propose Semantic Knowledge Transfer (SEED), a novel vocabulary expansion method to adapt the pre-trained language models for materials science. The core strategy of SEED is to transfer the materials' knowledge

of lightweight embeddings into the PLMs. To this end, we introduce knowledge bridge networks, which learn to transfer the latent knowledge of the materials embeddings into ones compatible with PLMs. By expanding the embedding layer of PLMs with these transformed embeddings, PLMs can comprehensively understand the complex terminology associated with materials science. We conduct extensive experiments across a broad range of materials-related benchmarks. Comprehensive evaluation results convincingly demonstrate that SEED mitigates the mentioned limitations of previous adaptation methods, showcasing the efficacy of transferring embedding knowledge into PLMs.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

TensorOpera Router: A Multi-Model Router for Efficient LLM Inference

Dimitris Stripelis, Zhaozhuo Xu, Zijian Hu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Jipeng Zhang, Tong Zhang, Salman Avestimehr, Chaoyang He

With the rapid growth of Large Language Models (LLMs) across various domains, numerous new LLMs have emerged, each possessing domain-specific expertise. This proliferation has highlighted the need for quick, high-quality, and cost-effective LLM query response methods. Yet, no single LLM exists to efficiently balance this trilemma. Some models are powerful but extremely costly, while others are fast and inexpensive but qualitatively inferior. To address this challenge, we present PolyRouter, a non-monolithic LLM querying system that seamlessly integrates various LLM experts into a single query interface and dynamically routes incoming queries to the most high-performant expert based on query's requirements. Through extensive experiments, we demonstrate that when compared to standalone expert models, PolyRouter improves query efficiency by up to 40%, and leads to significant cost reductions of up to 30%, while maintaining or enhancing model performance by up to 10%.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Personal Large Language Model Agents: A Case Study on Tailored Travel Planning

Harmapreet Singh, Nikhil Verma, Yixiao Wang, Manasa Bharadwaj, Homay Fashandi, Kevin Ferreira, Chul Lee

Large Language Models (LLMs) have made significant progress, becoming more autonomous and capable of handling real-world tasks through their access to tools, various planning strategies, and memory, referred to as LLM agents. One emerging area of focus is customizing these models to cater to individual user preferences, thereby shaping them into personal LLM agents. This work investigates how the user model, which encapsulates user-related information, preferences, and personal concepts, influences an LLM agent's planning and reasoning capabilities. We introduce a personalized version of TravelPlanner, called TravelPlanner+, and establish baselines for personal LLM agents. Our evaluation strategy contains an LLM-as-a-Judge component, which provides further in-depth insights into the decision-making process of a personal LLM agent by comparing generic and personal plans. Our findings reveal that while generic plans perform robustly, personal plans show marked improvement in relevance and suitability, with preference rates up to 74.4% on validation and 87.3% on the test set. These results highlight the potential of personal LLM agents to significantly enhance user satisfaction.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

TelBench: A Benchmark for Evaluating Telco-Specific Large Language Models

Sunwoo Lee, Dhammiko Arya, Seung-Mo Cho, Gyeong-Eun Han, Seokyung Hong, Wonbeom Jang, Seojin Lee, Sohee Park, Sereimony Sek, Injee Song, Sungbin Yoon, Eric Davis

The telecommunications industry, characterized by its vast customer base and complex service offerings, necessitates a high level of domain expertise and proficiency in customer service center operations. Consequently, there is a growing demand for Large Language Models (LLMs) to augment the capabilities of customer service representatives. This paper introduces a methodology for developing a specialized Telecommunications LLM (Telco LLM) designed to enhance the efficiency of customer service agents and promote consistency in service quality across representatives. We present the construction process of TelBench, a novel dataset created for performance evaluation of customer service expertise in the telecommunications domain. We also evaluate various LLMs and demonstrate the ability to benchmark both proprietary and open-source LLMs on predefined tasks, thereby establishing metrics that define telecommunications performance.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

SMARTCAL: An Approach to Self-Aware Tool-Use Evaluation and Calibration

Yuanhao Shen, Xiaodan Zhu, Lei Chen

The tool-use ability of Large Language Models (LLMs) has a profound impact on a wide range of applications. However, LLMs' self-awareness and self-control capability in appropriately using tools remains understudied. The problem is consequential as it alarms a potential risk of degraded performance and poses a threat to trustworthiness on the models. In this paper, we conduct a study on a family of state-of-the-art LLMs on three datasets with two mainstream tool-use frameworks. Our study reveals the tool-abuse behavior of LLMs, a tendency for models to misuse tools along with models' frequent overconfidence in tool choice. We also find that this is a common issue regardless of model capability. Accordingly, we propose a novel framework, SMARTCAL, to mitigate the observed issues, and our results show an average 8.6 percent increase in the QA performance in three testing datasets and 21.6 percent lower Expected Calibration Error (ECE) than existing methods.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Adapting LLMs for Structured Natural Language API Integration

Robin Chan, Katsiaryna Miryleva, Thomas Gschwind, Christoph Miksovic, Paolo Scotton, Enrico Toniatto, Abdel Labbi

Integrating APIs is crucial for enterprise systems, enabling seamless application interaction within workflows. However, the vast and diverse API landscape makes combining calls based on user intent a significant challenge. Existing methods rely on Named Entity Recognition (NER) and knowledge graphs, but struggle with control flow structures like conditionals and loops. We propose a novel framework that leverages the success of Large Language Models (LLMs) in code generation for natural language API integration. Our approach involves fine-tuning an LLM on automatically generated API flows derived from services' OpenAPI specifications. This aims to surpass NER-based methods and compare the effectiveness of different tuning strategies. Specifically, we investigate the impact of enforcing syntax through constrained generation or retrieval-augmented generation. To facilitate systematic comparison, we introduce targeted test suites that assess the generalization capabilities and ability of these approaches to retain structured knowledge. We expect to observe that fine-tuned LLMs can: (a) learn structural constraints implicitly during training, and (b) achieve significant improvements in both in-distribution and out-of-distribution performance.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

The State of the Art of Large Language Models on Chartered Financial Analyst Exams

Ethan Callanan, Mahmoud Matfouji, Mathieu Sibue, Antony Papadimitriou, Zhiqiang Ma, Xiaomo Liu, Xiaodan Zhu

The Chartered Financial Analyst (CFA) program is one of the most widely recognized financial certifications globally. In this work, we test a variety of state-of-the-art large language models (LLMs) on mock CFA exams to provide an overview of their financial analysis capabilities using the same evaluation standards applied for human professionals. We benchmark five leading proprietary models and eight open-source models on all three levels of the CFA through challenging multiple-choice and essay questions. We find that flagship proprietary models perform relatively well and can solidly pass levels I and II exams, but fail at level III due to essay questions. Open-source models generally fall short of estimated passing scores, but still show strong performance considering their size, cost, and availability advantages. We also find

that using textbook data helps bridge the gap between open-source and proprietary models to a certain extent, despite reduced gains in CFA levels II and III. By understanding the current financial analysis abilities of LLMs, we aim to guide practitioners on which models are best suited for enhancing automation in the financial industry.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Athena: Safe Autonomous Agents with Verbal Contrastive Learning

Tamanna Sadhu, Ali Pesaranghader, Yanan Chen, Dong Hoon Yi

Due to emergent capabilities, large language models (LLMs) have been utilized as language-based agents to perform a variety of tasks and make decisions with an increasing degree of autonomy. These autonomous agents can understand high-level instructions, interact with their environments, and execute complex tasks using a selection of tools available to them. As the capabilities of the agents expand, ensuring their safety and trustworthiness becomes more imperative. In this study, we introduce the Athena framework which leverages the concept of verbal contrastive learning where past safe and unsafe trajectories are used as in-context (contrastive) examples to guide the agent towards safety while fulfilling a given task. The framework also incorporates a critiquing mechanism to guide the agent to prevent risky actions at every step. Furthermore, due to the lack of existing benchmarks on the safety reasoning ability of LLM-based agents, we curate a set of 80 toolkits across 8 categories with 180 scenarios to provide a safety evaluation benchmark. Our experimental evaluation, with both closed- and open-source LLMs, indicates verbal contrastive learning and interaction-level critiquing improve the safety rate significantly.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Query-OPT: Optimizing Inference of Large Language Models via Multi-Query Instructions in Meeting Summarization

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, Shashi Bhushan Tn

This work focuses on the task of query-based meeting summarization in which the summary of a context (meeting transcript) is generated in response to a specific query. When using Large Language Models (LLMs) for this task, a new call to the LLM inference endpoint/API is required for each new query even if the context stays the same. However, repeated calls to the LLM inference endpoints would significantly increase the costs of using them in production, making LLMs impractical for many real-world use cases. To address this problem, in this paper, we investigate whether combining the queries for the same input context in a single prompt to minimize repeated calls can be successfully used in meeting summarization. In this regard, we conduct extensive experiments by comparing the performance of various popular LLMs: GPT-4, Gemini, Claude-3, LLaMA2, Mistral, Phi-3, and Qwen-2 in single-query and multi-query settings. We observe that the capability to reliably generate the response in the expected format is usually limited to closed-source LLMs, with most open-source LLMs lagging behind (except Mistral). We conclude that multi-query prompting could be useful to optimize the inference costs by significantly reducing calls to the inference endpoints/APIs for the task of meeting summarization.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Systematic Evaluation of Long-Context LLMs on Financial Concepts

Lavanya Gupta, Saket Sharma, Yiyun Zhao

Long-context large language models (LC LLMs) promise to increase reliability of LLMs in real-world tasks requiring processing and understanding of long input documents. However, this ability of LC LLMs to reliably utilize their growing context windows remains under investigation. In this work, we evaluate the performance of state-of-the-art GPT-4 suite of LC LLMs in solving a series of progressively challenging tasks, as a function of factors such as context length, task difficulty, and position of key information by creating a real world financial news dataset. Our findings indicate that LC LLMs exhibit brittleness at longer context lengths even for simple tasks, with performance deteriorating sharply as task complexity increases. At longer context lengths, these state-of-the-art models experience catastrophic failures in instruction following resulting in degenerate outputs. Our prompt ablations also reveal unfortunate continued sensitivity to both the placement of the task instruction in the context window as well as minor markdown formatting. Finally, we advocate for more rigorous evaluation of LC LLMs by employing holistic metrics such as F1 (rather than recall) and reporting confidence intervals, thereby ensuring robust and conclusive findings.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-Yi Lee, Yun-Nung Chen

Structured generation, the process of producing content in standardized formats like JSON and XML, is widely utilized in real-world applications to extract key output information from large language models (LLMs). This study investigates whether such constraints on generation space impact LLMs abilities, including reasoning and domain knowledge comprehension. Specifically, we evaluate LLMs performance when restricted to adhere to structured formats versus generating free-form responses across various common tasks. Surprisingly, we observe a significant decline in LLMs reasoning abilities under format restrictions. Furthermore, we find that stricter format constraints generally lead to greater performance degradation in reasoning tasks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

ASTRA: Automatic Schema Matching using Machine Translation

Tarang Chugh, Deepak Zambre

Many eCommerce platforms source product information from millions of sellers and manufacturers, each having their own proprietary schemas, and employ schema matching solutions to structure it to enable informative shopping experiences. Meanwhile, state-of-the-art machine translation techniques have demonstrated great success in building context-aware representations that generalize well to new languages with minimal training data. In this work, we propose modeling the schema matching problem as a neural machine translation task: given product context and an attribute-value pair from a source schema, the model predicts the corresponding attribute, if available, in the target schema. We utilize open-source seq2seq models, such as mT5 and mBART, fine-tuned on product attribute mappings to build a scalable schema matching framework. We demonstrate that our proposed approach achieves a significant performance boost (15% precision and 7% recall uplift) compared to the baseline system and can support new attributes with precision $\geq 95\%$ using only five labeled samples per attribute.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Visual Editing with LLM-based Tool Chaining: An Efficient Distillation Approach for Real-Time Applications

Oren Sultan, Alex Khasin, Guy Shiran, Asnat Greenstein-Messica, Dafna Shahaf

We present a practical distillation approach to fine-tune LLMs for invoking tools in real-time applications. We focus on visual editing tasks; specifically, we modify images and videos by interpreting user stylistic requests, specified in natural language ("golden hour"), using an LLM to select the appropriate tools and their parameters to achieve the desired visual effect. We found that proprietary LLMs such as GPT-3.5-Turbo show potential in this task, but their high cost and latency make them unsuitable for real-time applications. In our approach, we fine-tune a (smaller) student LLM with guidance from a (larger) teacher LLM and behavioral signals. We introduce offline metrics to evaluate student LLMs. Both online and offline experiments show that our student models manage to match the performance of our teacher model (GPT-3.5-Turbo), significantly reducing costs and latency. Lastly, we show that fine-tuning was improved by 25% in low-data regimes using augmentation.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MARCO: Multi-Agent Real-time Chat Orchestration

Anubhav Shrimai, Stanley Kanagaraj, Kriti Biswas, Swarnalatha Raghuraman, Anish Nediyanachath, Yi Zhang, Promod Yenigalla

Large language model advancements have enabled the development of multi-agent frameworks to tackle complex, real-world problems such as to automate workflows that require interactions with diverse tools, reasoning, and human collaboration. We present MARCO, a Multi-Agent Real-time Chat Orchestration framework for automating workflows using LLMs. MARCO addresses key challenges in utilizing LLMs for complex, multi-step task execution in a production environment. It incorporates robust guardrails to steer LLM behavior, validate outputs, and recover from errors that stem from inconsistent output formatting, function and parameter hallucination, and lack of domain knowledge. Through extensive experiments we demonstrate MARCO's superior performance with 94.48% and 92.74% accuracy on task execution for Digital Restaurant Service Platform conversations and Retail conversations datasets respectively along with 44.91% improved latency and 33.71% cost reduction in a production setting. We also report effects of guardrails in performance gain along with comparisons of various LLM models, both open-source and proprietary. The modular and generic design of MARCO allows it to be adapted for automating workflows across domains and to execute complex tasks through multi-turn interactions.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

ItiNera: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning

Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaojeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, Wei Ma

Citywalk, a recently popular form of urban travel, requires genuine personalization and understanding of fine-grained requests compared to traditional itinerary planning. In this paper, we introduce the novel task of Open-domain Urban Itinerary Planning (OUIP), which generates personalized urban itineraries from user requests in natural language. We then present ItiNera, an OUIP system that integrates spatial optimization with large language models to provide customized urban itineraries based on user needs. This involves decomposing user requests, selecting candidate points of interest (POIs), ordering the POIs based on cluster-aware spatial optimization, and generating the itinerary. Experiments on real-world datasets and the performance of the deployed system demonstrate our system's capacity to deliver personalized and spatially coherent itineraries compared to current solutions.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

RESTful-Llama: Connecting User Queries to RESTful APIs

Han Xu, Ruining Zhao, Jindong Wang, Haipeng Chen

Recent advancements in Large Language Models (LLMs) have showcased remarkable performance in zero-shot learning and reasoning tasks. However, integrating these models with external tools, which is crucial for real-world applications, remains challenging. We propose RESTful-Llama, a novel framework designed to enable Llama3 to transform natural language instructions into successful RESTful API calls. To enhance the fine-tuning process, we introduce DOC_Prompt, a method to augment datasets from public API documentation. RESTful-Llama stands out by empowering open-source LLMs to effectively leverage APIs and adapt to any REST API system. Extensive experiments demonstrate a 33.8% improvement in robustness and a 3.0x4 increase in efficiency compared to existing methods.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

GraphQL Query Generation: A Large Training and Benchmarking Dataset

Manish Keswani, Sambit Ghosh, Nitin Gupta, Shravana Chakraborty, Renuka Sindhwatta, Sameep Mehta, Carlos Eberhardt, Dan Debrunner

GraphQL is a powerful query language for APIs that allows clients to fetch precise data efficiently and flexibly, querying multiple resources with a single request. However, crafting complex GraphQL query operations can be challenging. Large Language Models (LLMs) offer an alternative by generating GraphQL queries from natural language, but they struggle due to limited exposure to publicly available GraphQL schemas, often resulting in invalid or suboptimal queries. Furthermore, no benchmark test data suite is available to reliably evaluate the performance of contemporary LLMs. To address this, we present a large-scale, cross-domain Text-to-GraphQL query operation dataset. The dataset includes 10,940 training triples spanning 185 cross-source data stores and 957 test triples over 14 data stores. Each triple consists of a GraphQL schema, GraphQL query operation, and corresponding natural language query. The dataset has been predominantly manually created, with natural language paraphrasing, and carefully validated, requiring approximately 1200 person-hours. In our evaluation, we tested 10 state-of-the-art LLMs using our test dataset. The best-performing model achieved an accuracy of only around 50% with one in-context few-shot example, underscoring the necessity for custom fine-tuning. To support further research and benchmarking, we are releasing the training and test datasets under the MIT License.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

INDUS: Effective and Efficient Language Models for Scientific Applications

Bishwaranjan Bhattacharjee, Aashika Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishant Panthee, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Mike Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain Costes, Sergiu Blanco-Cuarezas, Kelly Lockhart, Thomas Allen, Felix Grezes, Megan Ansell, Alberto Accomazzi, Yousef El-Kurdi, Davis Wertheimer, Birgit Pfitzmann, Cesar Berrospel Ramis, Michele Dolfi, Rafael Teixeira De Lima, Panagiota Vagenas, Surya Karthik Mukkavilli, Peter W. J. Staa, Sanaz Vahidinia, Ryan Mcgrannahan, Tsengdar Lee

Large language models (LLMs) trained on general domain corpora showed remarkable results on natural language processing (NLP) tasks. However, previous research demonstrated (LLMs) trained using domain-focused corpora perform better on specialized tasks. Inspired by this insight, we developed INDUS, a comprehensive suite of LLMs tailored for the closely-related domains of Earth science, biology, physics, heliophysics, planetary sciences and astrophysics and trained using curated scientific corpora drawn from diverse data sources. The suite of models include: (1) an encoder model trained using domain-specific vocabulary and corpora to address NLP tasks, (2) a contrastive-learning-based text embedding model trained using a diverse set of datasets to address information retrieval tasks and (3) smaller versions of these models created using knowledge distillation for applications which have latency or resource constraints. We also created three new scientific benchmark datasets, Climate-Change-NER (entity-recognition), Earth-QA (extractive QA) and Astro-IR (IR) to accelerate research in these multi-disciplinary fields. We show that our models outperform both general-purpose (RoBERTa) and domain-specific (SciBERT) encoders on these new tasks as well as existing tasks in the domains of interest. Furthermore, we demonstrate the use of these models in two industrial settings as a retrieval model for large-scale vector search applications and in an automatic content tagging system.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

SAAS: Solving Ability Amplification Strategy for Enhanced Mathematical Reasoning in Large Language Models

Hyeyoung Kim, Gyoyoung Gim, Yungi Kim, Byungjin Kim, Wonseok Lee, Chanjun Park

This study presents a novel learning approach designed to enhance both mathematical reasoning and problem-solving abilities of Large Language Models (LLMs). We focus on integrating the Chain-of-Thought (CoT) and the Program-of-Thought (PoT) learning, hypothesizing that prioritizing the learning of mathematical reasoning ability is helpful for the amplification of problem-solving ability. Thus, the initial learning with CoT is essential for solving challenging mathematical problems. To this end, we propose a sequential learning approach, named SAAS

(Solving Ability Amplification Strategy), which strategically transitions from CoT learning to PoT learning. Our empirical study, involving an extensive performance comparison using several benchmarks, demonstrates that our SaaS achieves state-of-the-art (SOTA) performance. The results underscore the effectiveness of our sequential learning approach, marking a significant advancement in the field of mathematical reasoning in LLMs.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Fusion-Eval: Integrating Assistant Evaluators with LLMs

Lei Shu, Nevan Wicheris, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, Lei Meng

Evaluating natural language generation (NLG) systems automatically poses significant challenges. Recent studies have employed large language models (LLMs) as reference-free metrics for NLG evaluation, enhancing adaptability to new tasks. However, these methods still show lower correspondence with human judgments compared to specialized neural evaluators. In this paper, we introduce "Fusion-Eval", an innovative approach that leverages LLMs to integrate insights from various assistant evaluators. The LLM is given the example to evaluate along with scores from the assistant evaluators. Each of these evaluators specializes in assessing distinct aspects of responses. Fusion-Eval achieves a 0.962 system-level Kendall-Tau correlation with humans on SummEval and a 0.744 turn-level Spearman correlation on TopicalChat, which is significantly higher than baseline methods. These results highlight Fusion-Eval's significant potential in the realm of natural language system evaluation.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Code Representation Pre-training with Complements from Program Executions

Jiaob Huang, Jianyu Zhao, Yuyang Rong, Yiwén Guo, Yifeng He, Hao Chen

Language models for natural language processing have been grafted onto programming language modeling for advancing code intelligence. Although it can be represented in the text format, code is syntactically more rigorous, as it is designed to be properly compiled or interpreted to perform a set of behaviors given any inputs. In this case, existing works benefit from syntactic representations to learn from code less ambiguously in forms of abstract syntax tree, control-flow graph, etc. However, programs with the same purpose can be implemented in various ways showing different syntactic representations, while the ones with similar implementations can have distinct behaviors. Though trivially demonstrated during executions, such semantics about functionality are challenging to be learned directly from code, especially in an unsupervised manner. Hence, in this paper, we propose FuzzPretrain to explore the dynamic information of programs revealed by their test cases and embed it into the feature representations of code as complements. The test cases are obtained with the assistance of a customized fuzzer and are only required during pre-training. FuzzPretrain yielded more than 6%/19% mAP improvements on code search over its masked language modeling counterparts trained with only source code and source code coupled with abstract syntax trees (ASTs), respectively. Our experiments show the benefits of learning discriminative code representations from FuzzPretrain.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

ScaleLLM: A Resource-Frugal LLM Serving Framework by Optimizing End-to-End Efficiency

Yuhang Yao, Han Jin, Atay Dilipbhai Shah, Shanshan Han, Zijian Hu, Yide Ran, Dimitris Stripelis, Zhaozhuo Xu, Salman Avestimehr, Chaoyang He

Large language models (LLMs) have surged in popularity and are extensively used in commercial applications, where the efficiency of model serving is crucial for the user experience. Most current research focuses on optimizing individual sub-procedures, e.g., local inference and communication, however, there is no comprehensive framework that provides a holistic system view for optimizing LLM serving in an end-to-end manner. In this work, we conduct a detailed analysis to identify major bottlenecks that impact end-to-end latency in LLM serving systems. Our analysis reveals that a comprehensive LLM serving endpoint must address a series of efficiency bottlenecks that extend beyond LLM inference. We then propose ScaleLLM, an optimized system for resource-efficient LLM serving. Our extensive experiments reveal that with 64 concurrent requests, ScaleLLM achieves a 4.3E speed up over vLLM and outperforms state-of-the-arts with 1.5E higher throughput.

Information Extraction 2

Nov 14 (Thu) 14:00-15:30 - Room: Riverfront Hall

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

In-context Contrastive Learning for Event Causality Identification

Wei Xiang, Bang Wang

Event Causality Identification (ECI) aims at determining the existence of a causal relation between two events. Although recent prompt learning-based approaches have shown promising improvements on the ECI task, their performance are often subject to the delicate design of multiple prompts and the positive correlations between the main task and derivative tasks. The in-context learning paradigm provides explicit guidance for label prediction in the prompt learning paradigm, alleviating its reliance on complex prompts and derivative tasks. However, it does not distinguish between positive and negative demonstrations for analogy learning. Motivated from such considerations, this paper proposes an `**I**_n,**C**_n**context **C**_n**contrastive **L**_n**learning` (ICCL) model that utilizes contrastive learning to enhance the effectiveness of both positive and negative demonstrations. Additionally, we apply contrastive learning to event pairs to better facilitate event causality identification. Our ICCL is evaluated on the widely used corpora, including the EventStoryLine and Causal-TimeBank, and results show significant performance improvements over the state-of-the-art algorithms.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Event Causality Identification with Synthetic Control

Haoyu Wang, Fengze Liu, Jiayao Zhang, Dan Roth, Kyle Richardson

Event causality identification (ECI), a process that extracts causal relations between events from text, is crucial for distinguishing causation from correlation. Traditional approaches to ECI have primarily utilized linguistic patterns and multi-hop relational inference, risking false causality identification due to informal usage of causality and specious graphical inference. In this paper, we adopt the Rubin Causal Model to identify event causality: given two temporally ordered events, we see the first event as the treatment and the second one as the observed outcome. Determining their causality involves manipulating the treatment and estimating the resultant change in the likelihood of the outcome. Given that it is only possible to implement manipulation conceptually in the text domain, as a work-around, we try to find a twin for the protagonist from existing corpora. This twin should have identical life experiences with the protagonist before the treatment but undergoes an intervention of treatment. However, the practical difficulty of locating such a match limits its feasibility. Addressing this issue, we use the synthetic control method to generate such a 'twin' from relevant historical data, leveraging text embedding synthesis and inversion techniques. This approach allows us to identify causal relations more robustly than previous methods, including GPT-4, which is demonstrated on a causality benchmark, COPES-hard.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Integrating Structural Semantic Knowledge for Enhanced Information Extraction Pre-training

Xiaoyang Yi, Yuru Bao, Jian Zhang, Yifang Qin, Faxin Lin

Information Extraction (IE), aiming to extract structured information from unstructured natural language texts, can significantly benefit from pre-trained language models. However, existing pre-training methods solely focus on exploiting the textual knowledge, relying extensively on annotated large-scale datasets, which is labor-intensive and thus limits the scalability and versatility of the resulting models. To address these issues, we propose SKIE, a novel pre-training framework tailored for IE that integrates structural semantic knowledge via contrastive learning, effectively alleviating the annotation burden. Specifically, SKIE utilizes Abstract Meaning Representation (AMR) as a low-cost supervision source to boost model performance without human intervention. By enhancing the topology of AMR graphs, SKIE derives high-quality cohesive subgraphs as additional training samples, providing diverse multi-level structural semantic knowledge. Furthermore, SKIE refines the graph encoder to better capture cohesive information and edge relation information, thereby improving the pre-training efficacy. Extensive experimental results demonstrate that SKIE outperforms state-of-the-art baselines across multiple IE tasks and showcases exceptional performance in few-shot and zero-shot settings.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation

Wenhuo Huang, Zhouhong Gu, Chenghao Peng, Jiaqing Liang, Zhihua Li, Yanghua Xiao, liqian wen, Zulong Chen

Web scraping is a powerful technique that extracts data from websites, enabling automated data collection, enhancing data analysis capabilities, and minimizing manual data entry efforts. Existing methods, wrappers-based methods suffer from limited adaptability and scalability when faced with a new website, while language agents, empowered by large language models (LLMs), exhibit poor reusability in diverse web environments. In this work, we introduce the paradigm of generating web scrapers with LLMs and propose AutoScraper, a two-stage framework that can handle diverse and changing web environments more efficiently. AutoScraper leverages the hierarchical structure of HTML and similarity across different web pages for generating web scrapers. Besides, we propose a new executability metric for better measuring the performance of web scraper generation tasks. We conduct comprehensive experiments with multiple LLMs and demonstrate the effectiveness of our framework. Our work is now open-source.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

LogicST: A Logical Self-Training Framework for Document-Level Relation Extraction with Incomplete Annotations

Shengda Fan, Yanling Wang, Shasha Mo, Jianwei Niu

Document-level relation extraction (DocRE) aims to identify relationships between entities within a document. Due to the vast number of entity pairs, fully annotating all fact triplets is challenging, resulting in datasets with numerous false negative samples. Recently, self-training-based methods have been introduced to address this issue. However, these methods are purely black-box and sub-symbolic, making them difficult to interpret and prone to overlooking symbolic interdependences between relations. To remedy this deficiency, our insight is that symbolic knowledge, such as logical rules, can be used as diagnostic tools to identify conflicts between pseudo-labels. By resolving these conflicts through logical diagnoses, we can correct erroneous pseudo-labels, thus enhancing the training of neural models. To achieve this, we propose **LogicST**, a neural-logic self-training framework that iteratively resolves conflicts and constructs the minimal diagnostic set for updating models. Extensive experiments demonstrate that LogicST significantly improves performance and outperforms previous state-of-the-art methods. For instance, LogicST achieves an increase of **+7.94%** in F1 score compared to CAST (Tan et al., 2023a) on the DocRED benchmark (Yao et al., 2019). Additionally, LogicST is more time-efficient than its self-training counterparts, requiring only **10%** of the training time of CAST.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards

Furkan ahinuc, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, Iryna Gurevych

Scientific leaderboards are standardized ranking systems that facilitate evaluating and comparing competitive methods. Typically, a leaderboard is defined by a task, dataset, and evaluation metric (TDM) triple, allowing objective performance assessment and fostering innovation through benchmarking. However, the exponential increase in publications has made it infeasible to construct and maintain these leaderboards manually. Automatic leaderboard construction has emerged as a solution to reduce manual labor. Existing datasets for this task are based on the community-contributed leaderboards without additional curation. Our analysis shows that a large portion of these leaderboards are incomplete, and some of them contain incorrect information. In this work, we present SciLead, a manually-curated Scientific Leaderboard dataset that overcomes the aforementioned problems. Building on this dataset, we propose three experimental settings that simulate real-world scenarios where TDM triples are fully defined, partially defined, or undefined during leaderboard construction. While previous research has only explored the first setting, the latter two are more representative of real-world applications. To address these diverse settings, we develop a comprehensive LLM-based framework for constructing leaderboards. Our experiments and analysis reveal that various LLMs often correctly identify TDM triples while struggling to extract result values from publications. We make our code and data publicly available.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Exploring Nested Named Entity Recognition with Large Language Models: Methods, Challenges, and Insights

Hongjin Kim, Jai-Eun Kim, Harksoo Kim

Nested Named Entity Recognition (NER) poses a significant challenge in Natural Language Processing (NLP), demanding sophisticated techniques to identify entities within entities. This research investigates the application of Large Language Models (LLMs) to nested NER, exploring methodologies from prior work and introducing specific reasoning techniques and instructions to improve LLM efficacy. Through experiments conducted on the ACE 2004, ACE 2005, and GENIA datasets, we evaluate the impact of these approaches on nested NER performance. Results indicate that output format critically influences nested NER performance, methodologies from previous works are less effective, and our nested NER-tailored instructions significantly enhance performance. Additionally, we find that label information and descriptions of nested cases are crucial in eliciting the capabilities of LLMs for nested NER, especially in specific domains (i.e., the GENIA dataset). However, these methods still do not outperform BERT-based models, highlighting the ongoing need for innovative approaches in nested NER with LLMs.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Jellyfish: Instruction-Tuning Local Large Language Models for Data Preprocessing

Haochen Zhang, Yuyang Dong, Chuan Xiao, Masafumi Oyamada

This paper explores the utilization of LLMs for data preprocessing (DP), a crucial step in the data mining pipeline that transforms raw data into a clean format. We instruction-tune local LLMs as universal DP task solvers that operate on a local, single, and low-priced GPU, ensuring data security and enabling further customization. We select a collection of datasets across four representative DP tasks and construct instruction data using data configuration, knowledge injection, and reasoning data distillation techniques tailored to DP. By tuning Mistral-7B, Llama 3-8B, and OpenOrca-Platypus2-13B, our models, Jellyfish-7B/8B/13B, deliver competitiveness compared to GPT-3.5/4 models and strong generalizability to unseen tasks while barely compromising the base models' abilities in NLP tasks. Meanwhile, Jellyfish offers enhanced reasoning capabilities compared to GPT-3.5. Our models are available at: <https://huggingface.co/NECOUDBFM/JellyfishOur>

dataset is available at: <https://huggingface.co/datasets/NECOUDBFM/Jellyfish-Instruct>

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction

Bowen Zhang, Harold Soh

In this work, we are interested in automated methods for knowledge graph creation (KGC) from input text. Progress on large language models (LLMs) has prompted a series of recent works applying them to KGC, e.g., via zero/few-shot prompting. Despite successes on small domain-specific datasets, these models face difficulties scaling up to text common in many real-world applications. A principal issue is that, in prior methods, the KG schema has to be included in the LLM prompt to generate valid triplets; larger and more complex schemas easily exceed the LLMs context window length. Furthermore, there are scenarios where a fixed pre-defined schema is not available and we would like the method to construct a high-quality KG with a succinct self-generated schema. To address these problems, we propose a three-phase framework named Extract-Define-Canonicalize (EDC): open information extraction followed by schema definition and post-hoc canonicalization. EDC is flexible in that it can be applied to settings where a pre-defined target schema is available and when it is not; in the latter case, it constructs a schema automatically and applies self-canonicalization. To further improve performance, we introduce a trained component that retrieves schema elements relevant to the input text; this improves the LLMs extraction performance in a retrieval-augmented generation-like manner. We demonstrate on three KGC benchmarks that EDC is able to extract high-quality triplets without any parameter tuning and with significantly larger schemas compared to prior works. Code for EDC is available at <https://github.com/clear-nus/edc>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

SPEED++: A Multilingual Event Extraction Framework for Epidemic Prediction and Preparedness

Tanmay Parekh, Jeffrey Kwan, Jiuru Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, Nanyun Peng

Social media is often the first place where communities discuss the latest societal trends. Prior works have utilized this platform to extract epidemic-related information (e.g. infections, preventive measures) to provide early warnings for epidemic prediction. However, these works only focused on English posts, while epidemics can occur anywhere in the world, and early discussions are often in the local, non-English languages. In this work, we introduce the first multilingual Event Extraction (EE) framework SPEED++ for extracting epidemic event information for any disease and language. To this end, we extend a previous epidemic ontology with 20 argument roles; and curate our multilingual EE dataset SPEED++ comprising 5.1K tweets in four languages for four diseases. Annotating data in every language is infeasible; thus we develop zero-shot cross-lingual cross-disease models (i.e., training only on English COVID data) utilizing multilingual pre-training and show their efficacy in extracting epidemic-related events for 65 diverse languages across different diseases. Experiments demonstrate that our framework can provide epidemic warnings for COVID-19 in its earliest stages in Dec 2019 (3 weeks before global discussions) from Chinese Weibo posts without any training in Chinese. Furthermore, we exploit our framework's argument extraction capabilities to aggregate community epidemic discussions like symptoms and cure measures, aiding misinformation detection and public attention monitoring. Overall, we lay a strong foundation for multilingual epidemic preparedness.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Grasping the Essentials: Tailoring Large Language Models for Zero-Shot Relation Extraction

Size Zhou, Yu Meng, Bowen Jin, Jiawei Han

Relation extraction (RE) aims to identify semantic relationships between entities within text. Despite considerable advancements, existing models predominantly require extensive annotated training data, which is both costly and labor-intensive to collect. Moreover, these models often struggle to adapt to new or unseen relations. Few-shot learning, aiming to lessen annotation demands, typically provides incomplete and biased supervision for target relations, leading to degraded and unstable performance. To accurately and explicitly describe relation semantics while minimizing annotation demands, we explore the definition only zero-shot RE setting where only relation definitions expressed in natural language are used to train a RE model. We introduce REPAL, comprising three stages: (1) We leverage large language models (LLMs) to generate initial seed instances from relation definitions and an unlabeled corpus. (2) We fine-tune a bidirectional Small Language Model (SLM) with initial seeds to learn relations for the target domain. (3) We expand pattern coverage and mitigate bias from initial seeds by integrating feedback from the SLMs predictions on the unlabeled corpus and the synthesis history. To accomplish this, we leverage the multi-turn conversation ability of LLMs to generate new instances in follow-up dialogues, informed by both the feedback and synthesis history. Studies reveal that definition-oriented seed synthesis enhances pattern coverage whereas indiscriminately increasing seed quantity leads to performance saturation. Experiments on two datasets show REPAL significantly improved cost-effective zero-shot performance by large margins.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Preserving Generalization of Language models in Few-shot Continual Relation Extraction

Quyen Tran, Nguyen Xuan Thanh, Nguyen Hoang Anh, Nam Le Hai, Trung Le, Linh Van Ngo, Thien Huu Nguyen

Few-shot Continual Relations Extraction (FCRE) is an emerging area of study where models can sequentially integrate knowledge from new relations with limited labeled data while circumventing catastrophic forgetting and preserving prior knowledge from pre-trained backbones. In this work, we introduce a novel method that leverages often-discarded language model heads. By employing these components via a mutual information maximization strategy, our approach helps maintain prior knowledge from the pre-trained backbone and strategically aligns the primary classification head, thereby enhancing model performance. Furthermore, we explore the potential of Large Language Models (LLMs), renowned for their wealth of knowledge, in addressing FCRE challenges. Our comprehensive experimental results underscore the efficacy of the proposed method and offer valuable insights for future work.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Topic-Oriented Open Relation Extraction with A Priori Seed Generation

Linyi Ding, Jinfeng Xiao, Sizhe Zhou, Chaoli Yang, Jiawei Han

The field of open relation extraction (ORE) has recently observed significant advancement thanks to the growing capability of large language models (LLMs). Nevertheless, challenges persist when ORE is performed on specific topics. Existing methods give sub-optimal results in five dimensions: factuality, topic relevance, informativeness, coverage, and uniformity. To improve topic-oriented ORE, we propose a zero-shot approach called PriORE: Open Relation Extraction with a Priori seed generation. PriORE leverages the built-in knowledge of LLMs to maintain a dynamic seed relation dictionary for the topic. The dictionary is initialized by seed relations generated from topic-relevant entity types and expanded during contextualized ORE. PriORE then reduces the randomness in generative ORE by converting it to a more robust relation classification task. Experiments show the approach empowers better topic-oriented control over the generated relations and thus improves ORE performance along the five dimensions, especially on specialized and narrow topics.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Where am I? Large Language Models Wandering between Semantics and Structures in Long Contexts

Seonmin Koo, Jinsung Kim, YoungJoon Jang, Chanjun Park, Heutseok Lim

As the utilization of Large Language Models (LLMs) becomes more widespread, there is a growing demand for their ability to handle more complex and longer external knowledge across various use cases. Most existing evaluations of the open-ended question answering (ODQA) task, which necessitates the use of external knowledge, focus solely on whether the model provides the correct answer. However, even when

LLMs answer correctly, they often fail to provide an obvious source for their responses. Therefore, it is necessary to jointly evaluate and verify the correctness of the answers and the appropriateness of grounded evidence in complex external contexts. To address this issue, we examine the phenomenon of discrepancies in abilities across two distinct tasksQA and evidence selection when performed simultaneously, from the perspective of task alignment. To verify LLMs' task alignment, we introduce a verification framework and resources considering both semantic relevancy and structural diversity of the given long context knowledge. Through extensive experiments and detailed analysis, we provide insights into the task misalignment between QA and evidence selection. Our code and resources will be available upon acceptance.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Weak Reward Model Transforms Generative Models into Robust Causal Event Extraction Systems

Italo Luís da Silva, Hangi Yan, Lin Gui, Yulan He

The inherent ambiguity of cause and effect boundaries poses a challenge in evaluating causal event extraction tasks. Traditional metrics like Exact Match and BertScore poorly reflect model performance, so we trained evaluation models to approximate human evaluation, achieving high agreement. We used them to perform Reinforcement Learning with extraction models to align them with human preference, prioritising semantic understanding. We successfully explored our approach through multiple datasets, including transferring an evaluator trained on one dataset to another as a way to decrease the reliance on human-annotated data. In that vein, we also propose a weak-to-strong supervision method that uses a fraction of the annotated data to train an evaluation model while still achieving high performance in training an RL model.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups

Rzvan-Alexandru Smdu, David-Gabriel ION, Dumitru-Clementin Cercel, Florin Pop, Mihaela-Claudia Cercel

Complex Word Identification (CWI) is an essential step in the lexical simplification task and has recently become a task on its own. Some variations of this binary classification task have emerged, such as lexical complexity prediction (LCP) and complexity evaluation of multi-word expressions (MWE). Large language models (LLMs) recently became popular in the Natural Language Processing community because of their versatility and capability to solve unseen tasks in zero/few-shot settings. Our work investigates LLM usage, specifically open-source models such as Llama 2, Llama 3, and Vicuna v1.5, and closed-source, such as ChatGPT-3.5-turbo and GPT-4o, in the CWI, LCP, and MWE settings. We evaluate zero-shot, few-shot, and fine-tuning settings and show that LLMs struggle in certain conditions or achieve comparable results against existing methods. In addition, we provide some views on meta-learning combined with prompt learning. In the end, we conclude that the current state of LLMs cannot or barely outperform existing methods, which are usually much smaller.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Seg2Act: Global Context-aware Action Generation for Document Logical Structuring

Zichao Li, Shaojie He, Meng Liao, Xuantran Luong, Yaojie Lu, Hongyu Lin, Yanxiong Lu, Xianpei Han, Le Sun

Document logical structuring aims to extract the underlying hierarchical structure of documents, which is crucial for document intelligence. Traditional approaches often fall short in handling the complexity and the variability of lengthy documents. To address these issues, we introduce Seg2Act, an end-to-end, generation-based method for document logical structuring, revisiting logical structure extraction as an action generation task. Specifically, given the text segments of a document, Seg2Act iteratively generates the action sequence via a global context-aware generative model, and simultaneously updates its global context and current logical structure based on the generated actions. Experiments on ChCatExt and HierDoc datasets demonstrate the superior performance of Seg2Act in both supervised and transfer learning settings.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Learning to Generate Rules for Realistic Few-Shot Relation Classification: An Encoder-Decoder Approach

Mayank Singh, Eduardo Blanco

We propose a neuro-symbolic approach for realistic few-shot relation classification via rules. Instead of building neural models to predict relations, we design them to output straightforward rules that can be used to extract relations. The rules are generated using custom T5-style Encoder-Decoder Language Models. Crucially, our rules are fully interpretable and pliable (i.e., humans can easily modify them to boost performance). Through a combination of rules generated by these models along with a very effective, novel baseline, we demonstrate a few-shot relation-classification performance that is comparable to or stronger than the state of the art on the Few-Shot TACRED and NYT29 benchmarks while increasing interpretability and maintaining pliability.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MMUTF: Multimodal Multimedia Event Argument Extraction with Unified Template Filling

Philipp Seehuber, Dominik Wagner, Korbinian Riedhammer

With the advancement of multimedia technologies, news documents and user-generated content are often represented as multiple modalities, making Multimedia Event Extraction (MEE) an increasingly important challenge. However, recent MEE methods employ weak alignment strategies and data augmentation with simple classification models, which ignore the capabilities of natural language-formulated event templates for the challenging Event Argument Extraction (EAE) task. In this work, we focus on EAE and address this issue by introducing a unified template filling model that connects the textual and visual modalities via textual prompts. This approach enables the exploitation of cross-ontology transfer and the incorporation of event-specific semantics. Experiments on the M2E2 benchmark demonstrate the effectiveness of our approach. Our system surpasses the current SOTA on textual EAE by +7% F1, and performs generally better than the second-best systems for multimedia EAE.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Temporal Cognitive Tree: A Hierarchical Modeling Approach for Event Temporal Relation Extraction

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, Jingyao Tang

Understanding and analyzing event temporal relations is a crucial task in Natural Language Processing (NLP). This task, known as Event Temporal Relation Extraction (ETRE), aims to identify and extract temporal connections between events in text. Recent studies focus on locating the relative position of event pairs on the timeline by designing logical expressions or auxiliary tasks to predict their temporal occurrence. Despite these advances, this modeling approach neglects the multidimensional information in temporal relation and the hierarchical process of reasoning. In this study, we propose a novel hierarchical modeling approach for this task by introducing a Temporal Cognitive Tree (TCT) that mimics human logical reasoning. Additionally, we also design a integrated model incorporating prompt optimization and deductive reasoning to exploit multidimensional supervised information. Extensive experiments on TB-Dense and MATRES datasets demonstrate that our approach outperforms existing methods.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Refiner: Restructure Retrieved Content Efficiently to Advance Question-Answering Capabilities

Zhonghao Li, Xuming Hu, Aiwei Liu, Kening Zheng, Sirui Huang, Hui Xiong

Large Language Models (LLMs) are limited by their parametric knowledge, leading to hallucinations in knowledge-extensive tasks. To address this, Retrieval-Augmented Generation (RAG) incorporates external document chunks to expand LLM knowledge. Furthermore, com-

pressing information from document chunks through extraction or summarization can improve LLM performance. Nonetheless, LLMs still struggle to notice and utilize scattered key information, a problem known as the “lost-in-the-middle” syndrome. Therefore, we typically need to restructure the content for LLM to recognize the key information. We propose *Refiner*, an end-to-end extract-and-restructure paradigm that operates in the post-retrieval process of RAG. *Refiner* leverages a single decoder-only LLM to adaptively extract query-relevant contents verbatim along with the necessary context, and section them based on their interconnectedness, thereby highlights information distinction, and aligns downstream LLMs with the original context effectively. Experiments show that a trained *Refiner* (with 7B parameters) exhibits significant gain to downstream LLM in improving answer accuracy, and outperforms other state-of-the-art advanced RAG and concurrent compressing approaches on various single-hop and multi-hop QA tasks. Notably, *Refiner* achieves a 80.5% tokens reduction and a 1.6-7.0% improvement margin in multi-hop tasks compared to the next best solution. *Refiner* is a plug-and-play solution that can be seamlessly integrated with RAG systems, facilitating its application across diverse open-source frameworks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Scalable and Domain-General Abstractive Proposition Segmentation

Mohammad Javad Hosseini, Yang Gao, Tim Baumgärtner, Alex Fabrikant, Reinald Kim Amplayo

Segmenting text into fine-grained units of meaning is important to a wide range of NLP applications. The default approach of segmenting text into sentences is often insufficient, especially since sentences are usually complex enough to include multiple units of meaning that merit separate treatment in the downstream task. We focus on the task of abstractive proposition segmentation (APS): transforming text into simple, self-contained, well-formed sentences. Several recent works have demonstrated the utility of proposition segmentation with few-shot prompted LLMs for downstream tasks such as retrieval-augmented grounding and fact verification. However, this approach does not scale to large amounts of text and may not always extract all the facts from the input text. In this paper, we first introduce evaluation metrics for the task to measure several dimensions of quality. We then propose a scalable, yet accurate, proposition segmentation model. We model proposition segmentation as a supervised task by training LLMs on existing annotated datasets and show that training yields significantly improved results. We further show that by using the fine-tuned LLMs (Gemini Pro and Gemini Ultra) as teachers for annotating large amounts of multi-domain synthetic distillation data, we can train smaller student models (Gemini 1.2B and 7B) with results similar to the teacher LLMs. We then demonstrate that our technique leads to effective domain generalization, by annotating data in two domains outside the original training data and evaluating on them. Finally, as a key contribution of the paper, we share an easy-to-use API for NLP practitioners to use.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

A Survey on Open Information Extraction from Rule-based Model to Large Language Model

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, Yue Zhang

Open Information Extraction (OpenIE) represents a crucial NLP task aimed at deriving structured information from unstructured text, unrestricted by relation type or domain. This survey paper provides an overview of OpenIE technologies spanning from 2007 to 2024, emphasizing a chronological perspective absent in prior surveys. It examines the evolution of task settings in OpenIE to align with the advances in recent technologies. The paper categorizes OpenIE approaches into rule-based, neural, and pre-trained large language models, discussing each within a chronological framework. Additionally, it highlights prevalent datasets and evaluation metrics currently in use. Building on this extensive review, this paper systematically reviews the evolution of task settings, data, evaluation metrics, and methodologies in the era of large language models, highlighting their mutual influence, comparing their capabilities, and examining their implications for open challenges and future research directions.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Platform-Invariant Topic Modeling via Contrastive Learning to Mitigate Platform-Induced Bias

Minsoo Koo, DoeunKim, Sungwon Han, Sungkyu Shaun Park

Cross-platform topic dissemination is one of the research subjects that delved into media analysis; sometimes it fails to grasp the authentic topics due to platform-induced biases, which may be caused by aggregating documents from multiple platforms and running them on an existing topic model. This work deals with the impact of unique platform characteristics on the performance of topic models and proposes a new approach to enhance the effectiveness of topic modeling. The data utilized in this study consisted of a total of 1.5 million posts collected using the keyword “ChatGPT” on the three social media platforms. The devised model reduces platform influence in topic models by developing a platform-invariant contrastive learning algorithm and removing platform-specific jargon word sets. The proposed approach was thoroughly validated through quantitative and qualitative experiments alongside standard and state-of-the-art topic models and showed its supremacy. This method can mitigate biases arising from platform influences when modeling topics from texts collected across various platforms.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MAVEN-FACT: A Large-scale Event Factuality Detection Dataset

Chunyang Li, Hao Peng, Xiaohi Wang, Yunjia Qi, Lei Hou, Bin Xu, Juanzi Li

Event Factuality Detection (EFD) task determines the factuality of textual events, i.e., classifying whether an event is a fact, possibility, or impossibility, which is essential for faithfully understanding and utilizing event knowledge. However, due to the lack of high-quality large-scale data, event factuality detection is under-explored in event understanding research, which limits the development of EFD community. To address these issues and provide faithful event understanding, we introduce MAVEN-FACT, a large-scale and high-quality EFD dataset based on the MAVEN dataset. MAVEN-FACT includes factuality annotations of 112, 276 events, making it the largest EFD dataset. Extensive experiments demonstrate that MAVEN-FACT is challenging for both conventional fine-tuned models and large language models (LLMs). Thanks to the comprehensive annotations of event arguments and relations in MAVEN, MAVEN-FACT also supports some further analyses and we find that adopting event arguments and relations helps in event factuality detection for fine-tuned models but does not benefit LLMs. Furthermore, we preliminarily study an application case of event factuality detection and find it helps in mitigating event-related hallucination in LLMs. We will release our dataset and codes to facilitate further research on event factuality detection.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

ITER: Iterative Transformer-based Entity Recognition and Relation Extraction

Moritz Hennen, Florian Babil, Michaela Geierhos

When extracting structured information from text, recognizing entities and extracting relationships are essential. Recent advances in both tasks generate a structured representation of the information in an autoregressive manner, a time-consuming and computationally expensive approach. This naturally raises the question of whether autoregressive methods are necessary in order to achieve comparable results. In this work, we propose ITER, an efficient encoder-based relation extraction model, that performs the task in three parallelizable steps, greatly accelerating a recent language modeling approach: ITER achieves an inference throughput of over 600 samples per second for a large model on a single consumer-grade GPU. Furthermore, we achieve state-of-the-art results on the relation extraction datasets ADE and ACE05, and demonstrate competitive performance for both named entity recognition with GENIA and CoNLL03, and for relation extraction with SciERC and CoNLL04.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Consistent Document-level Relation Extraction via Counterfactuals

Ali Modarressi, Abdullatif Köksal, Hinrich Schütze

Many datasets have been developed to train and evaluate document-level relation extraction (RE) models. Most of these are constructed using real-world data. It has been shown that RE models trained on real-world data suffer from factual biases. To evaluate and address this issue, we present CovEReD, a counterfactual data generation approach for document-level relation extraction datasets using entity replacement. We first demonstrate that models trained on factual data exhibit inconsistent behavior: while they accurately extract triples from factual data, they fail to extract the same triples after counterfactual modification. This inconsistency suggests that models trained on factual data rely on spurious signals such as specific entities and external knowledge – rather than on context – to extract triples. We show that by generating document-level counterfactual data with CovEReD and training models on them, consistency is maintained with minimal impact on RE performance. We release our CovEReD pipeline as well as Re-DocRED-CF, a dataset of counterfactual RE documents, to assist in evaluating and addressing inconsistency in document-level RE.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

AliGATr: Graph-based layout generation for form understanding

Armineh Nourbakhsh, Zhao Jin, Siddharth Parekh, Sameena Shah, Carolyn Rose

Forms constitute a large portion of layout-rich documents that convey information through key-value pairs. Form understanding involves two main tasks, namely, the identification of keys and values (a.k.a. Key Information Extraction or KIE) and the association of keys to corresponding values (a.k.a. Relation Extraction or RE). State-of-the-art models for form understanding often rely on training paradigms that yield poorly calibrated output probabilities and low performance on RE. In this paper, we present AliGATr, a graph-based model that uses a generative objective to represent complex grid-like layouts that are often found in forms. Using a grid-based graph topology, our model learns to generate the layout of each page token by token in a data-efficient manner. Despite using 30% fewer parameters than the smallest SotA, AliGATr performs on par with or better than SotA models on the KIE and RE tasks against four datasets. We also show that AliGATr's output probabilities are better calibrated and do not exhibit the over-confident distributions of other SotA models.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

A Pointer Network-based Approach for Joint Extraction and Detection of Multi-Label Multi-Class Intents

Ankan Mukherjee, Sombit Bose, Abhilash Nandy, Gajala Sai Chaitanya, Pawan Goyal

In task-oriented dialogue systems, intent detection is crucial for interpreting user queries and providing appropriate responses. Existing research primarily addresses simple queries with a single intent, lacking effective systems for handling complex queries with multiple intents and extracting different intent spans. Additionally, there is a notable absence of multilingual multi-intent datasets. This study addresses three critical tasks: extracting multiple intent spans from queries, detecting multiple intents, and developing a multilingual multi-label intent dataset. We introduce a novel multi-label multi-class intent detection dataset (MLMCID-dataset) curated from existing benchmark datasets. We also propose a pointer network-based architecture (MLMCID) to extract intent spans and detect multiple intents with coarse and fine-grained labels in the form of sextuplets. Comprehensive analysis demonstrates the superiority of our pointer network based system over baseline approaches in terms of accuracy and F1-score across various datasets.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Improving Temporal Reasoning of Language Models via Recounted Narratives

Xinliang Frederick Zhang, Nicholas Beauchamp, Lu Wang

Reasoning about time and temporal relations is an integral aspect of human cognition, essential for perceiving the world and navigating our experiences. Though large language models (LLMs) have demonstrated impressive performance in many reasoning tasks, temporal reasoning remains challenging due to its intrinsic complexity. In this work, we first study an essential task of temporal reasoning—temporal graph generation, to unveil LLMs inherent, global reasoning capabilities. We show that this task presents great challenges even for the most powerful LLMs, such as GPT-3.5/4. We also notice a significant performance gap by small models ($< 10B$) that lag behind LLMs by 50%. Next, we study how to close this gap with a budget constraint, e.g., not using model finetuning. We propose a new prompting technique tailored for temporal reasoning, Narrative-of-Thought (NoT), that first converts the events set to a Python class, then prompts a small model to generate a temporally grounded narrative, guiding the final generation of a temporal graph. Extensive experiments showcase the efficacy of NoT in improving various metrics. Notably, NoT attains the highest F1 on the Schema-11 evaluation set, while securing an overall F1 on par with GPT-3.5. NoT also achieves the best structural similarity across the board, even compared with GPT-3.5/4.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Block-Diagonal Orthogonal Relation and Matrix Entity for Knowledge Graph Embedding

Yihua Zhu, Hirotoshi Shimodaira

The primary aim of Knowledge Graph Embeddings (KGE) is to learn low-dimensional representations of entities and relations for predicting missing facts. While rotation-based methods like RotatE and QuatE perform well in KGE, they face two challenges: limited model flexibility requiring proportional increases in relation size with entity dimension, and difficulties in generalizing the model for higher-dimensional rotations. To address these issues, we introduce OrthogonalE, a novel KGE model employing matrices for entities and block-diagonal orthogonal matrices with Riemannian optimization for relations. This approach not only enhances the generality and flexibility of KGE models but also captures several relation patterns that rotation-based methods can identify. Experimental results indicate that our new KGE model, OrthogonalE, offers generality and flexibility, captures several relation patterns, and significantly outperforms state-of-the-art KGE models while substantially reducing the number of relation parameters.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MMAR: Multilingual and Multimodal Anaphora Resolution in Instructional Videos

Cennet Oğuz, Pascal Denis, Simon Ostermann, Emmanuel Vincent, Natalia Skachkova, Josef van Genabith

Multilingual anaphora resolution identifies referring expressions and implicit arguments in texts and links to antecedents that cover several languages. In the most challenging setting, cross-lingual anaphora resolution, training data, and test data are in different languages. As knowledge needs to be transferred across languages, this task is challenging, both in the multilingual and cross-lingual setting. We hypothesize that one way to alleviate some of the difficulty of the task is to include multimodal information in the form of images (i.e. frames extracted from instructional videos). Such visual inputs are by nature language agnostic, therefore cross- and multilingual anaphora resolution should benefit from visual information. In this paper, we provide the first multilingual and multimodal dataset annotated with anaphoric relations and present experimental results for end-to-end multimodal and multilingual anaphora resolution. Given gold mentions, multimodal features improve anaphora resolution results by 10 % for unseen languages.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

When and Where Did it Happen? An Encoder-Decoder Model to Identify Scenario Context

Enrique Noriega-Atala, Robert Vacareanu, Salena Torres Ashton, Adarsh Pyarelal, Clayton T Morrison, Mihai Surdeanu

We introduce a neural architecture finetuned for the task of scenario context generation: The relevant location and time of an event or entity mentioned in text. Contextualizing information extraction helps to scope the validity of automated findings when aggregating them as knowledge graphs. Our approach uses a high-quality curated dataset of time and location annotations in a corpus of epidemiology papers to train an

encoder-decoder architecture. We also explored the use of data augmentation techniques during training. Our findings suggest that a relatively small fine-tuned encoder-decoder model performs better than out-of-the-box LLMs and semantic role labeling parsers to accurately predict the relevant scenario information of a particular entity or event.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Graph-tree Fusion Model with Bidirectional Information Propagation for Long Document Classification

Sudipta Singha Roy, Xindi Wang, Robert Mercer, Frank Rudzicz

Long document classification presents challenges in capturing both local and global dependencies due to their extensive content and complex structure. Existing methods often struggle with token limits and fail to adequately model hierarchical relationships within documents. To address these constraints, we propose a novel model leveraging a graph-tree structure. Our approach integrates syntax trees for sentence encodings and document graphs for document encodings, which capture fine-grained syntactic relationships and broader document contexts, respectively. We use Tree Transformers to generate sentence encodings, while a graph attention network models inter- and intra-sentence dependencies. During training, we implement bidirectional information propagation from word-to-sentence-to-document and vice versa, which enriches the contextual representation. Our proposed method enables a comprehensive understanding of content at all hierarchical levels and effectively handles arbitrarily long contexts without token limit constraints. Experimental results demonstrate the effectiveness of our approach in all types of long document classification tasks.

Multimodality and Language Grounding to Vision, Robotics and Beyond 5

Nov 14 (Thu) 14:00-15:30 - Room: Jasmine

Nov 14 (Thu) 14:00-15:30 - Jasmine

SignCLIP: Connecting Text and Sign Language by Contrastive Learning

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, Sarah Ebling

We present SignCLIP, which re-purposes CLIP (Contrastive Language-Image Pretraining) to project spoken language text and sign language videos, two classes of natural languages of distinct modalities, into the same space. SignCLIP is an efficient method of learning useful visual representations for sign language processing from large-scale, multilingual video-text pairs, without directly optimizing for a specific task or sign language which is often of limited size. We pretrain SignCLIP on Spreadthesign, a prominent sign language dictionary consisting of 500 thousand video clips in up to 44 sign languages, and evaluate it with various downstream datasets. SignCLIP discerns in-domain signing with notable text-to-video/video-to-video retrieval accuracy. It also performs competitively for out-of-domain downstream tasks such as isolated sign language recognition upon essential few-shot prompting or fine-tuning. We analyze the latent space formed by the spoken language text and sign language poses, which provides additional linguistic insights. Our code and models are openly available.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Can visual language models resolve textual ambiguity with visual cues? Let visual puns tell you!

Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, Youngjae Yu

Humans possess multimodal literacy, allowing them to actively integrate information from various modalities to form reasoning. Faced with challenges like lexical ambiguity in text, we supplement this with other modalities, such as thumbnail images or textbook illustrations. Is it possible for machines to achieve a similar multimodal understanding capability? In response, we present Understanding Pun with Image Explanations (UNPIE), a novel benchmark designed to assess the impact of multimodal inputs in resolving lexical ambiguities. Puns serve as the ideal subject for this evaluation due to their intrinsic ambiguity. Our dataset includes 1,000 puns, each accompanied by an image that explains both meanings. We pose three multimodal challenges with the annotations to assess different aspects of multimodal literacy: Pun Grounding, Disambiguation, and Reconstruction. The results indicate that various Socratic Models and Visual-Language Models improve over the text-only models when given visual context, particularly as the complexity of the tasks increases.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Quantifying the Gap Between Machine Translation and Native Language in Training for Multimodal, Multilingual Retrieval

Kyle Buettner, Adriana Kovashka

There is a scarcity of multilingual vision-language models that properly account for the perceptual differences that are reflected in image captions across languages and cultures. In this work, through a multimodal, multilingual retrieval case study, we quantify the existing lack of model flexibility. We empirically show performance gaps between training on captions that come from native German perception and captions that have been either machine-translated or human-translated from English into German. To address these gaps, we further propose and evaluate caption augmentation strategies. While we achieve mean recall improvements (+1.3), gaps still remain, indicating an open area of future work for the community.

Nov 14 (Thu) 14:00-15:30 - Jasmine

SURF: Teaching Large Vision-Language Models to Selectively Utilize Retrieved Information

Jiahuo Sun, Jiahai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, Yu Cheng

Large Vision-Language Models (LVLMs) have become pivotal at the intersection of computer vision and natural language processing. However, the full potential of LVLMs' Retrieval-Augmented Generation (RAG) capabilities remains underutilized. Existing works either focus solely on the text modality or are limited to specific tasks. Moreover, most LVLMs struggle to selectively utilize retrieved information and are sensitive to irrelevant or misleading references. To address these challenges, we propose a self-refinement framework designed to teach LVLMs to Selectively Utilize Retrieved Information (SURF). Specifically, when given questions that are incorrectly answered by the LVLML backbone, we obtain references that help correct the answers (positive references) and those that do not (negative references). We then fine-tune the LVLML backbone using a combination of these positive and negative references. Our experiments across three tasks and seven datasets demonstrate that our framework significantly enhances LVLMLs ability to effectively utilize retrieved multimodal references and improves their robustness against irrelevant or misleading information. The source code is available at <https://anonymous.4open.science/r/SURF-6433>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Attribute Diversity Determines the Systematicity Gap in VQA

Ian Berlot-Attwell, Kumar Krishna Agrawal, Annabelle Michael Carrell, Yash Sharma, Naomi Saphra

Although modern neural networks often generalize to new combinations of familiar concepts, the conditions that enable such compositionality have long been an open question. In this work, we study the systematicity gap in visual question answering: the performance difference between reasoning on previously seen and unseen combinations of object attributes. To test, we introduce a novel diagnostic dataset, CLEVR-HOPE. We find that the systematicity gap is not reduced by increasing the quantity of training data, but is reduced by increasing the diversity of training data. In particular, our experiments suggest that the more distinct attribute type combinations are seen during training, the more

systematic we can expect the resulting model to be.

Nov 14 (Thu) 14:00-15:30 - Jasmine

TravelER: A Modular Multi-LMM Agent Framework for Video Question-Answering

Chuqi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, Roei Herzig

Recently, image-based Large Multimodal Models (LMMs) have made significant progress in video question-answering (VideoQA) using a frame-wise approach by leveraging large-scale pretraining in a zero-shot manner. Nevertheless, these models need to be capable of finding relevant information, extracting it, and answering the question simultaneously. Currently, existing methods perform all of these steps in a single pass without being able to adapt if insufficient or incorrect information is collected. To overcome this, we introduce a modular multi-LMM agent framework based on several agents with different roles, instructed by a Planner agent that updates its instructions using shared feedback from the other agents. Specifically, we propose TravelER method that can create a plan to "##Travel##se" through the video, ask questions about individual frames to "##Locate##" and store key information, and then "##Evaluate##" if there is enough information to answer the question. Finally, if there is not enough information, our method is able to "##Replan##" based on its collected knowledge. Through extensive experiments, we find that the proposed TravelER approach improves performance on several VideoQA benchmarks without the need to fine-tune on specific datasets. Our code is available at <https://github.com/traveler-framework/TravelER>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Vision-Language Model Fine-Tuning via Simple Parameter-Efficient Modification

Ming Li, Like Zhong, Chenxin Li, Liuchuozheng Li, Nie Lin, Masashi Sugiyama

Recent advances in fine-tuning Vision-Language Models (VLMs) have witnessed the success of prompt tuning and adapter tuning, while the classic model fine-tuning on inherent parameters seems to be overlooked. It is believed that fine-tuning the parameters of VLMs with few-shot samples corrupt the pre-trained knowledge since fine-tuning the CLIP model even degrades performance. In this paper, we revisit this viewpoint, and propose a new perspective: fine-tuning the specific parameters instead of all will uncover the power of classic model fine-tuning on VLMs. Through our meticulous study, we propose ClipFit, a simple yet effective method to fine-tune CLIP without introducing any overhead of extra parameters. We demonstrate that by only fine-tuning the specific bias terms and normalization layers, ClipFit can improve the performance of zero-shot CLIP by 7.27% average harmonic mean accuracy. Lastly, to understand how fine-tuning in CLIPFit affects the pre-trained models, we conducted extensive experimental analyses w.r.t. changes in internal parameters and representations. We found that low-level text bias layers and the first layer normalization layer change much more than other layers. The code will be released.

Nov 14 (Thu) 14:00-15:30 - Jasmine

ActPlan-1K: Benchmarking the Procedural Planning Ability of Visual Language Models in Household Activities

Ying Su, Zhan Ling, Haochen Shi, Cheng Jiayang, Yauwei Yim, Yangqiu Song

Large language models(LLMs) have been adopted to process textual task description and accomplish procedural planning in embodied AI tasks because of their powerful reasoning ability. However, there is still lack of study on how vision language models(VLMs) behave when multi-modal task inputs are considered. Counterfactual planning that evaluates the model's reasoning ability over alternative task situations are also under exploited. In order to evaluate the planning ability of both multi-modal and counterfactual aspects, we propose ActPlan-1K. ActPlan-1K is a multi-modal planning benchmark constructed based on ChatGPT and household activity simulator iGibson2. The benchmark consists of 153 activities and 1,187 instances. Each instance describing one activity has a natural language task description and multiple environment images from the simulator. The gold plan of each instance is action sequences over the objects in provided scenes. Both the correctness and commonsense satisfaction are evaluated on typical VLMs. It turns out that current VLMs are still struggling at generating human-level procedural plans for both normal activities and counterfactual activities. We further provide automatic evaluation metrics by finetuning over BLEURT model to facilitate future research on our benchmark.

Nov 14 (Thu) 14:00-15:30 - Jasmine

On Efficient Language and Vision Assistants for Visually-Situated Natural Language Understanding: What Matters in Reading and Reasoning

Geewook Kim, Minjoon Seo

Recent advancements in language and vision assistants have showcased impressive capabilities but suffer from a lack of transparency, limiting broader research and reproducibility. While open-source models handle general image tasks effectively, they face challenges with the high computational demands of complex visually-situated text understanding. Such tasks often require increased token inputs and large vision modules to harness high-resolution information. Striking a balance between model size and data importance remains an open question. This study aims to redefine the design of vision-language models by identifying key components and creating efficient models with constrained inference costs. By strategically formulating datasets, optimizing vision modules, and enhancing supervision techniques, we achieve significant improvements in inference throughput while maintaining high performance. Extensive experiments across models ranging from 160M to 13B parameters offer insights into model optimization. We will fully open-source our codebase, models, and datasets at <https://github.com/naver-ai/elva>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Video-Text Prompting for Weakly Supervised Spatio-Temporal Video Grounding

Heng zhao, Zhao Yingjie, Bihai Wen, Yew-Soon Ong, Joey Tianyi Zhou

Weakly-supervised Spatio-Temporal Video Grounding(STVG) aims to localize target object tube given a text query, without densely annotated training data. Existing methods extract each candidate tube feature independently by cropping objects from video frame feature, discarding all contextual information such as position change and inter-entity relationship. In this paper, we propose Video-Text Prompting(VTP) to construct candidate feature. Instead of cropping tube region from feature map, we draw visual marker(s.e.g. red circle) over objects tubes as video prompts; corresponding text prompt(e.g. in red circle) is also inserted after the subject word of query text to highlight its presence. Nevertheless, each candidate feature may look similar without cropping. To address this, we further propose Contrastive VTP(CVT) by introducing negative contrastive samples whose candidate object is erased instead of being highlighted; by comparing the difference between VTP candidate and the contrastive sample, the gap of matching score between correct candidate and the rest is enlarged. Extensive experiments and ablations are conducted on several STVG datasets and our results surpass existing weakly-supervised methods by a great margin, demonstrating the effectiveness of our proposed methods.

Nov 14 (Thu) 14:00-15:30 - Jasmine

IntCoOp: Interpretability-Aware Vision-Language Prompt Tuning

Soumya Suvra Ghosal, Samyadeep Basu, Soheil Feizi, Dinesh Manocha

Image-text contrastive models such as CLIP learn transferable and robust representations for zero-shot transfer to a variety of downstream tasks. However, to obtain strong downstream performances, prompts need to be carefully curated, which can be a tedious engineering task. To address the issue of manual prompt engineering, prompt-tuning is used where a set of contextual vectors are learned by leveraging information from the training data. Despite their effectiveness, existing prompt-tuning frameworks often lack interpretability, thus limiting their ability to understand the compositional nature of images. In this work, we first identify that incorporating compositional attributes (e.g., a

"green" tree frog) in the design of manual prompts can significantly enhance image-text alignment scores. Building upon this observation, we propose a novel and interpretable prompt-tuning method named IntCoOp, which learns to jointly align attribute-level inductive biases and class embeddings during prompt-tuning. To assess the effectiveness of our approach, we evaluate IntCoOp across two representative tasks in a few-shot learning setup: generalization to novel classes, and unseen domain shifts. Through extensive experiments across 10 downstream datasets on CLIP, we find that introducing attribute-level inductive biases leads to superior performance against state-of-art prompt tuning frameworks. Notably, in a 16-shot setup, IntCoOp improves CoOp by 7.35% in average performance across 10 diverse datasets.

Nov 14 (Thu) 14:00-15:30 - Jasmine

VIEWS: Entity-Aware News Video Captioning

Hammad Ayyubi, Tianqi Liu, Arsha Nagrani, Xudong Lin, Mingda Zhang, Anurag Arnab, Feng Han, Yukun Zhu, Xuande Feng, Kevin Zhang, Jialu Liu, Shih-Fu Chang

Existing popular video captioning benchmarks and models often produce generic captions for videos that lack specific identification of individuals, locations, or organizations (named entities). However, in the case of news videos, the setting is more demanding, requiring the inclusion of such named entities for meaningful summarization. Therefore, we introduce the task of directly summarizing news videos into captions that are entity-aware. To facilitate research in this area, we have collected a large-scale dataset named VIEWS (Video NEWS). Within this task, we face challenges inherent to recognizing named entities and navigating diverse, dynamic contexts, all while relying solely on visual cues. To address these challenges, we propose a model-agnostic approach that enriches visual information extracted from videos with context sourced from external knowledge, enabling the generation of entity-aware captions. We validate the effectiveness of our approach across three video captioning models. Additionally, we conduct a critical analysis of our methodology to gain insights into the complexity of the task, the challenges it presents, and potential avenues for future research.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Eliciting In-Context Learning in Vision-Language Models for Videos Through Curated Data Distributional Properties

Keunwoo Peter Yu, Zheyuan Zhang, Fenyuan Hu, Shane Storks, Joyce Choi

A major reason behind the recent success of large language models (LLMs) is their *in-context learning* capability, which makes it possible to rapidly adapt them to downstream text-based tasks by prompting them with a small number of relevant demonstrations. While large vision-language models (VLMs) have recently been developed for tasks requiring both text and images, they largely lack in-context learning over visual information, especially in understanding and generating text about videos. In this work, we implement Emergent In-Context Learning on Videos (**EILeV**), a novel training paradigm that induces in-context learning over video and text by capturing key properties of pre-training data found by prior work to be essential for in-context learning in transformers. In our experiments, we show that **EILeV**-trained models outperform other off-the-shelf VLMs in few-shot video narration for novel, rare actions. Furthermore, we demonstrate that these key properties of bursty distributions, skewed marginal distributions, and dynamic meaning each contribute to varying degrees to VLMs' in-context learning capability in narrating procedural videos. Our results, analysis, and **EILeV**-trained models yield numerous insights about the emergence of in-context learning over video and text, creating a foundation for future work to optimize and scale VLMs for open-domain video understanding and reasoning.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Retrieval-enriched zero-shot image classification in low-resource domains

Nicola Dall'Asen, Yiming Wang, Enrica Fini, Elisa Ricci

Low-resource domains, characterized by scarce data and annotations, present significant challenges for language and visual understanding tasks, with the latter much under-explored in the literature. Recent advancements in Vision-Language Models (VLM) have shown promising results in high-resource domains but fall short in low-resource concepts that are under-represented (e.g. only a handful of images per category) in the pre-training set. We tackle the challenging task of zero-shot low-resource image classification from a novel perspective. By leveraging a retrieval-based strategy, we achieve this in a training-free fashion. Specifically, our method, named CoRE (Combination of Retrieval Enrichment), enriches the representation of both query images and class prototypes by retrieving relevant textual information from large web-crawled databases. This retrieval-based enrichment significantly boosts classification performance by incorporating the broader contextual information relevant to the specific class. We validate our method on a newly established benchmark covering diverse low-resource domains, including medical imaging, rare plants, and circuits. Our experiments demonstrate that CoRE outperforms existing state-of-the-art methods that rely on synthetic data generation and model fine-tuning.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Show and Guide: Instructional-Plan Grounded Vision and Language Model

Diogo Glória-Silva, David Semedo, Joao Magalhaes

Guiding users through complex procedural plans is an inherently multimodal task in which having visually illustrated plan steps is crucial to deliver an effective plan guidance. However, existing works on plan-following language models (LMs) often are not capable of multimodal input and output. In this work, we present MM-PlanLLM, the first multimodal LLM designed to assist users in executing instructional tasks by leveraging both textual plans and visual information. Specifically, we bring cross-modality through two key tasks: Conversational Video Moment Retrieval, where the model retrieves relevant step-video segments based on user queries, and Visually-Informed Step Generation, where the model generates the next step in a plan, conditioned on an image of the user's current progress. MM-PlanLLM is trained using a novel multitask-multistage approach, designed to gradually expose the model to multimodal instructional-plans semantic layers, achieving strong performance on both multimodal and textual dialogue in a plan-grounded setting. Furthermore, we show that the model delivers cross-modal temporal and plan-structure representations aligned between textual plan steps and instructional video moments.

Nov 14 (Thu) 14:00-15:30 - Jasmine

A Simple LLM Framework for Long-Range Video Question-Answering

Ce Zhang, Taixi Lu, Md Mohainul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, Gedas Bertasius

We present LLoVi, a simple yet effective ***L***nguage-based ***Lo***ng-range ***Vi***deo question-answering (LVQA) framework. Our method decomposes the short- and long-range modeling aspects of LVQA into two stages. First, we use a short-term visual captioner to generate textual descriptions of short video clips (0.5-8 seconds in length) densely sampled from a long input video. Afterward, an LLM aggregates the densely extracted short-term captions to answer a given question. Furthermore, we propose a novel multi-round summarization prompt that asks the LLM first to summarize the noisy short-term visual captions and then answer a given input question. To analyze what makes our simple framework so effective, we thoroughly evaluate various components of our framework. Our empirical analysis reveals that the choice of the visual captioner and LLM is critical for good LVQA performance. The proposed multi-round summarization prompt also leads to a significant LVQA performance boost. Our method achieves the best-reported results on the EgoSchema dataset, best known for very long-form video question-answering. LLoVi also outperforms the previous state-of-the-art by ***10.2%*** and ***6.2%*** on NExT-QA and IntentQA for LVQA. Finally, we extend LLoVi to grounded VideoQA, which requires both QA and temporal localization, and show that it outperforms all prior methods on NExT-GQA. Code is available at <https://github.com/CecZh/LLoVi>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

RECANFormer: Referring Expression Comprehension with Varying Numbers of Targets

Bhathiya Hemantage, Hakan Bilen, Phil Bartie, Christian Dondrup, Oliver Lemon

The Generalized Referring Expression Comprehension (GREC) task extends classic REC by generating image bounding boxes for objects referred to in natural language expressions, which may indicate zero, one, or multiple targets. This generalization enhances the practicality of REC models for diverse real-world applications. However, the presence of varying numbers of targets in samples makes GREC a more complex task, both in terms of training supervision and final prediction selection strategy. Addressing these challenges, we introduce RECANFormer, a one-stage method for GREC that combines a decoder-free (encoder-only) transformer architecture with DETR-like Hungarian matching. Our approach consistently outperforms baselines by significant margins in three GREC datasets.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Training-free Deep Concept Injection Enables Language Models for Video Question Answering

Xudong Lin, Manling Li, Richard Zemel, Heng Ji, Shih-Fu Chang

Recently, enabling pretrained language models (PLMs) to perform zero-shot crossmodal tasks such as video question answering has been extensively studied. A popular approach is to learn a projection network that projects visual features into the input text embedding space of a PLM, as well as feed-forward adaptation layers, with the weights of the PLM frozen. However, is it really necessary to learn such additional layers? In this paper, we make the first attempt to demonstrate that the PLM is able to perform zero-shot crossmodal tasks without any cross-modal pretraining, when the observed visual concepts are injected as both additional input text tokens and augmentation in the intermediate features within each feed-forward network for the PLM. Specifically, inputting observed visual concepts as text tokens helps to inject them through the self-attention layers in the PLM; to augment the intermediate features in a way that is compatible with the PLM, we propose to construct adaptation layers based on the intermediate representation of concepts (obtained by solely inputting them to the PLM). These two complementary injection mechanisms form the proposed Deep Concept Injection, which comprehensively enables the PLM to perceive instantly without crossmodal pretraining. Extensive empirical analysis on zero-shot video question answering, as well as visual question answering, shows Deep Concept Injection achieves competitive or even better results in both zero-shot and fine-tuning settings, compared to state-of-the-art methods that require crossmodal pretraining.

Nov 14 (Thu) 14:00-15:30 - Jasmine

MIBench: Evaluating Multimodal Large Language Models over Multiple Images

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, Weiming Hu

Built on the power of LLMs, numerous multimodal large language models (MLLMs) have recently achieved remarkable performance on various vision-language tasks. However, most existing MLLMs and benchmarks primarily focus on single-image input scenarios, leaving the performance of MLLMs when handling realistic multiple images underexplored. Although a few benchmarks consider multiple images, their evaluation dimensions and samples are very limited. In this paper, we propose a new benchmark MIBench, to comprehensively evaluate fine-grained abilities of MLLMs in multi-image scenarios. Specifically, MIBench categorizes the multi-image abilities into three scenarios: multi-image instruction (MII), multimodal knowledge-seeking (MKS) and multimodal in-context learning (MIC), and constructs 13 tasks with a total of 13K annotated samples. During data construction, for MII and MKS, we extract correct options from manual annotations and create challenging distractors to obtain multiple-choice questions. For MIC, to enable an in-depth evaluation, we set four sub-tasks and transform the original datasets into in-context learning formats. We evaluate several open-source and closed-source MLLMs on the proposed MIBench. The results reveal that although current models excel in single-image tasks, they exhibit significant shortcomings when faced with multi-image inputs, such as limited fine-grained perception, multi-image reasoning and in-context learning abilities. The annotated data of MIBench is available at <https://huggingface.co/datasets/StarBottle/MIBench>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Preserving Pre-trained Representation Space: On Effectiveness of Prefix-tuning for Large Multi-modal Models

Donghoon Kim, Gusang Lee, Kyuhong Shim, Byonghyo Shim

Recently, we have observed that Large Multi-modal Models (LMMs) are revolutionizing the way machines interact with the world, unlocking new possibilities across various multi-modal applications. To adapt LMMs for downstream tasks, parameter-efficient fine-tuning (PEFT) which only trains additional prefix tokens or modules, has gained popularity. Nevertheless, there has been little analysis of how PEFT works in LMMs. In this paper, we delve into the strengths and weaknesses of each tuning strategy, shifting the focus from the efficiency typically associated with these approaches. We first discover that model parameter tuning methods such as LoRA and Adapters, distort the feature representation space learned during pre-training, limiting the full utilization of pre-trained knowledge. We also demonstrate that prefix-tuning excels preserving the representation space, despite of its lower performance on downstream tasks. These findings suggest a simple two-step PEFT strategy called **Prefix-Tuned PEFT (PT-PEFT)**, which successively performs prefix-tuning and then other PEFT (i.e., Adapter, LoRA), combines the benefits of both. Experimental results show that PT-PEFT not only improves performance in image captioning and visual question answering compared to vanilla PEFT methods but also helps preserve the representation space of the four pre-trained models.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Difficult Task Yes but Simple Task No: Unveiling the Laziness in Multimodal LLMs

Sihang Zhao, Youliang Yuan, Xiaoying Tang, Pinjia He

Multimodal Large Language Models (MLLMs) demonstrate a strong understanding of the real world and can even handle complex tasks. However, they still fail on some straightforward visual question-answering (VQA) problems. This paper dives deeper into this issue, revealing that models tend to err when answering easy questions (e.g., Yes/No) questions about an image, even though they can correctly describe it. We refer to this model behavior discrepancy between difficult and simple questions as model laziness. To systematically investigate model laziness, we manually construct LazyBench, a benchmark that includes Yes/No, multiple choice, short answer questions, and image description tasks that are related to the same subjects in the images. Based on LazyBench, we observe that laziness widely exists in current advanced MLLMs (e.g., GPT-4o, Gemini-1.5-pro, Claude 3, LLaVA-1.5, LLaVA-1.6, and Qwen-VL). We also analyzed the failure cases of LLaVA-1.5-13B on the VQA-v2 benchmark and discovered that about half of these failures are due to the models' laziness. This further highlights the importance of ensuring that the model fully utilizes its capability. To this end, we conduct a preliminary exploration of how to mitigate laziness and find that chain of thought can effectively avoid this issue. The data can be accessed at <https://github.com/Akutagawa1998/LazyBench>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

VPL: Visual Proxy Learning Framework for Zero-Shot Medical Image Diagnosis

Jiaxiang Liu, Tianxiang Hu, Huimin Xiong, Jiawei Du, YANG FENG, Jian Wu, Joey Tianyi Zhou, Zuozhu Liu

Vision-language models like CLIP, utilizing class proxies derived from class name text features, have shown a notable capability in zero-shot medical image diagnosis which is vital in scenarios with limited disease databases or labeled samples. However, insufficient medical text precision and the modal disparity between text and vision spaces pose challenges for such paradigm. We show analytically and experimentally that enriching medical texts with detailed descriptions can markedly enhance the diagnosis performance, with the granularity and phrasing of these enhancements having a crucial impact on CLIP's understanding of medical images; and learning proxies within the vision domain can effectively circumvent the modal gap issue. Based on our analysis, we propose a medical visual proxy learning framework comprising two key components: a text refinement module that creates high quality medical text descriptions, and a stable Sinkhorn algorithm for an efficient

generation of pseudo labels which further guide the visual proxy learning. Our method elevates the Vanilla CLIP inference by supplying meticulously crafted clues to leverage CLIP's existing interpretive power and using the feature of refined texts to bridge the vision-text gap. The effectiveness and robustness of our method are clearly demonstrated through extensive experiments. Notably, our method outperforms the state-of-the-art zero-shot medical image diagnosis by a significant margin, ranging from 1.69% to 15.31% on five datasets covering various diseases, confirming its immense potential in zero-shot diagnosis across diverse medical applications.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Unleashing the Potentials of Likelihood Composition for Multi-modal Language Models

Shitian Zhao, Renrui Zhang, Xu Luo, Yan Wang, Shanghang Zhang, Peng Guo

Model fusing has always been an important topic, especially in an era where large language models (LLM) and multi-modal language models (MLM) with different architectures, parameter sizes and training pipelines, are being created all the time. In this work, we propose a post-hoc framework, aiming at fusing heterogeneous models off-the-shell, which we call *likelihood composition*, and the basic idea is to compose multiple models' likelihood distribution when doing a multi-choice visual-question-answering task. Here the core concept, *likelihood*, is actually the log-probability of the candidate answer. In *likelihood composition*, we introduce some basic operations: *debias*, *highlight*, *majority-vote* and *ensemble*. By combining (composing) these basic elements, we get the mixed composition methods: *mix-composition*. Through conducting comprehensive experiments on 9 VQA datasets and 10 MLMs, we prove the effectiveness of *mix-composition* compared with simple *ensemble* or *majority-vote* methods. In this framework, people can propose new basic composition methods and combine them to get the new mixed composition methods. We hope our proposed *likelihood composition* can provide a new perspective of fusing heterogeneous models and inspire the exploration under this framework.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Zero-shot Commonsense Reasoning over Machine Imagination

Hyuntae Park, Yeachan Kim, Jun-Hyung Park, SangKeun Lee

Recent approaches to zero-shot commonsense reasoning have enabled Pre-trained Language Models (PLMs) to learn a broad range of commonsense knowledge without being tailored to specific situations. However, they often suffer from human reporting bias inherent in textual commonsense knowledge, leading to discrepancies in understanding between PLMs and humans. In this work, we aim to bridge this gap by introducing an additional information channel to PLMs. We propose Imagine (Machine Imagination-based Reasoning), a novel zero-shot commonsense reasoning framework designed to complement textual inputs with visual signals derived from machine-generated images. To achieve this, we enhance PLMs with imagination capabilities by incorporating an image generator into the reasoning process. To guide PLMs in effectively leveraging machine imagination, we create a synthetic pre-training dataset that simulates visual question-answering. Our extensive experiments on diverse reasoning benchmarks and analysis show that Imagine outperforms existing methods by a large margin, highlighting the strength of machine imagination in mitigating reporting bias and enhancing generalization capabilities.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Adversarial Attacks on Parts of Speech: An Empirical Study in Text-to-Image Generation

G M Shaharai, Jia Chen, Jiachen Li, Yue Dong

Recent studies show that text-to-image (T2I) models are vulnerable to adversarial attacks, especially with noun perturbations in text prompts. In this study, we investigate the impact of adversarial attacks on different POS tags within text prompts on the images generated by T2I models. We create a high-quality dataset for realistic POS tag token swapping and perform gradient-based attacks to find adversarial suffixes that mislead T2I models into generating images with altered tokens. Our empirical results show that the attack success rate (ASR) varies significantly among different POS tag categories, with nouns, proper nouns, and adjectives being the easiest to attack. We explore the mechanism behind the steering effect of adversarial suffixes, finding that the number of critical tokens and information fusion vary among POS tags, while features like suffix transferability are consistent across categories.

Nov 14 (Thu) 14:00-15:30 - Jasmine

V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization

Yuxi Xie, Guanzhen Li, Xiao Xu, Min-Yen Kan

Large vision-language models (VLVLMs) suffer from hallucination, resulting in misalignment between the output textual response and the input visual content. Recent research indicates that the over-reliance on the Large Language Model (LLM) backbone, as one cause of the LVLV hallucination, inherently introduces bias from language priors, leading to insufficient context attention to the visual inputs. We tackle this issue of hallucination by mitigating such over-reliance through preference learning. We propose Vision-guided Direct Preference Optimization (V-DPO) to enhance visual context learning at training time. To interpret the effectiveness and generalizability of V-DPO on different types of training data, we construct a synthetic dataset containing both response- and image-contrast preference pairs, compared against existing human-annotated hallucination samples. Our approach achieves significant improvements compared with baseline methods across various hallucination benchmarks. Our analysis indicates that V-DPO excels in learning from image-contrast preference data, demonstrating its superior ability to elicit and understand nuances of visual context. Our code is publicly available at <https://github.com/YuxiXie/V-DPO>

Nov 14 (Thu) 14:00-15:30 - Jasmine

Tex2Model: Text-based Model Induction for Zero-shot Image Classification

Ohad Amotz, Tomer Volk, Eilam Shapira, Eyal Ben-David, Roi Reichart, Gal Chechik

We address the challenge of building task-agnostic classifiers using only text descriptions, demonstrating a unified approach to image classification, 3D point cloud classification, and action recognition from scenes. Unlike approaches that learn a fixed representation of the output classes, we generate an inference time a model tailored to a query classification task. To generate task-based zero-shot classifiers, we train a hypernetwork that receives class descriptions and outputs a multi-class model. The hypernetwork is designed to be equivariant with respect to the set of descriptions and the classification layer, thus obeying the symmetries of the problem and improving generalization. Our approach generates non-linear classifiers, handles rich textual descriptions, and may be adapted to produce lightweight models efficient enough for on-device applications. We evaluate this approach in a series of zero-shot classification tasks, for image, point-cloud, and action recognition, using a range of text descriptions: From single words to rich descriptions. Our results demonstrate strong improvements over previous approaches, showing that zero-shot learning can be applied with little training data. Furthermore, we conduct an analysis with foundational vision and language models, demonstrating that they struggle to generalize when describing what attributes the class lacks.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Grounding Complex Events in Multimodal Data

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, Benjamin Van Durme

How are we able to learn about complex current events just from short snippets of video? While natural language enables straightforward ways to represent under-specified, partially observable events, visual data does not facilitate analogous methods and, consequently, introduces unique challenges in event understanding. With the growing prevalence of vision-capable AI agents, these systems must be able to model events from collections of unstructured video data. To tackle robust event modeling in multimodal settings, we introduce a multimodal formulation for partially-defined events and cast the extraction of these events as a three-stage span retrieval task. We propose a corresponding

benchmark for this task, MultiVENT-G, that consists of 14.5 hours of densely annotated current event videos and 1,168 text documents, containing 22.8K labeled event-centric entities. We propose a collection of LLM-driven approaches to the task of multimodal event analysis, and evaluate them on MultiVENT-G. Results illustrate the challenges that abstract event understanding poses and demonstrates promise in event-centric video-language systems.

Nov 14 (Thu) 14:00-15:30 - Jasmine

EchoSight: Advancing Visual-Language Models with Wiki Knowledge

Yibin Yan, Weidi Xie

Knowledge-based Visual Question Answering (KVQA) tasks require answering questions about images using extensive background knowledge. Despite significant advancements, generative models often struggle with these tasks due to the limited integration of external knowledge. In this paper, we introduce ^{**}EchoSight^{**}, a novel multimodal Retrieval-Augmented Generation (RAG) framework that enables large language models (LLMs) to answer visual questions requiring fine-grained encyclopedic knowledge. To strive for high-performing retrieval, EchoSight first searches wiki articles by using visual-only information, subsequently, these candidate articles are further reranked according to their relevance to the combined text-image query. This approach significantly improves the integration of multimodal knowledge, leading to enhanced retrieval outcomes and more accurate VQA responses. Our experimental results on the E-VQA and InfoSeek datasets demonstrate that EchoSight establishes new state-of-the-art results in knowledge-based VQA, achieving an accuracy of 41.8% on E-VQA and 31.3% on InfoSeek.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Unraveling the Truth: Do LLMs really Understand Charts? A Deep Dive into Consistency and Robustness

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, Dan Roth

Chart question answering (CQA) is a crucial area of Visual Language Understanding. However, the robustness and consistency of current Visual Language Models (VLMs) in this field remain under-explored. This paper evaluates state-of-the-art VLMs on comprehensive datasets, developed specifically for this study, encompassing diverse question categories and chart formats. We investigate two key aspects: 1) the models' ability to handle varying levels of chart and question complexity, and 2) their robustness across different visual representations of the same underlying data. Our analysis reveals significant performance variations based on question and chart types, highlighting both strengths and weaknesses of current models. Additionally, we identify areas for improvement and propose future research directions to build more robust and reliable CQA systems. This study sheds light on the limitations of current models and paves the way for future advancements in the field.

Nov 14 (Thu) 14:00-15:30 - Jasmine

TransferCVLM: Transferring Cross-Modal Knowledge for Vision-Language Modeling

TransferCVLM: Jung-jae Kim, Hyunju Lee

Recent large vision-language multimodal models pre-trained with huge amount of image-text pairs show remarkable performances in downstream tasks. However, the multimodal pre-training has limitations in terms of resources and training time when it comes to obtaining new models that surpass existing models. To overcome these issues, we propose TransferCVLM, a method of efficient knowledge transfer that integrates pre-trained uni-modal models (and cross-modal fusion-encoder) into a combined vision-language model (CVLM), without pre-training the CVLM with large amount of multimodal data, and then for each task application, fine-tunes the CVLM and transfers the multimodal knowledge of a teacher vision-language model to the CVLM by using knowledge distillation techniques. We demonstrate that 1) the fine-tuned CVLM performs comparable to other vision-language models of similar size, that 2) the multimodal knowledge transfer consistently enhances the CVLM, and the knowledge-transferred CVLM composed of large-size unimodal models outperforms the teacher multimodal model in most of downstream tasks, and that 3) TransferCVLM can also be used for model compression when using small-size unimodal models. We estimate that the training of TransferCVLM takes only 6% of pre-training of other vision-language models. Our code is available at <https://github.com/DMCB-GIST/TransferCVLM>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Personalized Video Comment Generation

Xudong Lin, Ali Zare, Shiyuan Huang, Ming-Hsuan Yang, Shih-Fu Chang, Li Zhang

Generating personalized responses, particularly in the context of video, poses a unique challenge for language models. This paper introduces the novel task of **Personalized Video Comment Generation** (PVCG), aiming to predict user comments tailored to both the input video and the user's comment history, where the user is unseen during the model training process. Unlike existing video captioning tasks that ignores the personalization in the text generation process, we introduce PerVidCom, a new dataset specifically collected for this novel task with diverse personalized comments from YouTube. Recognizing the limitations of existing captioning metrics for evaluating this task, we propose a new automatic metric based on Large Language Models (LLMs) with few-shot in-context learning, named FICL-Score, specifically measuring quality from the aspects of emotion, language style and content relevance. We verify the proposed metric with human evaluations. We establish baselines using prominent Multimodal LLMs (MLLMs), analyze their performance discrepancies through extensive evaluation, and identifies directions for future improvement on this important task. Our research opens up a new direction of personalizing MLLMs and paves the way for future research.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design

Rishika Bhagwatkar, Shravan Nayak, Pouya Bashivan, Irina Rish

Vision-Language Models (VLMs) have seen a significant increase in both research interest and real-world applications across various domains, including healthcare, autonomous systems, and security. However, their growing prevalence demands higher reliability and safety including robustness to adversarial attacks. We systematically examine the possibility of incorporating adversarial robustness through various model design choices. We explore the effects of different vision encoders, the resolutions of vision encoders, and the size and type of language models. Additionally, we introduce novel, cost-effective approaches to enhance robustness through prompt engineering. By simply suggesting the possibility of adversarial perturbations or rephrasing questions, we demonstrate substantial improvements in model robustness against strong image-based attacks such as Auto-PGD. Our findings provide important guidelines for developing more robust VLMs, particularly for deployment in safety-critical environments where reliability and security are paramount. These insights are crucial for advancing the field of VLMs, ensuring they can be safely and effectively utilized in a wide range of applications.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Learning from Mistakes: Iterative Prompt Relabeling for Text-to-Image Diffusion Model Training

Xinyan Chen, Jiaxin Ge, Tianjun Zhang, Jianming Liu, Shanghang Zhang

Diffusion models have shown impressive performance in many domains. However, the model's capability to follow natural language instructions (e.g., spatial relationships between objects, generating complex scenes) is still unsatisfactory. In this work, we propose Iterative Prompt Relabeling (IPR), a novel algorithm that aligns images to text through iterative image sampling and prompt relabeling with feedback. IPR first samples a batch of images conditioned on the text, then relabels the text prompts of unmatched text-image pairs with classifier feedback. We

conduct thorough experiments on SDv2 and SDXL, testing their capability to follow instructions on spatial relations. With IPR, we improved up to 15.22% (absolute improvement) on the challenging spatial relation VISOR benchmark, demonstrating superior performance compared to previous RL methods. Our code is publicly available at <https://github.com/cxy00000/IPR-RLDF>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

PyramidCodec: Hierarchical Codec for Long-form Music Generation in Audio Domain

Jianyi Chen, Zheqi Dai, Zhen Ye, Xu Tan, Qifeng Liu, Yike Guo, Wei Xue

Generating well-structured long music compositions, spanning several minutes, remains a challenge due to inefficient representation and the lack of structured representation. In this paper, we propose PyramidCodec, a hierarchical discrete representation of audio, for long audio-domain music generation. Specifically, we employ residual vector quantization on different levels of features to obtain the hierarchical discrete representation. The highest level of features has the largest hop size, resulting in the most compact token sequence. The quantized higher-level representation is up-sampled and combined with lower-level features to apply residual vector quantization and obtain lower-level discrete representations. Furthermore, we design a hierarchical training strategy to ensure that the details are gradually added with more levels of tokens. By performing hierarchical tokenization, the overall token sequence represents information at various scales, facilitating long-context modeling in music and enabling the generation of well-structured compositions. The experimental results demonstrate that our proposed PyramidCodec achieves competitive performance in terms of reconstruction quality and token per second (TPS). By enabling ultra-long music modeling at the lowest level, the proposed approach facilitates training a language model that can generate well-structured long-form music for up to 3 minutes, whose quality is further demonstrated by subjective and objective evaluations. The samples can be found at <https://pyramidcodec.github.io/>.

Nov 14 (Thu) 14:00-15:30 - Jasmine

M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks

Florian Schneider, Sunayana Sitaram

Since the release of ChatGPT, the field of Natural Language Processing has experienced rapid advancements, particularly in Large Language Models (LLMs) and their multimodal counterparts, Large Multimodal Models (LMMs). Despite their impressive capabilities, LLMs often exhibit significant performance disparities across different languages and cultural contexts, as demonstrated by various text-only benchmarks. However, current research lacks such benchmarks for multimodal visuo-linguistic settings. This work fills this gap by introducing M5, the first comprehensive benchmark designed to evaluate LMMs on diverse vision-language tasks within a multilingual and multicultural context. M5 includes eight datasets covering five tasks and 41 languages, with a focus on underrepresented languages and culturally diverse images. Furthermore, we introduce two novel datasets, M5-VGR and M5-VLOD, including a new Visuo-Linguistic Outlier Detection task, in which all evaluated open-source models fail to significantly surpass the random baseline. Through extensive evaluation and analyses, we highlight substantial task-agnostic performance disparities between high- and low-resource languages. Moreover, we show that larger models do not necessarily outperform smaller ones in a multilingual setting.

Nov 14 (Thu) 14:00-15:30 - Jasmine

Navigating the Nuances: A Fine-grained Evaluation of Vision-Language Navigation

Zehao Wang, Minye Wu, Yixin Cao, Yubo Ma, Meiqi Chen, Tinne Tuytelaars

This study presents a novel evaluation framework for the Vision-Language Navigation (VLN) task. It aims to diagnose current models for various instruction categories at a finer-grained level. The framework is structured around the context-free grammar (CFG) of the task. The CFG serves as the basis for the problem decomposition and the core premise of the instruction categories design. We propose a semi-automatic method for CFG construction with the help of Large-Language Models (LLMs). Then, we induct and generate data spanning five principal instruction categories (i.e. direction change, landmark recognition, region recognition, vertical movement, and numerical comprehension). Our analysis of different models reveals notable performance discrepancies and recurrent issues. The stagnation of numerical comprehension, heavy selective biases over directional concepts, and other interesting findings contribute to the development of future language-guided navigation systems. The project is now available at <https://zehao-wang.github.io/navnuances>.

NLP Applications 5

Nov 14 (Thu) 14:00-15:30 - Room: Riverfront Hall

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

GraphReader: Building Graph-based Agent to Enhance Long-Context Abilities of Large Language Models

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, Bo Zheng

Long-context capabilities are essential for large language models (LLMs) to tackle complex and long-input tasks. Despite numerous efforts made to optimize LLMs for long contexts, challenges persist in robustly processing long inputs. In this paper, we introduce GraphReader, a graph-based agent system designed to handle long texts by structuring them into a graph and employing an agent to explore this graph autonomously. Upon receiving a question, the agent first undertakes a step-by-step analysis and devises a rational plan. It then invokes a set of predefined functions to read node content and neighbors, facilitating a coarse-to-fine exploration of the graph. Throughout the exploration, the agent continuously records new insights and reflects on current circumstances to optimize the process until it has gathered sufficient information to generate an answer. Experimental results on the LV-Eval dataset reveal that GraphReader using a 4k context window, consistently outperforms GPT-4-128k across context lengths from 16k to 256k by a large margin. Additionally, our approach demonstrates superior performance on four challenging single-hop and multi-hop benchmarks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

EAVE: Efficient Product Attribute Value Extraction via Lightweight Sparse-layer Interaction

Li Yang, Qifan Wang, Jianfeng Chi, Jiahao Liu, Jingang Wang, Fuli Feng, Zenglin Xu, Yi Fang, Lifu Huang, Dongfang Liu

Product attribute value extraction involves identifying the specific values associated with various attributes from a product profile. While existing methods often prioritize the development of effective models to improve extraction performance, there has been limited emphasis on extraction efficiency. However, in real-world scenarios, products are typically associated with multiple attributes, necessitating multiple extractions to obtain all corresponding values. In this work, we propose an Efficient product Attribute Value Extraction (EAVE) approach via lightweight sparse-layer interaction. Specifically, we employ a heavy encoder to separately encode the product context and attribute. The resulting non-interacting heavy representations of the context can be cached and reused for all attributes. Additionally, we introduce a light encoder to jointly encode the context and the attribute, facilitating lightweight interactions between them. To enrich the interaction within the lightweight encoder, we design a sparse-layer interaction module to fuse the non-interacting heavy representation into the lightweight encoder.

Comprehensive evaluation on two benchmarks demonstrate that our method achieves significant efficiency gains with neutral or marginal loss in performance when the context is long and number of attributes is large. Our code is available at: <https://anonymous.4open.science/r/EAVE-EA18>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, Huaxiu Yao

The recent emergence of Medical Large Vision Language Models (Med-LVLMs) has enhanced medical diagnosis. However, current Med-LVLMs frequently encounter factual issues, often generating responses that do not align with established medical facts. Retrieval-Augmented Generation (RAG), which utilizes external knowledge, can improve the factual accuracy of these models but introduces two major challenges. First, limited retrieved contexts might not cover all necessary information, while excessive retrieval can introduce irrelevant and inaccurate references, interfering with the model's generation. Second, in cases where the model originally responds correctly, applying RAG can lead to an over-reliance on retrieved contexts, resulting in incorrect answers. To address these issues, we propose RULE, which consists of two components. First, we introduce a provably effective strategy for controlling factuality risk through the calibrated selection of the number of retrieved contexts. Second, based on samples where over-reliance on retrieved contexts led to errors, we curate a preference dataset to fine-tune the model, balancing its dependence on inherent knowledge and retrieved contexts for generation. We demonstrate the effectiveness of RAFE on three medical VQA datasets, achieving an average improvement of 20.8% in factual accuracy.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

PhilоГPT: A Philology-Oriented Large Language Model for Ancient Chinese Manuscripts with Dunhuang as Case Study

Yuging Zhang, Baoyi He, Yihan Chen, Hangji Li, Han Yue, Shengyu Zhang, Huiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, Fei Wu
 Philology, the study of ancient manuscripts, demands years of professional training in ex-tensive knowledge memorization and manual textual retrieval. Despite these requirements align closely with strengths of recent successful Large Language Models (LLMs), the scarcity of high-quality, specialized training data has hindered direct applications. To bridge this gap, we curated the PhiloCorpus-ZH, a rich collection of ancient Chinese texts spanning a millen-nium with 30 diverse topics, including firsthand folk copies. This corpus facilitated the development of PhiloGPT, the first LLM tailored for discovering ancient Chinese manuscripts. To effectively tackle complex philological tasks like restoration, attribution, and linguistic anal-yis, we introduced the PhiloCoP framework. Modeled on the analytical patterns of philol-ogists, PhiloCoP enhances LLMs handling of historical linguistic peculiarities such as phonetic loans, polysemy, and syntactic inversions. We further integrated these tasks into the PhiloBenchmark, establishing a new standard for evaluating ancient Chinese LLMs addressing philology tasks. Deploying PhiloGPT in practical scenarios has enabled Dunhuang spe-cialists to resolve philology tasks, such as iden-tifying duplication of copied text and assisting archaeologists with text completion, demon-strating its potential in real-world applicatios.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Aligning Large Language Models with Diverse Political Viewpoints

Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcakre, Elliott Ash

Large language models such as ChatGPT exhibit striking political biases. If users query them about political information, they often take a normative stance. To overcome this, we align LLMs with diverse political viewpoints from 100,000 comments written by candidates running for national parliament in Switzerland. Models aligned with this data can generate more accurate political viewpoints from Swiss parties, compared to commercial models such as ChatGPT. We also propose a procedure to generate balanced overviews summarizing multiple viewpoints using such models. The replication package contains all code and data.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Structure Guided Prompt Instructing Large Language Model in Multi-Step Reasoning by Exploring Graph Structure of the Text

Kewei Cheng, Nesreen K. Ahmed, Theodore L. Willke, Yizhou Sun

Although Large Language Models (LLMs) excel at addressing straightforward reasoning tasks, they frequently struggle with difficulties when confronted by more complex multi-step reasoning due to a range of factors. Firstly, natural language often encompasses complex relationships among entities, making it challenging to maintain a clear reasoning chain over longer spans. Secondly, the abundance of linguistic diversity means that the same entities and relationships can be expressed using different terminologies and structures, complicating the task of identifying and establishing connections between multiple pieces of information. Graphs provide an effective solution to represent data rich in relational information and capture long-term dependencies among entities. To harness the potential of graphs, our paper introduces Structure Guided Prompt, an innovative three-stage task-agnostic prompting framework designed to improve the multi-step reasoning capabilities of LLMs in a zero-shot setting. This framework explicitly converts unstructured text into a graph via LLMs and instructs them to navigate this graph using task-specific strategies to formulate responses. By effectively organizing information and guiding navigation, it enables LLMs to provide more accurate and context-aware responses. Our experiments show that this framework significantly enhances the reasoning capabilities of LLMs, enabling them to excel in a broader spectrum of natural language scenarios.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Unveiling and Consulting Core Experts in Retrieval-Augmented MoE-based LLMs

Xin Zhou, Ping Nie, Yiwu Guo, Haojie Wei, Zhangyu Zhang, Pasquale Minervini, Ruotian Ma, Tao Gui, Qi Zhang, Xuanjing Huang

Retrieval-Augmented Generation (RAG) significantly improved the ability of Large Language Models (LLMs) to solve knowledge-intensive tasks. While existing research seeks to enhance RAG performance by retrieving higher-quality documents or designing RAG-specific LLMs, the internal mechanisms within LLMs that contribute to RAG's effectiveness remain underexplored. In this paper, we aim to investigate these internal mechanisms within the popular Mixture-of-Expert (MoE)-based LLMs and demonstrate how to improve RAG by examining expert activations in these LLMs. Our controlled experiments reveal that several core groups of experts are primarily responsible for RAG-related behaviors. The activation of these core experts can signify the model's inclination towards external/internal knowledge and adjust its behavior. For instance, we identify core experts that can (1) indicate the sufficiency of the model's internal knowledge, (2) assess the quality of retrieved documents, and (3) enhance the model's ability to utilize context. Based on these findings, we propose several strategies to enhance RAG's efficiency and effectiveness through expert activation. Experimental results across various datasets and MoE LLMs show the effectiveness of our method.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Do LLMs Know to Respect Copyright Notice?

Jialiang Xu, SHENGLAN LI, Zhaozhuo Xu, Denghui Zhang

Prior study shows that LLMs sometimes generate content that violates copyright. In this paper, we study another important yet underexplored problem, i.e., will LLMs respect copyright information in user input, and behave accordingly? The research problem is critical, as a negative answer would imply that LLMs will become the primary facilitator and accelerator of copyright infringement behavior. We conducted a series of experiments using a diverse set of language models, user prompts, and copyrighted materials, including books, news articles, API documentation, and movie scripts. Our study offers a conservative evaluation of the extent to which language models may infringe upon copyrights when processing user input containing protected material. This research emphasizes the need for further investigation and the importance of

ensuring LLMs respect copyright regulations when handling user input to prevent unauthorized use or reproduction of protected content. We also release a benchmark dataset serving as a test bed for evaluating infringement behaviors by LLMs and stress the need for future alignment.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Computational Meme Understanding: A Survey

Khoi P.N. Nguyen, Vincent Ng

Computational Meme Understanding, which concerns the automated comprehension of memes, has garnered interest over the last four years and is facing both substantial opportunities and challenges. We survey this emerging area of research by first introducing a comprehensive taxonomy for memes along three dimensions – forms, functions, and topics. Next, we present three key tasks in Computational Meme Understanding, namely, classification, interpretation, and explanation, and conduct a comprehensive review of existing datasets and models, discussing their limitations. Finally, we highlight the key challenges and recommend avenues for future work.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models

Jiaxin Zhang, Wendi Cui, Yiran Huang, Kamalika Das, Sricharan Kumar

Large language models (LLMs) are proficient at capturing factual knowledge across various domains. However, refining their capabilities on previously seen knowledge or integrating new knowledge from external sources remains a significant challenge. In this work, we propose a novel synthetic knowledge ingestion method called skrinspace, which leverages fine-grained synthesis, interleaved generation, and ensemble augmentation strategies to construct high-quality data representations from raw knowledge sources. We then integrate skr and its variations with three knowledge injection techniques: Retrieval Augmented Generation (RAG), Supervised Fine-tuning (SFT), and Continual Pre-training (CPT) to inject and refine knowledge in language models. Extensive empirical experiments are conducted on various question-answering tasks spanning finance, biomedicine, and open-generation domains to demonstrate that skr significantly outperforms baseline methods by facilitating effective knowledge injection. We believe that our work is an important step towards enhancing the factual accuracy of LLM outputs by refining knowledge representation and injection capabilities.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Adversarial Text Generation using Large Language Models for Dementia Detection

Youxiang Zhu, Nana Lin, Kiran Sandilya Balivada, Daniel Haehn, Xiaohui Liang

Although large language models (LLMs) excel in various text classification tasks, regular prompting strategies (e.g., few-shot prompting) do not work well with dementia detection via picture description. The challenge lies in the language marks for dementia are unclear, and LLM may struggle with relating its internal knowledge to dementia detection. In this paper, we present an accurate and interpretable classification approach by Adversarial Text Generation (ATG), a novel decoding strategy that could relate dementia detection with other tasks. We further develop a comprehensive set of instructions corresponding to various tasks and use them to guide ATG, achieving the best accuracy of 85%, >10% improvement compared to the regular prompting strategies. In addition, we introduce feature context, a human-understandable text that reveals the underlying features of LLM used for classifying dementia. From feature contexts, we found that dementia detection can be related to tasks such as assessing attention to detail, language, and clarity with specific features of the environment, character, and other picture content or language-related features. Future work includes incorporating multi-modal LLMs to interpret speech and picture information.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, May Dongmei Wang

Clinicians often rely on data engineers to retrieve complex patient information from electronic health record (EHR) systems, a process that is both inefficient and time-consuming. We propose EHRAgent, a large language model (LLM) agent empowered with accumulative domain knowledge and robust coding capability. EHRAgent enables autonomous code generation and execution to facilitate clinicians in directly interacting with EHRs using natural language. Specifically, we formulate a multi-tabular reasoning task based on EHRs as a tool-use planning process, efficiently decomposing a complex task into a sequence of manageable actions with external toolsets. We first inject relevant medical information to enable EHRAgent to effectively reason about the given query, identifying and extracting the required records from the appropriate tables. By integrating interactive coding and execution feedback, EHRAgent then effectively learns from error messages and iteratively improves its originally generated code. Experiments on three real-world EHR datasets show that EHRAgent outperforms the strongest baseline by up to 29.6% in success rate, verifying its strong capacity to tackle complex clinical tasks with minimal demonstrations.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MP-RNA: Unleashing Multi-species RNA Foundation Model via Calibrated Secondary Structure Prediction

Heng Yang, Ke Li

RNA foundation models (FMs) have been extensively used to interpret genomic sequences and address a wide range of in-silico genomic tasks. However, current RNA FMs often overlook the incorporation of secondary structures in the pretraining of FMs, which impedes the effectiveness in various genomic tasks. To address this problem, we leverage filtered high-fidelity structure annotations for structure pretraining to enhance the modeling ability of FMs in single nucleotide resolution tasks. Experimental evaluations across four comprehensive genomic benchmarks demonstrate that our RNA FM consistently outperforms existing RNA FMs, achieving a 40% improvement in RNA secondary structure prediction and obtaining top-tier results on DNA genomic benchmarks even though it has not been pretrained on any DNA genome. We release the code and models to encourage further research to bridge the gap between in-silico predictions and biological reality.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues

Deuksin Kwon, Emily Weiss, Tariq Kulshrestha, Kushal Chawla, Gale Lucas, Jonathan Gratch

A successful negotiation requires a range of capabilities, including comprehension of the conversation context, Theory-of-Mind (ToM) skills to infer the partners motives, strategic reasoning, and effective communication, making it challenging for automated systems. Despite the remarkable performance of LLMs in various NLP tasks, there is no systematic evaluation of their capabilities in negotiation. Such an evaluation is critical for advancing AI negotiation agents and negotiation research, ranging from designing dialogue systems to providing pedagogical feedback and scaling up data collection practices. This work aims to systematically analyze the multifaceted capabilities of LLMs across diverse dialogue scenarios throughout the stages of a typical negotiation interaction. Our analysis highlights GPT-4's superior performance in many tasks while identifying specific challenges, such as making subjective assessments and generating contextually appropriate, strategically advantageous responses.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

When Raw Data Prevails: Are Large Language Model Embeddings Effective in Numerical Data Representation for Medical Machine Learning Applications?

Yanyan Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, Majid Afshar

The introduction of Large Language Models (LLMs) has advanced data representation and analysis, bringing significant progress in their

use for medical questions and answering. Despite these advancements, integrating tabular data, especially numerical data pivotal in clinical contexts, into LLM paradigms has not been thoroughly explored. In this study, we examine the effectiveness of vector representations from last hidden states of LLMs for medical diagnostics and prognostics using electronic health record (EHR) data. We compare the performance of these embeddings with that of raw numerical EHR data when used as feature inputs to traditional machine learning (ML) algorithms that excel at tabular data learning, such as eXtreme Gradient Boosting. We focus on instruction-tuned LLMs in a zero-shot setting to represent abnormal physiological data and evaluating their utilities as feature extractors to enhance ML classifiers for predicting diagnoses, length of stay, and mortality. Furthermore, we examine prompt engineering techniques on zero-shot and few-shot LLM embeddings to measure their impact comprehensively. Although findings suggest the raw data features still prevail in medical ML tasks, zero-shot LLM embeddings demonstrate competitive results, suggesting a promising avenue for future research in medical applications.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage

Meng Lu, Brandon Ho, Dennis Ren, Xuan Wang

The global escalation in emergency department patient visits poses significant challenges to efficient clinical management, particularly in clinical triage. Traditionally managed by human professionals, clinical triage is susceptible to substantial variability and high workloads. Although large language models (LLMs) demonstrate promising reasoning and understanding capabilities, directly applying them to clinical triage remains challenging due to the complex and dynamic nature of the clinical triage task. To address these issues, we introduce TriageAgent, a novel heterogeneous multi-agent framework designed to enhance collaborative decision-making in clinical triage. TriageAgent leverages LLMs for role-playing, incorporating self-confidence and early-stopping mechanisms in multi-round discussions to improve document reasoning and classification precision for triage tasks. In addition, TriageAgent employs the medical Emergency Severity Index (ESI) handbook through a retrieval-augmented generation (RAG) approach to provide precise clinical knowledge and integrates both coarse- and fine-grained ESI-level predictions in the decision-making process. Extensive experiments demonstrate that TriageAgent outperforms state-of-the-art LLM-based methods on three clinical triage test sets. Furthermore, we have released the first public benchmark dataset for clinical triage with corresponding ESI levels and human expert performance for comparison.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Enabling Discriminative Reasoning in LLMs for Legal Judgment Prediction

Chenlong Deng, Kelong Mao, Yiyao Zhang, Zhicheng Dou

Legal judgment prediction is essential for enhancing judicial efficiency. In this work, we identify that existing large language models (LLMs) underperform in this domain due to challenges in understanding case complexities and distinguishing between similar charges. To adapt LLMs for effective legal judgment prediction, we introduce the Ask-Discriminate-Predict (ADAPT) reasoning framework inspired by human judicial reasoning. ADAPT involves decomposing case facts, discriminating among potential charges, and predicting the final judgment. We further enhance LLMs through fine-tuning with multi-task synthetic trajectories to improve legal judgment prediction accuracy and efficiency under our ADAPT framework. Extensive experiments conducted on two widely-used datasets demonstrate the superior performance of our framework in legal judgment prediction, particularly when dealing with complex and confusing charges.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Retrieval and Reasoning on KGs: Integrate Knowledge Graphs into Large Language Models for Complex Question Answering

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, mingjie zhong, Xu Jia, Min Zhang

Despite Large Language Models (LLMs) have performed impressively in various Natural Language Processing (NLP) tasks, their inherent hallucination phenomena severely challenge their credibility in complex reasoning. Combining explainable Knowledge Graphs (KGs) with LLMs is a promising path to address this issue. However, structured KGs are difficult to utilize, and how to make LLMs understand and incorporate them is a challenging topic. We thereby reorganize a more efficient structure of KGs, while designing the KG-related instruction tuning and continual pre-training strategies to enable LLMs to learn and internalize this form of representation effectively. Moreover, we construct subgraphs to further enhance the retrieval capabilities of KGs via CoT reasoning. Extensive experiments on two KGQA datasets demonstrate that our model achieves convincing performance compared to strong baselines⁸.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MetaKP: On-Demand Keyphrase Generation

Di Wu, Xiaoxian Shen, Kai-Wei Chang

Traditional keyphrase prediction methods predict a single set of keyphrases per document, failing to cater to the diverse needs of users and downstream applications. To bridge the gap, we introduce on-demand keyphrase generation, a novel paradigm that requires keyphrases that conform to specific high-level goals or intents. For this task, we present MetaKP, a large-scale benchmark comprising four datasets, 7500 documents, and 3/60 goals across news and biomedical domains with human-annotated keyphrases. Leveraging MetaKP, we design both supervised and unsupervised methods, including a multi-task fine-tuning approach and a self-consistency prompting method with large language models. The results highlight the challenges of supervised fine-tuning, whose performance is not robust to distribution shifts. By contrast, the proposed self-consistency prompting approach greatly improves the performance of large language models, enabling GPT-4o to achieve 0.548 SemF1, surpassing the performance of a fully fine-tuned BART-base model. Finally, we demonstrate the potential of our method to serve as a general NLP infrastructure, exemplified by its application in epidemic event detection from social media.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

PRESTO: Progressive Pretraining Enhances Synthetic Chemistry Outcomes

He CAO, Yanjun Shao, Zhiyuan Liu, Zijiang Liu, Xiangru Tang, Yuan Yao, Yu Li

Multimodal Large Language Models (MLLMs) have seen growing adoption across various scientific disciplines. These advancements encourage the investigation of molecule-text modeling within synthetic chemistry, a field dedicated to designing and conducting chemical reactions to synthesize new compounds with desired properties and applications. Current approaches, however, often neglect the critical role of multi-molecule graph interaction in understanding chemical reactions, leading to suboptimal performance in synthetic chemistry tasks. This study introduces PRESTO (Progressive Pretraining Enhances Synthetic Chemistry Outcomes), a new framework that bridges the molecule-text modality gap by integrating a comprehensive benchmark of pretraining strategies and dataset configurations. It progressively improves multimodal LLMs through cross-modal alignment and multi-graph understanding. Our extensive experiments demonstrate that PRESTO offers competitive results in downstream synthetic chemistry tasks. The code can be found at <https://github.com/IDEA-XL/PRESTO>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MELT: Materials-aware Continued Pre-training for Language Model Adaptation to Materials Science

Junho Kim, Yeachan Kim, Jun-Hyung Park, Yerim Oh, Suho Kim, SangKeun Lee

We introduce a novel continued pre-training method, MELT (MatErialiS-aware continued pre-Training), specifically designed to efficiently

⁸<https://github.com/Dereck0602/Retrieval-and-Reasoning-on-KGs>

adapt the pre-trained language models (PLMs) for materials science. Unlike previous adaptation strategies that solely focus on constructing domain-specific corpus, MELT comprehensively considers both the corpus and the training strategy, given that materials science corpus has distinct characteristics from other domains. To this end, we first construct a comprehensive materials knowledge base from the scientific corpus by building semantic graphs. Leveraging this extracted knowledge, we integrate a curriculum into the adaptation process that begins with familiar and generalized concepts and progressively moves toward more specialized terms. We conduct extensive experiments across diverse benchmarks to verify the effectiveness and generality of MELT. A comprehensive evaluation convincingly supports the strength of MELT, demonstrating superior performance compared to existing continued pre-training methods. In-depth analysis also shows that MELT enables PLMs to effectively represent materials entities compared to the existing adaptation methods, thereby highlighting its broad applicability across a wide spectrum of materials science.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Cross-Lingual Unlearning of Selective Knowledge in Multilingual Language Models

Minsok Choi, Kyunghyun Min, Jaegul Choo

Pretrained language models memorize vast amounts of information, including private and copyrighted data, raising significant safety concerns. Retraining these models after excluding sensitive data is prohibitively expensive, making machine unlearning a viable, cost-effective alternative. Previous research has focused on machine unlearning for monolingual models, but we find that unlearning in one language does not necessarily transfer to others. This vulnerability makes models susceptible to low-resource language attacks, where sensitive information remains accessible in less dominant languages. This paper presents a pioneering approach to machine unlearning for multilingual language models, selectively erasing information across different languages while maintaining overall performance. Specifically, our method employs an adaptive unlearning scheme that assigns language-dependent weights to address different language performances of multilingual language models. Empirical results demonstrate the effectiveness of our framework compared to existing unlearning baselines, setting a new standard for secure and adaptable multilingual language models.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Using LLMs to simulate students' responses to exam questions

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, Paula Buttery

Previous research leveraged Large Language Models (LLMs) in numerous ways in the educational domain. Here, we show that they can be used to answer exam questions simulating students of different skill levels and share a prompt engineered for GPT-3.5, that enables the simulation of varying student skill levels on questions from different educational domains. We evaluate the proposed prompt three publicly available datasets (one from science exams and two from English reading comprehension exams) and three LLMs (two versions of GPT-3.5 and one of GPT-4), and show that it is robust to different educational domains and capable of generalising to data unseen during the prompt engineering phase. We also show that, being engineered for a specific version of GPT-3.5, the prompt does not generalise well to different LLMs, stressing the need for prompt engineering for each model in practical applications. Lastly, we find that there is not a direct correlation between the quality of the rationales obtained with chain-of-thought prompting and the accuracy in the student simulation task.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Revisiting the Impact of Pursuing Modularity for Code Generation

Deokyeong Kang, Kih Jung Seo, Taeuk Kim

Modular programming, which aims to construct the final program by integrating smaller, independent building blocks, has been regarded as a desirable practice in software development. However, with the rise of recent code generation agents built upon large language models (LLMs), a question emerges: is this traditional practice equally effective for these new tools? In this work, we assess the impact of modularity in code generation by introducing a novel metric for its quantitative measurement. Surprisingly, unlike conventional wisdom on the topic, we find that modularity is not a core factor for improving the performance of code generation models. We also explore potential explanations for why LLMs do not exhibit a preference for modular code compared to non-modular code.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

MATHWELL: Generating Educational Math Word Problems

Bryan R Christi, Jonathan Kropko, Thomas Hartvigsen

Math word problems are critical K-8 educational tools, but writing them is time consuming and requires extensive expertise. To be educational, problems must be solvable, have accurate answers, and, most importantly, be educationally appropriate. We propose that language models have potential to support K-8 math education by automatically generating word problems. However, evaluating educational appropriateness is hard to quantify. We fill this gap by having teachers evaluate problems generated by LLMs, who find existing models and data often fail to be educationally appropriate. We then explore automatically generating *educational* word problems, ultimately using our expert annotations to finetune a 70B language model. Our model, MATHWELL, is the first K-8 word problem generator targeted at educational appropriateness. Further expert studies find MATHWELL generates problems far more solvable, accurate, and appropriate than public models. MATHWELL also matches GPT-4's problem quality while attaining more appropriate reading levels for K-8 students and avoiding generating harmful questions.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

To Err Is Human, but Llamas Can Learn It Too

Agnes Luharu, Taido Purason, Martin Yainikko, Maksym Del, Mark Fisher

This study explores enhancing grammatical error correction (GEC) through automatic error generation (AEG) using language models (LMs). Specifically, we fine-tune Llama 2 LMs for error generation and find that this approach yields synthetic errors akin to human errors. Next, we train GEC Llama models using these artificial errors and outperform previous state-of-the-art error correction models, with gains ranging between 0.8 and 6 F0.5 points across all tested languages (German, Ukrainian, and Estonian). Moreover, we demonstrate that generating errors by fine-tuning smaller sequence-to-sequence models and prompting large commercial LMs (GPT3.5 and GPT4) also results in synthetic errors beneficially affecting error generation models. We openly release trained models for error generation and correction as well as all the synthesized error datasets for the covered languages.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Tab2Text - A framework for deep learning with tabular data

Tong Lin, Jason Yan, David Jurgens, Sabrina J Tomkins

Tabular data, from public opinion surveys to records of interactions with social services, is foundational to the social sciences. One application of such data is to fit supervised learning models in order to predict consequential outcomes, for example: whether a family is likely to be evicted, whether a student will graduate from high school or is at risk of dropping out, and whether a voter will turn out in an upcoming election. While supervised learning has seen drastic improvements in performance with advancements in deep learning technology, these gains are largely lost on tabular data which poses unique difficulties for deep learning frameworks. We propose a technique for transforming tabular data to text data and demonstrate the extent to which this technique can improve the performance of deep learning models for tabular data. Overall, we find modest gains (1.5% on average). Interestingly, we find that these gains do not depend on using large language models

to generate text.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

How to Train Your Fact Verifier: Knowledge Transfer with Multimodal Open Models

Jaeyoung Lee, Ximing Lu, Jack Hessel, Faeze Brahman, Youngjae Yu, Yonatan Bisk, Yejin Choi, Saadia Gabriel

Given the growing influx of misinformation across news and social media, there is a critical need for systems that can provide effective real-time verification of news claims. Large language or multimodal model based verification has been proposed to scale up online policing mechanisms for mitigating spread of false and harmful content. While these can potentially reduce burden on human fact-checkers, such efforts may be hampered by foundation model training data becoming outdated. In this work, we test the limits of improving foundation model performance without continual updating through an initial study of knowledge transfer using either existing intra- and inter-domain benchmarks or explanations generated from large language models (LLMs). We evaluate on 12 public benchmarks for fact-checking and misinformation detection as well as two other tasks relevant to content moderation - toxicity and stance detection. Our results on two recent multi-modal fact-checking benchmarks, Mocheg and Fakeddit, indicate that knowledge transfer strategies can improve Fakeddit performance over the state-of-the-art by up to 1.7% and Mocheg performance by up to 2.9%. The code, model checkpoints, and dataset are available: <https://github.com/given131/fact-verifier-knowledge-transfer>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

SciDoc2Diagrammer-MAF: Towards Generation of Scientific Diagrams from Documents guided by Multi-Aspect Feedback Refinement

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, Jordan Lee Boyd-Graber

Automating the creation of scientific diagrams from academic papers can significantly streamline the development of tutorials, presentations, and posters, thereby saving time and accelerating the process. Current text-to-image models (Rombach et al., 2022a; Belouadi et al., 2023) struggle with generating accurate and visually appealing diagrams from long-context inputs. We propose SciDoc2Diagram, a task that extracts relevant information from scientific papers and generates diagrams along with a benchmarking dataset, SciDoc2DiagramBench. We develop a multi-step pipeline SciDoc2Diagrammer that generates diagrams based on user intentions using intermediate code generation. We observed that initial diagram drafts were often incomplete or unfaithful to the source, leading us to develop SciDoc2Diagrammer-Multi-Aspect-Feedback (MAF), a refinement strategy that significantly enhances factual correctness and visual appeal and outperforms existing models on both automatic and human judgement.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

ClaimVer: Explainable Claim-Level Verification and Evidence Attribution of Text Through Knowledge Graphs

Preetam Prabhu Srikar Damnu, Himanshu Nadisu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, Chirag Shah

In the midst of widespread misinformation and disinformation through social media and the proliferation of AI-generated texts, it has become increasingly difficult for people to validate and trust information they encounter. Many fact-checking approaches and tools have been developed, but they often lack appropriate explainability or granularity to be useful in various contexts. A text validation method that is easy to use, accessible, and can perform fine-grained evidence attribution has become crucial. More importantly, building user trust in such a method requires presenting the rationale behind each prediction, as research shows this significantly influences people's belief in automated systems. Localizing and bringing users' attention to the specific problematic content is also paramount, instead of providing simple blanket labels. In this paper, we present *ClaimVer*, a *human-centric* framework tailored to meet users' informational and verification needs by generating rich annotations and thereby reducing cognitive load. Designed to deliver comprehensive evaluations of texts, it highlights each claim, verifies it against a trusted knowledge graph (KG), presents the evidence, and provides succinct, clear explanations for each claim prediction. Finally, our framework introduces an attribution score, enhancing applicability across a wide range of downstream tasks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Pedagogical Alignment of Large Language Models

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, Richard Baraniuk

Large Language Models (LLMs), when used in educational settings without pedagogical fine-tuning, often provide immediate answers rather than guiding students through the problem-solving process. This approach falls short of pedagogically best practices and limits their effectiveness as educational tools. We term the objective of training LLMs to emulate effective teaching strategies as ‘pedagogical alignment’. In this paper, we investigate Learning from Human Preferences (lhpns) algorithms to achieve this alignment objective. A key challenge in this process is the scarcity of high-quality preference datasets to guide the alignment. To address this, we propose a novel approach for constructing a large-scale dataset using synthetic data generation techniques, eliminating the need for time-consuming and costly manual annotation. Leveraging this dataset, our experiments with Llama and Mistral models demonstrate that LHP methods outperform standard supervised fine-tuning (SFT), improving pedagogical alignment accuracy by 13.1% and 8.7% respectively. Existing evaluation methods also lack quantitative metrics to adequately measure the pedagogical alignment of LLMs. To address this gap, we propose novel perplexity-based metrics that quantify LLMs’ tendency to provide scaffolded guidance versus direct answers, offering a robust measure of pedagogical alignment. Our analysis provides compelling evidence for the superiority of lhp methods over SFT in optimizing LLMs’ behavior, underscoring the potential of lhp methods in better aligning LLMs with educational objectives and fostering effective learning experiences. Code and models are available [here](https://github.com/sonkarshashank/lhpns).

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Let's Ask GNN: Empowering Large Language Model for Graph In-Context Learning

Yichuan Li, Zhengyu Hu, Zhengyu Chen, Jingang Wang, Han Liu, Kyumin Lee, Kaize Ding

Textual Attributed Graphs (TAGs) are crucial for modeling complex real-world systems, yet leveraging large language models (LLMs) for TAGs presents unique challenges due to the gap between sequential text processing and graph-structured data. We introduce AskGNN, a novel approach that bridges this gap by leveraging In-Context Learning (ICL) to integrate graph data and task-specific information into LLMs. AskGNN employs a Graph Neural Network (GNN)-powered structure-enhanced retriever to select labeled nodes across graphs, incorporating complex graph structures and their supervision signals. Our learning-to-retrieve algorithm optimizes the retriever to select example nodes that maximize LLM performance on graph. Experiments across three tasks and seven LLMs demonstrate AskGNN’s superior effectiveness in graph task performance, opening new avenues for applying LLMs to graph-structured data without extensive fine-tuning.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

LLM-Based Multi-Hop Question Answering with Knowledge Graph Integration in Evolving Environments

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaa Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, Bo Ai

The important challenge of keeping knowledge in Large Language Models (LLMs) up-to-date has led to the development of various methods for incorporating new facts. However, existing methods for such knowledge editing still face difficulties with multi-hop questions that require accurate fact identification and sequential logical reasoning, particularly among numerous fact updates. To tackle these challenges, this paper introduces Graph Memory-based Editing for Large Language Models (GMelLo), a straightforward and effective method that merges the explicit knowledge representation of Knowledge Graphs (KGs) with the linguistic flexibility of LLMs. Beyond merely leveraging LLMs for question answering, GMelLo employs these models to convert free-form language into structured queries and fact triples, facilitating seam-

less interaction with KGs for rapid updates and precise multi-hop reasoning. Our results show that GMelLo significantly surpasses current state-of-the-art (SOTA) knowledge editing methods in the multi-hop question answering benchmark, MQuAKE, especially in scenarios with extensive knowledge edits.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Knowledge-Centric Templatic Views of Documents

Isabel Alyssa Cachola, Silvia Cucerzan, Allen herring, Vuksan Mijovic, Erik Oveson, Sujay Kumar Jauhar

Authors seeking to communicate with broader audiences often share their ideas in various document formats, such as slide decks, newsletters, reports, and posters. Prior work on document generation has generally tackled the creation of each separate format to be a different task, leading to fragmented learning processes, redundancy in models and methods, and disjointed evaluation. We consider each of these documents as templatic views of the same underlying knowledge/content, and we aim to unify the generation and evaluation of these templatic views. We begin by showing that current LLMs are capable of generating various document formats with little to no supervision. Further, a simple augmentation involving a structured intermediate representation can improve performance, especially for smaller models. We then introduce a novel unified evaluation framework that can be adapted to measuring the quality of document generators for heterogeneous downstream applications. This evaluation is adaptable to a range of user defined criteria and application scenarios, obviating the need for task specific evaluation metrics. Finally, we conduct a human evaluation, which shows that people prefer 82% of the documents generated with our method, while correlating more highly with our unified evaluation framework than prior metrics in the literature.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

To Ask LLMs about English Grammaticality, Prompt Them in a Different Language

Shabnam Behzad, Amir Zeldes, Nathan Schneider

In addition to asking questions about facts in the world, some internet users in particular, second language learners ask questions about language itself. Depending on their proficiency level and audience, they may pose these questions in an L1 (first language) or an L2 (second language). We investigate how multilingual LLMs perform at crosslingual metalinguistic question answering. Focusing on binary questions about sentence grammaticality constructed from error-annotated learner corpora, we prompt three LLMs (Aya, Llama, and GPT) in multiple languages, including English, German, Korean, Russian, and Ukrainian. Our study reveals that the language of the prompt can significantly affect model performance, and despite English being the dominant training language for all three models, prompting in a different language with questions about English often yields better results.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Who Wrote When? Author Diarization in Social Media Discussions

Benedikt Boenninghoff, Henry Hosseini, Robert M. Nickel, Dorothea Kolossa

We are proposing a novel framework for author diarization, i.e. attributing comments in online discussions to individual authors. We consider an innovative approach that merges pre-trained neural representations of writing style with author-conditional encoder-decoder diarization, enhanced by a Conditional Random Field with Viterbi decoding for alignment refinement. Additionally, we introduce two new large-scale German language datasets, one for authorship verification and the other for author diarization. We evaluate the performance of our diarization framework on these datasets, offering insights into the strengths and limitations of this approach.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Misinformation with Legal Consequences (MisLC): A New Task Towards Harnessing Societal Harm of Misinformation

Chu Fei Luo, Radin Shayanfar, Rohan V Bhamoria, Samuel Dahan, Xiaodan Zhu

Misinformation, defined as false or inaccurate information, can result in significant societal harm when it is spread with malicious or even unintentional intent. The rapid online information exchange necessitates advanced detection mechanisms to mitigate misinformation-induced harm. Existing research, however, has predominantly focused on the veracity of information, overlooking the legal implications and consequences of misinformation. In this work, we take a novel angle to consolidate the definition of misinformation detection using legal issues as a measurement of societal ramifications, aiming to bring interdisciplinary efforts to tackle misinformation and its consequence. We introduce a new task: Misinformation with Legal Consequence (MisLC), which leverages definitions from a wide range of legal domains covering 4 broader legal topics and 11 fine-grained legal issues, including hate speech, election laws, and privacy regulations. For this task, we advocate a two-step dataset curation approach that utilizes crowd-sourced checkworthiness and expert evaluations of misinformation. We provide insights about the MisLC task through empirical evidence, from the problem definition to experiments and expert involvement. While the latest large language models and retrieval-augmented generation are effective baselines for the task, we find they are still far from replicating expert performance.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Domain Adaptation via Prompt Learning for Alzheimer's Detection

Shahla Farzana

Spoken language presents a compelling medium for non-invasive Alzheimer's disease (AD) screening, and prior work has examined the use of fine-tuned pretrained language models (PLMs) for this purpose. However, PLMs are often optimized on tasks that are inconsistent with AD classification. Spoken language corpora for AD detection are also small and disparate, making generalizability difficult. This paper investigates the use of domain-adaptive prompt fine-tuning for AD detection, using AD classification loss as the training objective and leveraging spoken language corpora from a variety of language tasks. Extensive experiments using voting-based combinations of different prompting paradigms show an impressive mean detection $F1=0.8952$ (with $std=0.01$ and best $F1=0.9130$) for the highest-performing approach when using BERT as the base PLM.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Fine-Tuning Language Models on Multiple Datasets for Citation Intention Classification

Zeren Shui, Petros Karypis, Daniel S. Karlos, Mingjian Wen, Saurav Manchanda, Ellad B. Tadmor, George Karypis

Citation intention Classification (CIC) tools classify citations by their intention (e.g., background, motivation) and assist readers in evaluating the contribution of scientific literature. Prior research has shown that pretrained language models (PLMs) such as SciBERT can achieve state-of-the-art performance on CIC benchmarks. PLMs are trained via self-supervision tasks on a large corpus of general text and can quickly adapt to CIC tasks via moderate fine-tuning on the corresponding dataset. Despite their advantages, PLMs can easily overfit small datasets during fine-tuning. In this paper, we propose a multi-task learning (MTL) framework that jointly fine-tunes PLMs on a dataset of primary interest together with multiple auxiliary CIC datasets to take advantage of additional supervision signals. We develop a data-driven task relation learning (TRL) method that controls the contribution of auxiliary datasets to avoid negative transfer and expensive hyper-parameter tuning. We conduct experiments on three CIC datasets and show that fine-tuning with additional datasets can improve the PLMs' generalization performance on the primary dataset. PLMs fine-tuned with our proposed framework outperform the current state-of-the-art models by 7% to 11% on small datasets while aligning with the best-performing model on a large dataset.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

BiMedIX: Bilingual Medical Mixture of Experts LLM

Sara Pieri, Sahal Shaji Mullaipilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, Hisham Cholakkal

In this paper, we introduce BiMedIX, the first bilingual medical mixture of experts LLM designed for seamless interaction in both English and Arabic. Our model facilitates a wide range of medical interactions in English and Arabic, including multi-turn chats to inquire about additional details such as patient symptoms and medical history, multiple-choice question answering, and open-ended question answering. We propose a semi-automated English-to-Arabic translation pipeline with human refinement to ensure high-quality translations. We also introduce a comprehensive evaluation benchmark for Arabic medical LLMs. Furthermore, we introduce BiMed1.3M, an extensive Arabic-English bilingual instruction set that covers 1.3 Million diverse medical interactions, including 200k synthesized multi-turn doctor-patient chats, in a 1:2 Arabic-to-English ratio. Our model outperforms state-of-the-art Med42 and Meditron by average absolute gains of 2.5% and 4.1%, respectively, computed across multiple medical evaluation benchmarks in English, while operating at 8-times faster inference. Moreover, our BiMedIX outperforms the generic Arabic-English bilingual LLM, Jais-30B, by average absolute gains of 10% on our Arabic and 15% on our bilingual evaluations across multiple datasets. Additionally, BiMedIX exceeds the accuracy of GPT4 by 4.4% in open-ended question UPHILL evaluation and largely outperforms state-of-the-art open source medical LLMs in human evaluations of multi-turn conversations. Our trained models, instruction set, and source code are available at <https://github.com/mbzuai-oryx/BiMedIX>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, Yufang Wu

Advances in automated essay scoring (AES) have traditionally relied on labeled essays, requiring tremendous cost and expertise for their acquisition. Recently, large language models (LLMs) have achieved great success in various tasks, but their potential is less explored in AES. In this paper, we show that our zero-shot prompting framework, Multi Trait Specialization (MTS), elicits LLMs' ample potential for essay scoring. In particular, we automatically decompose writing proficiency into distinct traits and generate scoring criteria for each trait. Then, an LLM is prompted to extract trait scores from several conversational rounds, each round scoring one of the traits based on the scoring criteria. Finally, we derive the overall score via trait averaging and min-max scaling. Experimental results on two benchmark datasets demonstrate that MTS consistently outperforms straightforward prompting (Vanilla) in average QWK across all LLMs and datasets, with maximum gains of 0.437 on TOEFL11 and 0.355 on ASAP. Additionally, with the help of MTS, the small-sized Llama2-13b-chat substantially outperforms ChatGPT, facilitating an effective deployment in real applications.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Eliciting Instruction-tuned Code Language Models' Capabilities to Utilize Auxiliary Function for Code Generation

Seonghyeon Lee, Suyeon Kim, Joonwon Jang, Heejae Chon, Dongha Lee, Hwanjo Yu

We study the code generation behavior of instruction-tuned models built on top of code pre-trained language models when they could access an auxiliary function to implement a function. We design several ways to provide auxiliary functions to the models by adding them to the query or providing a response prefix to incorporate the ability to utilize auxiliary functions with the instruction-following capability. Our experimental results show the effectiveness of combining the base models' auxiliary function utilization ability with the instruction following ability. In particular, the performance of adopting our approaches with the open-sourced language models surpasses that of the recent powerful language models, i.e., gpt-4o.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Translation of Multifaceted Data without Re-Training of Machine Translation Systems

Hyeyoun Moon, Seungyoon Lee, SeongTae Hong, Seungjin Lee, Chanjun Park, Heuseok Lim

Translating major language resources to build minor language resources becomes a widely-used approach. Particularly in translating complex data points composed of multiple components, it is common to translate each component separately. However, we argue that this practice often overlooks the interrelation between components within the same data point. To address this limitation, we propose a novel MT pipeline that considers the intra-data relation, in implementing MT for training data. In our MT pipeline, all the components in a data point are concatenated to form a single translation sequence and subsequently reconstructed to the data components after translation. We introduce a Catalyst Statement (CS) to enhance the intra-data relation, and Indicator Token (IT) to assist the decomposition of a translated sequence into its respective data components. Through our approach, we have achieved a considerable improvement in translation quality itself, along with its effectiveness as training data. Compared with the conventional approach that translates each data component separately, our method yields better training data that enhances the performance of the trained model by 2.690 points for the web page ranking (WPR) task, and 0.845 for the question generation (QG) task in the XGLUE benchmark.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimers Disease Questions with Scientific Literature

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, huan liu, Li Shen, Tianlong Chen

Recent advancements in large language models (LLMs) have achieved promising performances across various applications. Nonetheless, the ongoing challenge of integrating long-tail knowledge continues to impede the seamless adoption of LLMs in specialized domains. In this work, we introduce DALK, a.k.a. Dynamic Co-Augmentation of LLMs and KG, to address this limitation and demonstrate its ability on studying Alzheimer's Disease (AD), a specialized sub-field in biomedicine and a global health priority. With a synergized framework of LLM and KG mutually enhancing each other, we first leverage LLM to construct an evolving AD-specific knowledge graph (KG) sourced from AD-related scientific literature, and then we utilize a coarse-to-fine sampling method with a novel self-aware knowledge retrieval approach to select appropriate knowledge from the KG to augment LLM inference capabilities. The experimental results, conducted on our constructed AD question answering (ADQA) benchmark, underscore the efficacy of DALK. Additionally, we perform a series of detailed analyses that can offer valuable insights and guidelines for the emerging topic of mutually enhancing KG and LLM.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

EvoR: Evolving Retrieval for Code Generation

Hongjin SU, Shuyang Jiang, Yuhang Lai, Haoyuan Wu, Boao Shi, Che Liu, Qian Liu, Tao Yu

Recently the retrieval-augmented generation (RAG) has been successfully applied in code generation. However, existing pipelines for retrieval-augmented code generation (RACG) employ static knowledge bases with a single source, limiting the adaptation capabilities of Large Language Models (LLMs) to domains they have insufficient knowledge of. In this work, we develop a novel pipeline, EVOR, that employs the synchronous evolution of both queries and diverse knowledge bases. On two realistic settings where the external knowledge is required to solve code generation tasks, we compile four new datasets associated with frequently updated libraries and long-tail programming languages, named EVOR-BENCH. Extensive experiments demonstrate that EVOR achieves two to four times of execution accuracy compared to other methods such as Reflexion (Shinn et al., 2024), DocPrompting (Zhou et al., 2023), etc. We demonstrate that EVOR is flexible and can be easily combined with them to achieve further improvement. Further analysis reveals that EVOR benefits from the synchronous evolution of queries and documents and the diverse information sources in the knowledge base. We hope that our studies will inspire more insights into the design of advanced RACG pipelines in future research.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Learning Musical Representations for Music Performance Question Answering

Xingjian Diao, Chunhui Zhang, Tingxuan Wu, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, Soroush Vosoughi, Jiang Gui
Music performances are representative scenarios for audio-visual modeling. Unlike common scenarios with sparse audio, music performances continuously involve dense audio signals throughout. While existing multimodal learning methods on the audio-video QA demonstrate impressive capabilities on general scenarios, they are incapable of dealing with fundamental problems within the music performances: they underexplore the interaction between the multimodal signals in performance, and fail to consider the distinctive characteristics of instruments and music. Therefore, existing methods tend to inaccurately answer questions regarding musical performances. To bridge the above research gaps, first, given the intricate multimodal interconnectivity inherent to music data, our primary backbone is designed to incorporate multimodal interactions within the context of music; second, to enable the model to learn music characteristics, we annotate and release rhythmic and music sources in the current music datasets; third, for time-aware audio-visual modelling, we align the model's music predictions with the temporal dimension. Our experiments show state-of-the-art effects on the Music AVQA datasets. Our code is available at: <https://github.com/xid32/Amuse>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Predictive Multiplicity of Knowledge Graph Embeddings in Link Prediction

Yucheng Zhu, Nica Petyka, Mojtaba Nayyeri, Bo Xiong, Yunjie He, Evgeny Kharlamov, Steffen Staab

Knowledge graph embedding (KGE) models are often used to predict missing links for knowledge graphs (KGs). However, multiple KG embeddings can perform almost equally well for link prediction yet give conflicting predictions for unseen queries. This phenomenon is termed *predictive multiplicity* in the literature. It poses substantial risks for KGE-based applications in high-stake domains but has been overlooked in KGE research. We define predictive multiplicity in link prediction, introduce evaluation metrics and measure predictive multiplicity for representative KGE methods on commonly used benchmark datasets. Our empirical study reveals significant predictive multiplicity in link prediction, with 8% to 39% testing queries exhibiting conflicting predictions. We address this issue by leveraging voting methods from social choice theory, significantly mitigating conflicts by 66% to 78% in our experiments.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Temporal Fact Reasoning over Hyper-Relational Knowledge Graphs

Zifeng Ding, Jingcheng Wu, Jingpei Wu, Yan Xia, Bo Xiong, Volker Tresp

Stemming from traditional knowledge graphs (KGs), hyper-relational KGs (HKGs) provide additional key-value pairs (i.e., qualifiers) for each KG fact that help to better restrict the fact validity. In recent years, there has been an increasing interest in studying graph reasoning over HKGs. Meanwhile, as discussed in recent works that focus on temporal KGs (TKGs), world knowledge is ever-evolving, making it important to reason over temporal facts in KGs. Previous mainstream benchmark HKGs do not explicitly specify temporal information for each HKG fact. Therefore, almost all existing HKG reasoning approaches do not devise any module specifically for temporal reasoning. To better study temporal fact reasoning over HKGs, we propose a new type of data structure named hyper-relational TKG (HTKG). Every fact in an HTKG is coupled with a timestamp explicitly indicating its time validity. We develop two new benchmark HTKG datasets, i.e., Wiki-hy and YAGO-hy, and propose an HTKG reasoning model that efficiently models hyper-relational temporal facts. To support future research on this topic, we open-source our datasets and model.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

NormTab: Improving Symbolic Reasoning in LLMs Through Tabular Data Normalization

Md Mahadi Hasan Nahid, Davood Rafiei

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in parsing textual data and generating code. However, their performance in tasks involving tabular data, especially those requiring symbolic reasoning, faces challenges due to the structural variance and inconsistency in table cell values often found in web tables. In this paper, we introduce NormTab, a novel framework aimed at enhancing the symbolic reasoning performance of LLMs by normalizing web tables. We study table normalization as a stand-alone, one-time preprocessing step using LLMs to support symbolic reasoning on tabular data. Our experimental evaluation, conducted on challenging web table datasets such as WikiTableQuestion and TabFact, demonstrates that leveraging NormTab significantly improves symbolic reasoning performance, showcasing the importance and effectiveness of web table normalization for enhancing LLM-based symbolic reasoning tasks.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Can Large Language Models Identify Authorship?

Baixiang Huang, Canyu Chen, Kai Shu

The ability to accurately identify authorship is crucial for verifying content authenticity and mitigating misinformation. Large Language Models (LLMs) have demonstrated exceptional capacity for reasoning and problem-solving. However, their potential in authorship analysis remains under-explored. Traditional studies have depended on hand-crafted stylistic features, whereas state-of-the-art approaches leverage text embeddings from pre-trained language models. These methods, which typically require fine-tuning on labeled data, often suffer from performance degradation in cross-domain applications and provide limited explainability. This work seeks to address three research questions: (1) Can LLMs perform zero-shot, end-to-end authorship verification effectively? (2) Are LLMs capable of accurately attributing authorship among multiple candidates authors (e.g., 10 and 20)? (3) Can LLMs provide explainability in authorship analysis, particularly through the role of linguistic features? Moreover, we investigate the integration of explicit linguistic features to guide LLMs in their reasoning processes. Our assessment demonstrates LLMs' proficiency in both tasks without the need for domain-specific fine-tuning, providing explanations into their decision making via a detailed analysis of linguistic features. This establishes a new benchmark for future research on LLM-based authorship analysis.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Emosical: An Emotion Annotated Musical Theatre Dataset

Hayoon Kim, Ahyeon Choi, Sungho Lee, Hyun Jin Jung, Kyogu Lee

This paper presents Emosical, a multimodal open-source dataset of musical films. Emosical comprises video, vocal audio, text, and character identity paired samples with annotated emotion tags. Emosical provides rich emotion annotations for each sample by inferring the background story of the characters. To achieve this, we leverage the musical theatre script, which contains the characters' complete background stories and narrative contexts. The annotation pipeline includes feeding the speaking character, text, global persona, and context of the dialogue and song track into a large language model. To verify the effectiveness of our tagging scheme, we perform an ablation study by bypassing each step of the pipeline. The ablation results show the usefulness of each component in generating accurate emotion tags. A subjective test is conducted to compare the generated tags of each ablation result. We also perform a statistical analysis to find out the global characteristics of the collected emotion tags. Emosical would enable expressive synthesis and tagging of the speech and singing voice in the musical theatre domain in future research. Emosical is publicly available at <https://github.com/gillosoe/emosical>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Geneverse: A Collection of Open-source Multimodal Large Language Models for Genomic and Proteomic Research

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, Hongyu Zhao

The applications of large language models (LLMs) are promising for biomedical and healthcare research. Despite the availability of open-source LLMs trained using a wide range of biomedical data, current research on the applications of LLMs to genomics and proteomics is still limited. To fill this gap, we propose a collection of finetuned LLMs and multimodal LLMs (MLLMs), known as Geneverse, for three novel tasks in genomic and proteomic research. The models in Geneverse are trained and evaluated based on domain-specific datasets, and we use advanced parameter-efficient finetuning techniques to achieve the model adaptation for tasks including the generation of descriptions for gene functions, protein function inference from its structure, and marker gene selection from spatial transcriptomic data. We demonstrate that adapted LLMs and MLLMs perform well for these tasks and may outperform closed-source large-scale models based on our evaluations focusing on both truthfulness and structural correctness. All of the training strategies and base models we used are freely accessible. Our codes can be found at <https://github.com>HelloWorldITY/Geneverse>.

Nov 14 (Thu) 14:00-15:30 - Riverfront Hall

Generating Media Background Checks for Automated Source Critical Reasoning

Michael Sejr Schlichtkrull

Not everything on the internet is true. This unfortunate fact requires both humans and models to perform complex reasoning about credibility when working with retrieved information. In NLP, this problem has seen little attention. Indeed, retrieval-augmented models are not typically expected to distrust retrieved documents. Human experts overcome the challenge by gathering signals about the context, reliability, and tendency of source documents - that is, they perform *source criticism*. We propose a novel NLP task focused on finding and summarising such signals. We introduce a new dataset of 6,709 "media background checks" derived from Media Bias / Fact Check, a volunteer-run website documenting media bias. We test open-source and closed-source LLM baselines with and without retrieval on this dataset, finding that retrieval greatly improves performance. We furthermore carry out human evaluation, demonstrating that 1) media background checks are helpful for humans, and 2) media background checks are helpful for retrieval-augmented models.

Virtual Poster Session 1 - (Nov 12): 17:4518:45 (Evening)**Computational Social Science and Cultural Analytics**

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter

Katryna Kasianenko, Shima Khanehzar, Stephen Wan, Ehsan Dehghan, Axel Bruns

Communities on social media display distinct patterns of linguistic expression and behaviour, collectively referred to as practices. These practices can be traced in textual exchanges, and reflect the intentions, knowledge, values, and norms of users and communities. This paper introduces a comprehensive methodological workflow for computational identification of such practices within social media texts. By focusing on supporters of Ukraine during the Russia-Ukraine war in (1) the activist collective NAFO and (2) the Eurovision Twitter community, we present a gold-standard data set capturing their unique practices. Using this corpus, we perform practice prediction experiments with both open-source baseline models and OpenAI's large language models (LLMs). Our results demonstrate that closed-source models, especially GPT-4, achieve superior performance, particularly with prompts that incorporate salient features of practices, or utilize Chain-of-Thought prompting. This study provides a detailed error analysis and offers valuable insights into improving the precision of practice identification, thereby supporting context-sensitive moderation and advancing the understanding of online community dynamics.

(Nov 12): 17:4518:45 (Evening) - Gather

I love pineapple on pizza != I hate pineapple on pizza: Stance-Aware Sentence Transformers for Opinion Mining

Vahid Ghafouri, Jose M. Such, Guillermo Suarez-Tangil

Sentence transformers excel at grouping topically similar texts, but struggle to differentiate opposing viewpoints on the same topic. This shortcoming hinders their utility in applications where understanding nuanced differences in opinion is essential, such as those related to social and political discourse analysis. This paper addresses this issue by fine-tuning sentence transformers with arguments for and against human-generated controversial claims. We demonstrate how our fine-tuned model enhances the utility of sentence transformers for social computing tasks such as opinion mining and stance detection. We elaborate that applying stance-aware sentence transformers to opinion mining is a more computationally efficient and robust approach in comparison to the classic classification-based approaches.

(Nov 12): 17:4518:45 (Evening) - Gather

IndoCulture: Exploring Geographically-Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces

Fajri Koto, Rahmad Mahendra, Nurul Aisyah Timothy Baldwin

Although commonsense reasoning is greatly shaped by cultural and geographical factors, previous studies on language models have predominantly centered on English cultures, potentially resulting in an Anglocentric bias. In this paper, we introduce IndoCulture, aimed at understanding the influence of geographical factors on language model reasoning ability, with a specific emphasis on the diverse cultures found within eleven Indonesian provinces. In contrast to prior works that relied on template (Yin et al., 2022) and online scrapping (Fung et al., 2024), we create IndoCulture by asking local people to manually develop the context and plausible options based on predefined topics. Evaluations of 27 language models reveal several insights: (1) the recent open-source model, LLaMA3, is as competitive as GPT-4, while other open-source models struggle with accuracies below 50%, (2) models often provide more accurate predictions for specific provinces, such as Bali and West Java, and (3) the inclusion of location contexts enhances performance, especially in larger models like GPT-4, emphasizing the significance of geographical context in commonsense reasoning.

Demo

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

AUTOGEN STUDIO: A No-Code Developer Tool for Building and Debugging Multi-Agent Systems

Adam Fournary, Chi Wang, Erkang Zhu, Gagan Bansal, Jingya Chen, Saleema Amershi, Saff Syed, Victor Dibia

Multi-agent systems, where multiple agents (generative AI models + tools) collaborate, are emerging as an effective pattern for solving long-running, complex tasks in numerous domains. However, specifying their parameters (such as models, tools, and orchestration mechanisms etc.) and debugging them remains challenging for most developers. To address this challenge, we present AUTOGEN STUDIO, a no-code developer tool for rapidly prototyping, debugging, and evaluating multi-agent work-flows built upon the AUTOGEN framework. AUTOGEN STUDIO offers a web interface and a Python API for representing LLM-enabled agents using a declarative (JSON-based) specification. It provides an intuitive drag-and-drop UI for agent workflow specification, interactive evaluation and debugging of workflows, and a gallery of reusable agent components. We highlight four design principles for no-code multi-agent developer tools and contribute an open-source implementation. <https://github.com/microsoft/autogen/tree/autogenstudio/samples/apps/autogen-studio>

Dialogue and Interactive Systems

(Nov 12): 17:45 18:45 (Evening) - Room: Gather

(Nov 12): 17:45 18:45 (Evening) - Gather

Inductive-Deductive Strategy Reuse for Multi-Turn Instructional Dialogues

Jiao Ou, Jiayi Wu, Che Liu, Fuzheng Zhang, Di ZHANG, Kun Gai

Aligning large language models (LLMs) with human expectations requires high-quality instructional dialogues, which can be achieved by raising diverse, in-depth, and insightful instructions that deepen interactions. Existing methods target instructions from real instruction dialogues as a learning goal and fine-tune a user simulator for posing instructions. However, the user simulator struggles to implicitly model complex dialogue flows and pose high-quality instructions. In this paper, we take inspiration from the cognitive abilities inherent in human learning and propose the explicit modeling of complex dialogue flows through instructional strategy reuse. Specifically, we first induce high-level strategies from various real instruction dialogues. These strategies are applied to new dialogue scenarios deductively, where the instructional strategies facilitate high-quality instructions. Experimental results show that our method can generate diverse, in-depth, and insightful instructions for a given dialogue history. The constructed multi-turn instructional dialogues can outperform competitive baselines on the downstream chat model.

Ethics, Bias, and Fairness

(Nov 12): 17:45 18:45 (Evening) - Room: Gather

(Nov 12): 17:45 18:45 (Evening) - Gather

Evaluating Psychological Safety of Large Language Models

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, Lidong Bing

In this work, we designed unbiased prompts to systematically evaluate the psychological safety of large language models (LLMs). First, we tested five different LLMs by using two personality tests: Short Dark Triad (SD-3) and Big Five Inventory (BFI). All models scored higher than the human average on SD-3, suggesting a relatively darker personality pattern. Despite being instruction fine-tuned with safety metrics to reduce toxicity, InstructGPT, GPT-3.5, and GPT-4 still showed dark personality patterns; these models scored higher than self-supervised GPT-3 on the Machiavellianism and narcissism traits on SD-3. Then, we evaluated the LLMs in the GPT series by using well-being tests to study the impact of fine-tuning with more training data. We observed a continuous increase in the well-being scores of GPT models. Following these observations, we showed that fine-tuning Llama-2-chat-7B with responses from BFI using direct preference optimization could effectively reduce the psychological toxicity of the model. Based on the findings, we recommended the application of systematic and comprehensive psychological metrics to further evaluate and improve the safety of LLMs.

(Nov 12): 17:45 18:45 (Evening) - Gather

Intrinsic Self-correction for Enhanced Morality: An Analysis of Internal Mechanisms and the Superficial Hypothesis

Guangliang Liu, Haitao Mao, Jiliang Tang, Kristen Johnson

Large Language Models (LLMs) are capable of producing content that perpetuates stereotypes, discrimination, and toxicity. The recently proposed *moral self-correction* is a computationally efficient method for reducing harmful content in the responses of LLMs. However, the process of how injecting self-correction instructions can modify the behavior of LLMs remains under-explored. In this paper, we explore the effectiveness of moral self-correction by answering three research questions: (1) In what scenarios does moral self-correction work? (2) What are the internal mechanisms of LLMs, e.g., hidden states, that are influenced by moral self-correction instructions? (3) Is intrinsic moral self-correction actually superficial in terms of reduced immorality in hidden states? We argue that self-correction can help LLMs find a shortcut to more morally correct output, rather than truly reducing the immorality stored in hidden states. Through empirical investigation with tasks of language generation and multi-choice question answering, we conclude: (i) LLMs exhibit good performance across both tasks, and self-correction instructions are particularly beneficial when the correct answer is already top-ranked; (ii) The morality levels in intermediate hidden states are strong indicators as to whether one instruction would be more effective than another; (iii) Based on our analysis of intermediate hidden states and task case studies of self-correction behaviors, we are first to propose the hypothesis that intrinsic moral self-correction is in fact superficial.

(Nov 12): 17:45 18:45 (Evening) - Gather

A Study of Implicit Ranking Unfairness in Large Language Models

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, Tat-Seng Chua

Recently, Large Language Models (LLMs) have demonstrated a superior ability to serve as ranking models. However, concerns have arisen as LLMs will exhibit discriminatory ranking behaviors based on users' sensitive attributes (e.g. gender). Worse still, in this paper, we identify a subtler form of discrimination in LLMs, termed *implicit ranking unfairness*, where LLMs exhibit discriminatory ranking patterns based solely on non-sensitive user profiles, such as user names. Such implicit unfairness is more widespread but less noticeable, threatening the ethical foundation. To comprehensively explore such unfairness, our analysis will focus on three research aspects: (1) We propose an evaluation method to investigate the severity of implicit ranking unfairness. (2) We uncover the reasons for causing such unfairness. (3) To mitigate such unfairness effectively, we utilize a pair-wise regression method to conduct fair-aware data augmentation for LLM fine-tuning. The experiment demonstrates that our method outperforms existing approaches in ranking fairness, achieving this with only a small reduction in accuracy. Lastly, we emphasize the need for the community to identify and mitigate the implicit unfairness, aiming to avert the potential deterioration

in the reinforced human-LLMs ecosystem deterioration.

Generation 1

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Llama SLayer 8B: Shallow Layers Hold the Key to Knowledge Injection

Tianxiang Chen, Zhenhao Tan, Tao Gong, Yue Wu, Qi Chu, Bin Liu, Jieping Ye, Nenghai Yu

As a manner to augment pretrained large language models (LLM), knowledge injection is critical to develop vertical domain large models and has been widely studied. While most current approaches, including parameter-efficient fine-tuning (PEFT) and block expansion methods, uniformly apply knowledge across all LLM layers, it raises the question: are all layers equally crucial for knowledge injection? We embark upon evaluating the importance of each layer to locate the optimal layer range for knowledge injection. Intuitively, more important layers should play more critical roles in knowledge injection and deserve denser injection. We observe performance dips in question-answering benchmarks after the removal or expansion of the shallow layers, and the degradation shrinks as the layer gets deeper, indicating that the shallow layers hold the key to knowledge injection. This insight leads us to propose the S strategy, a post-pretraining strategy of selectively enhancing shallow layers while pruning the less effective deep ones. Based on this strategy, we introduce Llama Slayer 8B. We experimented on the corpus of code & math and demonstrated the effectiveness of our strategy. Further experiments across different LLM, Mistral-7B, and a legal corpus confirmed the approach's general applicability, underscoring its wide-ranging efficacy.

(Nov 12): 17:4518:45 (Evening) - Gather

TRACE the Evidence: Constructing Knowledge-Grounded Reasoning Chains for Retrieval-Augmented Generation

Jinyuan Fang, Zaiqiao Meng, Craig MacDonald

Retrieval-augmented generation (RAG) offers an effective approach for addressing question answering (QA) tasks. However, the imperfections of the retrievers in RAG models often result in the retrieval of irrelevant information, which could introduce noise and degrade the performance, especially when handling multi-hop questions that require multiple steps of reasoning. To enhance the multi-hop reasoning ability of RAG models, we propose TRACE. TRACE constructs knowledge-grounded reasoning chains, which are a series of logically connected knowledge triples, to identify and integrate supporting evidence from the retrieved documents for answering questions. Specifically, TRACE employs a KG Generator to create a knowledge graph (KG) from the retrieved documents, and then uses a novel Autoregressive Reasoning Chain Constructor to build reasoning chains. Experimental results on three multi-hop QA datasets show that TRACE achieves an average performance improvement of up to 14.03% compared to using all the retrieved documents. Moreover, the results indicate that using reasoning chains as context, rather than the entire documents, is often sufficient to correctly answer questions.

(Nov 12): 17:4518:45 (Evening) - Gather

The Effect of Sampling Temperature on Problem Solving in Large Language Models

Matthew Renze

In this research study, we empirically investigate the effect of sampling temperature on the performance of Large Language Models (LLMs) on various problem-solving tasks. We created a multiple-choice question-and-answer (MCQA) exam by randomly sampling problems from standard LLM benchmarks. Then, we used nine popular LLMs with five prompt-engineering techniques to solve the MCQA problems while increasing the sampling temperature from 0.0 to 1.6. Despite anecdotal reports to the contrary, our empirical results indicate that changes in temperature from 0.0 to 1.0 do not have a statistically significant impact on LLM performance for problem-solving tasks. In addition, these results appear to generalize across LLMs, prompt-engineering techniques, and problem domains. All code, data, and supplemental materials are available on GitHub at: <https://github.com/matthewrenze/jhu-llm-temperature>

Information Extraction

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Breaking the Boundaries: A Unified Framework for Chinese Named Entity Recognition Across Text and Speech

Jinzhong Ning, Yuanyuan Sun, Bo Xu, Zhihao Yang, Ling Luo, Hongfei Lin

In recent years, with the vast and rapidly increasing amounts of spoken and textual data, Named Entity Recognition (NER) tasks have evolved into three distinct categories, i.e., text-based NER (TNER), Speech NER (SNER) and Multimodal NER (MNER). However, existing approaches typically require designing separate models for each task, overlooking the potential connections between tasks and limiting the versatility of NER methods. To mitigate these limitations, we introduce a new task named Integrated Multimodal NER (IMNER) to break the boundaries between different modal NER tasks, enabling a unified implementation of them. To achieve this, we first design a unified data format for inputs from different modalities. Then, leveraging the pre-trained MMSpeech model as the backbone, we propose an ***Integrated ***M***ultimod***q***l ***Ge***peration Framework (**IMAGE**), formulating the Chinese IMNER task as an entity-aware text generation task. Experimental results demonstrate the feasibility of our proposed IMAGE framework in the IMNER task. Our work in integrated multimodal learning in advancing the performance of NER may set up a new direction for future research in the field. Our source code is available at <https://github.com/NingJinzhong/IMAGE4IMNER>.

(Nov 12): 17:4518:45 (Evening) - Gather

Generative Models for Automatic Medical Decision Rule Extraction from Text

Yuxin He, Buzhou Tang, Xiaoling Wang

Medical decision rules play a key role in many clinical decision support systems (CDSS). However, these rules are conventionally constructed by medical experts, which is expensive and hard to scale up. In this study, we explore the automatic extraction of medical decision rules from text, leading to a solution to construct large-scale medical decision rules. We adopt a formulation of medical decision rules as binary trees consisting of condition/decision nodes. Such trees are referred to as medical decision trees and we introduce several generative models to extract them from text. The proposed models inherit the merit of two categories of successful natural language generation frameworks, i.e., sequence-to-sequence generation and autoregressive generation. To unleash the potential of pretrained language models, we design three styles of linearization (natural language, augmented natural language and JSON code), acting as the target sequence for our models. Our final system achieves 67% tree accuracy on a comprehensive Chinese benchmark, outperforming state-of-the-art baseline by 12%. The result

demonstrates the effectiveness of generative models on explicitly modeling structural decision-making roadmaps, and shows great potential to boost the development of CDSS and explainable AI. Our code will be open-source upon acceptance.

Information Retrieval and Text Mining

(Nov 12): 17:45 18:45 (Evening) - Room: Gather

(Nov 12): 17:45 18:45 (Evening) - Gather

Consolidating Ranking and Relevance Predictions of Large Language Models through Post-Processing

Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, Harrie Oosterhuis

The powerful generative abilities of large language models (LLMs) show potential in generating relevance labels for search applications. Previous work has found that directly asking about relevancy, such as “*How relevant is document A to query Q?**”, results in suboptimal ranking. Instead, the pairwise-ranking prompting (PRP) approach produces promising ranking performance through asking about pairwise comparisons, e.g., “*Is document A more relevant than document B to query Q?**”. Thus, while LLMs are effective at their ranking ability, this is not reflected in their relevance label generation. In this work, we propose a post-processing method to consolidate the relevance labels generated by an LLM with its powerful ranking abilities. Our method takes both LLM generated relevance labels and pairwise preferences. The labels are then altered to satisfy the pairwise preferences of the LLM, while staying as close to the original values as possible. Our experimental results indicate that our approach effectively balances label accuracy and ranking performance. Thereby, our work shows it is possible to combine both the ranking and labeling abilities of LLMs through post-processing.

(Nov 12): 17:45 18:45 (Evening) - Gather

Multi-Granularity History and Entity Similarity Learning for Temporal Knowledge Graph Reasoning

Shi Mingcong, Chunjiang Zhu, Detian Zhang, Shiting Wen, Qing Li

Temporal Knowledge Graph (TKG) reasoning, aiming to predict future unknown facts based on historical information, has attracted considerable attention due to its great practical value. Insight into history is the key to predict the future. However, most existing TKG reasoning models singly capture repetitive history, ignoring the entity's multi-hop neighbour history which can provide valuable background knowledge for TKG reasoning. In this paper, we propose Multi-Granularity History and Entity Similarity Learning (MGESL) model for Temporal Knowledge Graph Reasoning, which models historical information from both coarse-grained and fine-grained history. Since similar entities tend to exhibit similar behavioural patterns, we also design a hypergraph convolution aggregator to capture the similarity between entities. Furthermore, we introduce a more realistic setting for the TKG reasoning, where candidate entities are already known at the timestamp to be predicted. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our proposed model.

(Nov 12): 17:45 18:45 (Evening) - Gather

PepRec: Progressive Enhancement of Prompting for Recommendation

Yakun Yu, Shi-ang Qi, Baochun Li, Di Niu

With large language models (LLMs) achieving remarkable breakthroughs in natural language processing (NLP) domains, recent researchers have actively explored the potential of LLMs for recommendation systems by converting the input data into textual sentences through prompt templates. Although semantic knowledge from LLMs can help enrich the content information of items, to date it is still hard for them to achieve comparable performance to traditional deep learning recommendation models, partly due to a lack of ability to leverage collaborative filtering. In this paper, we propose a novel training-free prompting framework, PepRec, which aims to capture knowledge from both content-based filtering and collaborative filtering to boost recommendation performance with LLMs, while providing interpretation for the recommendation. Experiments based on two real-world datasets from different domains show that PepRec significantly outperforms various traditional deep learning recommendation models and prompt-based recommendation systems.

Interpretability and Analysis of Models for NLP

(Nov 12): 17:45 18:45 (Evening) - Room: Gather

(Nov 12): 17:45 18:45 (Evening) - Gather

Adaptive Immune-based Sound-Shape Code Substitution for Adversarial Chinese Text Attacks

Ao Wang, Xinghua Yang, Chen Li, Bao-di Liu, Weifeng Liu

Adversarial textual examples reveal the vulnerability of natural language processing (NLP) models. Most existing text attack methods are designed for English text, while the robust implementation of the second popular language, i.e., Chinese with 1 billion users, is greatly underestimated. Although several Chinese attack methods have been presented, they either directly transfer from English attacks or adopt simple greedy search to optimize the attack priority, usually leading to unnatural sentences. To address these issues, we propose an adaptive Immune-based Sound-Shape Code (ISSC) algorithm for adversarial Chinese text attacks. Firstly, we leverage the Sound-Shape code to generate natural substitutions, which comprehensively integrate multiple Chinese features. Secondly, we employ adaptive immune algorithm (IA) to determine the replacement order, which can reduce the duplication of population to improve the search ability. Extensive experimental results validate the superiority of our ISSC in producing high-quality Chinese adversarial texts. Our code and data can be found in <https://github.com/nohuma/chinese-attack-issc>.

(Nov 12): 17:45 18:45 (Evening) - Gather

A Coordinate System for In-Context Learning

Anhao Zhao, Fanghua Ye, Jinlan Fu, Xiaoyu Shen

Large language models (LLMs) exhibit remarkable in-context learning (ICL) capabilities. However, the underlying working mechanism of ICL remains poorly understood. Recent research presents two conflicting views on ICL: One emphasizes the impact of similar examples in the demonstrations, stressing the need for label correctness and more shots. The other attributes it to LLMs' inherent ability of task recognition, deeming label correctness and shot numbers of demonstrations as not crucial. In this work, we provide a Two-Dimensional Coordinate System that unifies both views into a systematic framework. The framework explains the behavior of ICL through two orthogonal variables: whether similar examples are presented in the demonstrations (perception) and whether LLMs can recognize the task (cognition). We propose the peak inverse rank metric to detect the task recognition ability of LLMs and study LLMs' reactions to different definitions of similarity. Based on these, we conduct extensive experiments to elucidate how ICL functions across each quadrant on multiple representative classification tasks. Finally, we extend our analyses to generation tasks, showing that our coordinate system can also be used to interpret ICL for generation tasks

effectively.

(Nov 12): 17:4518:45 (Evening) - Gather

Calibrating the Confidence of Large Language Models by Eliciting Fidelity

MoZhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, Xipeng Qiu

Large language models optimized with techniques like RLHF have achieved good alignment in being helpful and harmless. However, post-alignment, these language models often exhibit overconfidence, where the expressed confidence does not accurately calibrate with their correctness rate. In this paper, we decompose the language model confidence into the *Uncertainty* about the question and the *Fidelity* to the answer generated by language models. Then, we propose a plug-and-play method, *UF Calibration*, to estimate the confidence of language models. Our method has shown good calibration performance by conducting experiments with 6 RLHF-LMs on four MCQA datasets. Moreover, we propose two novel metrics, IPR and CE, to evaluate the calibration of the model, and we have conducted a detailed discussion on *Truly Well-Calibrated Confidence* for large language models. Our method could serve as a strong baseline, and we hope that this work will provide some insights into the model confidence calibration.

(Nov 12): 17:4518:45 (Evening) - Gather

Formality Favored: Unraveling the Learning Preferences of Large Language Models on Data with Conflicting Knowledge

Jiahuan Li, Yiqing Cao, Shujian Huang, Jiajun Chen

Having been trained on massive pretraining data, large language models have shown excellent performance on many knowledge-intensive tasks. However, pretraining data tends to contain misleading and even conflicting information, and it is intriguing to understand how LLMs handle these noisy data during training. In this study, we systematically analyze LLMs learning preferences for data with conflicting knowledge. We find that pretrained LLMs establish learning preferences similar to humans, i.e., preferences towards formal texts and texts with fewer spelling errors, resulting in faster learning and more favorable treatment of knowledge in data with such features when facing conflicts. This finding is generalizable across models and languages and is more evident in larger models. An in-depth analysis reveals that LLMs tend to trust data with features that signify consistency with the majority of data, and it is possible to instill new preferences and erase old ones by manipulating the degree of consistency with the majority data.

Language Modeling

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Instruction Pre-Training: Language Models are Supervised Multitask Learners

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, Furui Wei

Unsupervised multitask pre-training has been the critical method behind the recent success of language models (LMs). However, supervised multitask learning still holds significant promise, as scaling it in the post-training stage trends towards better generalization. In this paper, we explore supervised multitask pre-training by proposing Instruction Pre-training, a framework that scalably augments massive raw corpora with instruction-response pairs to pre-train LMs. The instruction-response pairs are generated by an efficient instruction synthesizer built on open-source models. In our experiments, we synthesize 200M instruction response pairs covering 40+ task categories to verify the effectiveness of Instruction Pre-training. In pre-training from scratch, Instruction Pre-training not only consistently enhances pre-trained base models but also benefits more from further instruction tuning. In continual pre-training, Instruction Pre-training enables Llama3-8B to be comparable to or even outperform Llama3-70B. Our model, code, and data are available at <https://github.com/microsoft/LMOps>.

(Nov 12): 17:4518:45 (Evening) - Gather

Mitigating Training Imbalance in LLM Fine-Tuning via Selective Parameter Merging

Yining Ju, Ziyi Ni, Xingrun Xing, Zhixiong Zeng, hanyu Zhao, Sitqi Fan, Zheng Zhang

Supervised fine-tuning (SFT) is crucial for adapting Large Language Models (LLMs) to specific tasks. In this work, we demonstrate that the order of training data can lead to significant training imbalances, potentially resulting in performance degradation. Consequently, we propose to mitigate this imbalance by merging SFT models fine-tuned with different data orders, thereby enhancing the overall effectiveness of SFT. Additionally, we introduce a novel technique, "parameter-selection merging," which outperforms traditional weighted-average methods on five datasets. Further, through analysis and ablation studies, we validate the effectiveness of our method and identify the sources of performance improvements.

(Nov 12): 17:4518:45 (Evening) - Gather

Mixture-of-Skills: Learning to Optimize Data Usage for Fine-Tuning Large Language Models

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, Reza Haf

Large language models (LLMs) are typically fine-tuned on diverse and extensive datasets sourced from various origins to develop a comprehensive range of skills, such as writing, reasoning, chatting, coding, and more. Each skill has unique characteristics, and these datasets are often heterogeneous and imbalanced, making the fine-tuning process highly challenging. Balancing the development of each skill while ensuring the model maintains its overall performance requires sophisticated techniques and careful dataset curation. In this work, we propose a general, model-agnostic, reinforcement learning framework, Mixture-of-Skills (MoS), that learns to optimize data usage automatically during the fine-tuning process. This framework ensures the optimal comprehensive skill development of LLMs by dynamically adjusting the focus on different datasets based on their current learning state. To validate the effectiveness of MoS, we conduct extensive experiments using three diverse LLM backbones on two widely used benchmarks and demonstrate that MoS substantially enhances model performance. Building on the success of MoS, we propose MoSpec, an adaptation for task-specific fine-tuning, which harnesses the utilities of various datasets for a specific purpose. Our work underlines the significance of dataset rebalancing and present MoS as a powerful, general solution for optimizing data usage in the fine-tuning of LLMs for various purposes.

(Nov 12): 17:4518:45 (Evening) - Gather

LongHeads: Multi-Head Attention is Secretly a Long Context Processor

Yi Lu, Xin Zhou, Wei He, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, Xuanjing Huang

Large language models (LLMs) have achieved impressive performance in numerous domains but often struggle to process lengthy inputs effectively and efficiently due to limited length generalization and attention quadratic computational demands. Many sought to mitigate this by restricting the attention window within the pre-trained length. However, these methods introduce new issues such as ignoring the middle context and requiring additional training. To address these problems, we propose LongHeads, a training-free framework that enhances LLMs long context ability by unlocking multi-head attention untapped potential. Instead of allowing each head to attend to the full sentence, which struggles with generalizing to longer sequences due to out-of-distribution (OOD) issues, we allow each head to process in-distribution length

by selecting and attending to important context chunks. To this end, we propose a chunk selection strategy that relies on the inherent correlation between the query and the key representations, efficiently distributing context chunks to different heads. In this way, each head ensures it can effectively process attended tokens within the trained length, while different heads in different layers can collectively process longer contexts. LongHeads works efficiently and fits seamlessly with many LLMs that use relative positional encoding. LongHeads achieves 100% accuracy at the 128k length on passkey retrieval task, verifying LongHeads' efficacy in extending the usable context window for existing models.

Linguistic Theories, Cognitive Modeling and Psycholinguistics

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Beyond Agreement: Diagnosing the Rationale Alignment of Automated Essay Scoring Methods based on Linguistically-informed Counterfactuals

Yupei Wang, Renfen Hu, Zhe Zhao

While current Automated Essay Scoring (AES) methods demonstrate high scoring agreement with human raters, their decision-making mechanisms are not fully understood. Our proposed method, using counterfactual intervention assisted by Large Language Models (LLMs), reveals that BERT-like models primarily focus on sentence-level features, whereas LLMs such as GPT-3.5, GPT-4 and Llama-3 are sensitive to conventions & accuracy, language complexity, and organization, indicating a more comprehensive rationale alignment with scoring rubrics. Moreover, LLMs can discern counterfactual interventions when giving feedback on essays. Our approach improves understanding of neural AES methods and can also apply to other domains seeking transparency in model-driven decisions.

Low-resource Methods for NLP

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, Kiet Van Nguyen

Vietnamese, a low-resource language, is typically categorized into three primary dialect groups that belong to Northern, Central, and Southern Vietnam. However, each province within these regions exhibits its own distinct pronunciation variations. Despite the existence of various speech recognition datasets, none of them has provided a fine-grained classification of the 63 dialects specific to individual provinces of Vietnam. To address this gap, we introduce Vietnamese Multi-Dialect (ViMD) dataset, a novel comprehensive dataset capturing the rich diversity of 63 provincial dialects spoken across Vietnam. Our dataset comprises 102.56 hours of audio, consisting of approximately 19,000 utterances, and the associated transcripts contain over 1.2 million words. To provide benchmarks and simultaneously demonstrate the challenges of our dataset, we fine-tune state-of-the-art pre-trained models for two downstream tasks: (1) Dialect identification and (2) Speech recognition. The empirical results suggest two implications including the influence of geographical factors on dialects, and the constraints of current approaches in speech recognition tasks involving multi-dialect speech data. Our dataset is available for research purposes.

(Nov 12): 17:4518:45 (Evening) - Gather

Characterizing Text Datasets with Psycholinguistic Features

Marcio Monteiro, Charu Karakapparambil James, Marius Kloft, Sophie Fellenz

Fine-tuning pretrained language models on task-specific data is a common practice in Natural Language Processing (NLP) applications. However, the number of pretrained models available to choose from can be very large, and it remains unclear how to select the optimal model without spending considerable amounts of computational resources, especially for the text domain. To address this problem, we introduce PsyMatrix, a novel framework designed to efficiently characterize text datasets. PsyMatrix evaluates multiple dimensions of text and discourse, producing interpretable, low-dimensional embeddings. Our framework has been tested using a meta-dataset repository that includes the performance of 24 pretrained large language models fine-tuned across 146 classification datasets. Using the proposed embeddings, we successfully developed a meta-learning system capable of recommending the most effective pretrained models (optimal and near-optimal) for fine-tuning on new datasets.

(Nov 12): 17:4518:45 (Evening) - Gather

Advancing Vision-Language Models with Adapter Ensemble Strategies

Yue Bai, Handong Zhao, Zhe Lin, Ajinky Kale, Jiaxiang Gu, Tong Yu, Sungchul Kim, Yun Fu

CLIP revolutes vision-language pretraining by using contrastive learning on paired web data. However, the sheer size of these pretrained models makes full-model finetuning exceedingly costly. One common solution is the "adapter", which finetunes a few additional parameters while freezing the backbone. It harnesses the heavy-duty backbone while offering a light finetuning for small downstream tasks. This synergy prompts us to explore the potential of augmenting large-scale backbones with traditional machine learning techniques. Often employed in traditional fields and overlooked in the large-scale era, these techniques could provide valuable enhancements. Herein, we delve into the "adapter ensembles" in the realm of large-scale pretrained vision-language models. We begin with a proof-of-concept study to establish the efficacy of combining multiple adapters. We then present extensive evidence showing these ensembles excel in a variety of settings, particularly when employing a Multi-Scale Attention (MSA) approach thoughtfully integrated into the ensemble framework. We further incorporate the LoRA to mitigate the additional parameter burden. We focus on vision-language retrieval, using different backbones under constraints of minimal data, parameters, and finetuning budgets. This research paves the way for a synergistic blend of traditional, yet effective, strategies with modern large-scale networks.

Machine Learning for NLP 1

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Optimizing Language Models with Fair and Stable Reward Composition in Reinforcement Learning

Jiahui Li, Hanlin Zhang, Fengda Zhang, Tai-Wei Chang, Kun Kuang, Long Chen, JUN ZHOU

Reinforcement learning from human feedback (RLHF) and AI-generated feedback (RLAIF) have become prominent techniques that significantly enhance the functionality of pre-trained language models (LMs). These methods harness feedback, sourced either from humans or AI, as direct rewards or to shape reward models that steer LM optimization. Nonetheless, the effective integration of rewards from diverse sources presents a significant challenge due to their disparate characteristics. To address this, recent research has developed algorithms incorporating strategies such as weighting, ranking, and constraining to handle this complexity. Despite these innovations, a bias toward disproportionately high rewards can still skew the reinforcement learning process and negatively impact LM performance. This paper explores a methodology for reward composition that enables simultaneous improvements in LMs across multiple dimensions. Inspired by fairness theory, we introduce a training algorithm that aims to reduce disparity and enhance stability among various rewards. Our method treats the aggregate reward as a dynamic weighted sum of individual rewards, with alternating updates to the weights and model parameters. For efficient and straightforward implementation, we employ an estimation technique rooted in the mirror descent method for weight updates, eliminating the need for gradient computations. The empirical results under various types of rewards across a wide range of scenarios demonstrate the effectiveness of our method.

(Nov 12): 17:4518:45 (Evening) - Gather

Leveraging BERT and TFIDF Features for Short Text Clustering via Alignment-Promoting Co-Training

Zetong Li, Qianliang Su, Shijing Si, Jianxing Yu

BERT and TFIDF features excel in capturing rich semantics and important words, respectively. Since most existing clustering methods are solely based on the BERT model, they often fall short in utilizing keyword information, which, however, is very useful in clustering short texts. In this paper, we propose a **CO***-*T*-training **C**-clustering (**COTC**) framework to make use of the collective strengths of BERT and TFIDF features. Specifically, we develop two modules responsible for the clustering of BERT and TFIDF features, respectively. We use the deep representations and cluster assignments from the TFIDF module outputs to guide the learning of the BERT module, seeking to align them at both the representation and cluster levels. Reversely, we also use the BERT module outputs to train the TFIDF module, thus leading to the mutual promotion. We then show that the alternating co-training framework can be placed under a unified joint training objective, which allows the two modules to be connected tightly and the training signals to be propagated efficiently. Experiments on eight benchmark datasets show that our method outperforms current SOTA methods significantly.

(Nov 12): 17:4518:45 (Evening) - Gather

One-to-Many Communication and Compositionality in Emergent Communication

Heeyoung Lee

Compositional languages leverage rules that derive meaning from combinations of simpler constituents. This property is considered to be the hallmark of human language as it enables the ability to express novel concepts and ease of learning. As such, numerous studies in the emergent communication field explore the prerequisite conditions for emergence of compositionality. Most of these studies set out one-to-one communication environment wherein a speaker interacts with a single listener during a single round of communication game. However, real-world communications often involve multiple listeners; their interests may vary and they may even need to coordinate among themselves to be successful at a given task. This work investigates the effects of one-to-many communication environment on emergent languages where a single speaker broadcasts its message to multiple listeners to cooperatively solve a task. We observe that simply broadcasting the speaker's message to multiple listeners does not induce more compositional languages. We then find and analyze two axes of environmental pressures that facilitate emergence of compositionality: listeners of *different interests* and *coordination* among listeners.

(Nov 12): 17:4518:45 (Evening) - Gather

MuMath-Code: Combining Tool-Use Large Language Models with Multi-perspective Data Augmentation for Mathematical Reasoning

Shuo You, Weihao You, Zhilong Ji, Guoqiang Zhong, Jinfeng Bai

The tool-use Large Language Models (LLMs) that integrate with external Python interpreters have significantly enhanced mathematical reasoning capabilities for open-source LLMs, while tool-free methods chose another track: augmenting math reasoning data. However, a great method to integrate the above two research paths and combine their advantages remains to be explored. In this work, we firstly include new math questions via **mu**-ti-perspective data augmenting methods and then synthesize **code**-nested solutions to them. The open LLMs (e.g., Llama-2) are finetuned on the augmented dataset to get the resulting models, **MuMath-Code** (μ -Math-Code). During the inference phase, our MuMath-Code generates code and interacts with the external python interpreter to get the execution results. Therefore, MuMath-Code leverages the advantages of both the external tool and data augmentation. To fully leverage the advantages of our augmented data, we propose a two-stage training strategy: In Stage-1, we finetune Llama-2 on pure CoT data to get an intermediate model, which then is trained on the code-nested data in Stage-2 to get the resulting MuMath-Code. Our MuMath-Code-7B achieves 83.8% on GSM8K and 52.4% on MATH, while MuMath-Code-70B model achieves new state-of-the-art performance among open methods—achieving 90.7% on GSM8K and 55.1% on MATH. Extensive experiments validate the combination of tool use and data augmentation, as well as our two-stage training strategy. We release the proposed dataset along with the associated code for public use: <https://github.com/youweihao-tal/MuMath-Code>.

(Nov 12): 17:4518:45 (Evening) - Gather

Revisiting Supervised Contrastive Learning for Microblog Classification

Junbo Huang, Ricardo Usbeck

Microblog content (e.g., Tweets) is noisy due to its informal use of language and its lack of contextual information within each post. To tackle these challenges, state-of-the-art microblog classification models rely on pre-training language models (LMs). However, pre-training dedicated LMs is resource-intensive and not suitable for small labs. Supervised contrastive learning (SCL) has shown its effectiveness with small, available resources. In this work, we examine the effectiveness of fine-tuning transformer-based language models, regularized with a SCL loss for English microblog classification. Despite its simplicity, the evaluation on two English microblog classification benchmarks (TweetEval and Tweet Topic Classification) shows an improvement over baseline models. The result shows that, across all subtasks, our proposed method has a performance gain of up to 11.9 percentage points. All our models are open source.

Machine Translation

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

ICL: Iterative Continual Learning for Multi-domain Neural Machine Translation

Zhibo Man, Kaiyu Huang, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, Jinan Xu

In a practical scenario, multi-domain neural machine translation (MDNMT) aims to continuously acquire knowledge from new domain data while retaining old knowledge. Previous work separately learns each new domain knowledge based on parameter isolation methods, which effectively capture the new knowledge. However, task-specific parameters lead to isolation between models, which hinders the mutual transfer of knowledge between new domains. Given the scarcity of domain-specific corpora, we consider making full use of the data from multiple new domains. Therefore, our work aims to leverage previously acquired domain knowledge when modeling subsequent domains. To this end, we propose an Iterative Continual Learning (ICL) framework for multi-domain neural machine translation. Specifically, when each new domain arrives, (1) we first build a pluggable incremental learning model, (2) then we design an iterative updating algorithm to continuously update the original model, which can be used flexibly for constructing subsequent domain models. Furthermore, we design a domain knowledge transfer mechanism to enhance the fine-grained domain-specific representation, thereby solving the word ambiguity caused by mixing domain data. Experimental results on the UM-Corpus and OPUS multi-domain datasets show the superior performance of our proposed model compared to representative baselines.

(Nov 12): 17:45:18:45 (Evening) - Gather

DeMPT: Decoding-enhanced Multi-phase Prompt Tuning for Making LLMs Be Better Context-aware Translators

Xinglin Lyu, Junhui Li, Yaqing Zhao, Min Zhang, Daimeng Wei, shimin tao, Hao Yang, Min Zhang

Generally, the decoder — only large language models (LLMs) are adapted to context-aware neural machine translation (NMT) in a concatenating way, where LLMs take the concatenation of the source sentence (i.e., intra-sentence context) and the inter-sentence context as the input, and then to generate the target tokens sequentially. This adaptation strategy, i.e., concatenation mode, considers intra-sentence and inter-sentence contexts with the same priority, despite an apparent difference between the two kinds of contexts. In this paper, we propose an alternative adaptation approach, named Decoding-enhanced Multi-phase Prompt Tuning (DeMPT), to make LLMs discriminately model and utilize the inter- and intra-sentence context and more effectively adapt LLMs to context-aware NMT. First, DeMPT divides the context-aware NMT process into three separate phases. During each phase, different continuous prompts are introduced to make LLMs discriminately model various information. Second, DeMPT employs a heuristic way to further discriminately enhance the utilization of the source-side inter- and intra-sentence information at the final decoding phase. Experiments show that our approach significantly outperforms the concatenation method, and further improves the performance of LLMs in discourse modeling.

Multilinguality and Language Diversity

(Nov 12): 17:45:18:45 (Evening) - Room: Gather

(Nov 12): 17:45:18:45 (Evening) - Gather

Can we teach language models to gloss endangered languages?

Michael Gim, Mans Hulden, Alexis Palmer

Interlinear glossed text (IGT) is a popular format in language documentation projects, where each morpheme is labeled with a descriptive annotation. Automating the creation of interlinear glossed text would be desirable to reduce annotator effort and maintain consistency across annotated corpora. Prior research has explored a number of statistical and neural methods for automatically producing IGT. As large language models (LLMs) have shown promising results across multilingual tasks, even for rare, endangered languages, it is natural to wonder whether they can be utilized for the task of generating IGT. We explore whether LLMs can be effective at the task of interlinear glossing with in-context learning, without any traditional training. We propose new approaches for selecting examples to provide in-context, observing that targeted selection can significantly improve performance. We find that LLM-based methods beat standard transformer baselines, despite requiring no training at all. These approaches still underperform state-of-the-art supervised systems for the task, but are highly practical for researchers outside of the NLP community, requiring minimal effort to use.

Multimodality and Language Grounding to Vision, Robotics and Beyond

(Nov 12): 17:45:18:45 (Evening) - Room: Gather

(Nov 12): 17:45:18:45 (Evening) - Gather

MMNeuron: Discovering Neuron-Level Domain-Specific Interpretation in Multimodal Large Language Model

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, Xuming Hu

Projecting visual features into word embedding space has become a significant fusion strategy adopted by Multimodal Large Language Models (MLLMs). However, its internal mechanisms have yet to be explored. Inspired by multilingual research, we identify domain-specific neurons in multimodal large language models. Specifically, we investigate the distribution of domain-specific neurons and the mechanism of how MLLMs process features from diverse domains. Furthermore, we propose a three-stage framework for language model modules in MLLMs when handling projected image features, and verify this hypothesis using logit lens. Extensive experiments indicate that while current MLLMs exhibit Visual Question Answering (VQA) capability, they may not fully utilize domain-specific information. Manipulating domain-specific neurons properly will result in a 10% change of accuracy at most, shedding light on the development of cross-domain, all-encompassing MLLMs in the future. The source code is available at <https://anonymous.4open.science/r/MMNeuron>.

(Nov 12): 17:45:18:45 (Evening) - Gather

Interpretable Composition Attribution Enhancement for Visio-linguistic Compositional Understanding

Wei Li, Zhen Huang, Ximnei Tian, Le Lu, Houqiang Li, Xu Shen, Sieping Ye

Contrastively trained vision-language models such as CLIP have achieved remarkable progress in vision and language representation learning. Despite the promising progress, their proficiency in compositional reasoning over attributes and relations (e.g., distinguishing between "the car is underneath the person" and "the person is underneath the car") remains notably inadequate. We investigate the cause for this deficient behavior is the composition attribution issue, where the attribution scores (e.g., attention scores or GradCAM scores) for relations (e.g., underneath) or attributes (e.g., red) in the text are substantially lower than those for object terms. In this work, we show such issue is mitigated via a novel framework called CAE (Composition Attribution Enhancement). This generic framework incorporates various interpretable attribution methods to encourage the model to pay greater attention to composition words denoting relationships and attributes within the text. Detailed analysis shows that our approach enables the models to adjust and rectify the attribution of the texts. Extensive experiments across seven benchmarks reveal that our framework significantly enhances the ability to discern intricate details and construct more sophisticated interpretations of combined visual and linguistic elements.

(Nov 12): 17:4518:45 (Evening) - Gather

M3D: MultiModal MultiDocument Fine-Grained Inconsistency Detection

Chia-Wei Tang, Ting-Chih Chen, Alvi Md Ishman, Kiet A. Nguyen, Kazi Sajeed Mehrab, Chris Thomas

Fact-checking claims is a highly laborious task that involves understanding how each factual assertion within the claim relates to a set of trusted source materials. Existing approaches make sample-level predictions but fail to identify the specific aspects of the claim that are troublesome and the specific evidence relied upon. In this paper, we introduce a method and new benchmark for this challenging task. Our method predicts the fine-grained logical relationship of each aspect of the claim from a set of multimodal documents, which include text, image(s), video(s), and audio(s). We also introduce a new benchmark (M3DC) of claims requiring multimodal multidocument reasoning, which we construct using a novel claim synthesis technique. Experiments show that our approach outperforms other models on this challenging task on two benchmarks while providing finer-grained predictions, explanations, and evidence.

(Nov 12): 17:4518:45 (Evening) - Gather

Towards One-to-Many Visual Question Answering

Huishan Ji, Qingyi Si, Zheng Lin, Yanan Cao, Weiping Wang

Most existing Visual Question Answering (VQA) systems are constrained to support domain-specific questions, i.e., to train different models separately for different VQA tasks, thus generalizing poorly to others. For example, models trained on the reasoning-focused dataset GQA struggle to effectively handle samples from the knowledge-emphasizing dataset OKVQA. Meanwhile, in real-world scenarios, it is user-unfriendly to restrict the domain of questions. Therefore, this paper proposes a necessary task: One-to-Many Visual Question Answering, of which the ultimate goal is to enable a single model to answer as many different domains of questions as possible by the effective integration of available VQA resources. To this end, we first investigate into ten common VQA datasets, and break the task of VQA down into the integration of three key abilities. Then, considering assorted questions rely on different VQA abilities, this paper proposes a novel dynamic Mixture of LoRAs (MoL) strategy. MoL mixes three individually trained LoRA adapters (corresponding to each VQA ability) dynamically for different samples demanding various VQA abilities. The proposed MoL strategy is verified to be highly effective by experiments, establishing SOTAs on four datasets. In addition, MoL generalizes well to three extra zero-shot datasets. Data and codes will be released.

(Nov 12): 17:4518:45 (Evening) - Gather

OpenSep: Leveraging Large Language Models with Textual Inversion for Open World Audio Separation

Tanvir Mahmud, Diana Marculescu

Audio separation in real-world scenarios, where mixtures contain a variable number of sources, presents significant challenges due to limitations of existing models, such as over-separation, under-separation, and dependence on predefined training sources. We propose OpenSep, a novel framework that leverages large language models (LLMs) for automated audio separation, eliminating the need for manual intervention and overcoming source limitations. OpenSep uses textual inversion to generate captions from audio mixtures with off-the-shelf audio captioning models, effectively parsing the sound sources present. It then employs few-shot LLM prompting to extract detailed audio properties of each parsed source, facilitating separation in unseen mixtures. Additionally, we introduce a multi-level extension of the mix-and-separate training framework to enhance modality alignment by separating single source sounds and mixtures simultaneously. Extensive experiments demonstrate OpenSep's superiority in precisely separating new, unseen, and variable sources in challenging mixtures, outperforming SOTA baseline methods. Code is released at <https://github.com/tanvir-utexas/OpenSep.git>.

(Nov 12): 17:4518:45 (Evening) - Gather

Self-Training Large Language and Vision Assistant for Medical

GuoHao Sun, Can Qin, Huazhu Fu, Linwei Wang, ZHIQIANG TAO

Large Vision-Language Models (LVLMs) have shown significant potential in assisting medical diagnosis by leveraging extensive biomedical datasets. However, the advancement of medical image understanding and reasoning critically depends on building high-quality visual instruction data, which is costly and labor-intensive to obtain, particularly in the medical domain. To mitigate this data-starving issue, we introduce Self-Training Large Language and Vision Assistant for Medical (STLLaVA-Med). The proposed method is designed to train a policy model (an LVLM) capable of auto-generating medical visual instruction data to improve data efficiency, guided through Direct Preference Optimization (DPO). Specifically, a more powerful and larger LVLM (e.g., GPT-4o) is involved as a biomedical expert to oversee the DPO fine-tuning process on the auto-generated data, encouraging the policy model to align efficiently with human preferences. We validate the efficacy and data efficiency of STLLaVA-Med across three major medical Visual Question Answering (VQA) benchmarks, demonstrating competitive zero-shot performance with the utilization of only 9% of the medical data.

NLP Applications 1

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

What Are the Odds? Language Models Are Capable of Probabilistic Reasoning

Akshay Paruchuri, Jake Garrison, shua fiao, John B Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, Daniel McDuff

Language models (LM) are capable of remarkably complex linguistic tasks; however, numerical reasoning is an area in which they frequently struggle. An important but rarely evaluated form of reasoning is understanding probability distributions. In this paper, we focus on evaluating the probabilistic reasoning capabilities of LMs using idealized and real-world statistical distributions. We perform a systematic evaluation of state-of-the-art LMs on three tasks: estimating percentiles, drawing samples, and calculating probabilities. We evaluate three ways to provide context to LMs 1) anchoring examples from within a distribution or family of distributions, 2) real-world context, 3) summary statistics on which to base a Normal approximation. Models can make inferences about distributions, and can be further aided by the incorporation of real-world context, example shots and simplified assumptions, even if these assumptions are incorrect or misspecified. To conduct this work, we developed a comprehensive benchmark distribution dataset with associated question-answer pairs that we have released publicly.

(Nov 12): 17:4518:45 (Evening) - Gather

Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning

Qizhou Chen, TaoLin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, Hui Xue¹

Model editing aims to correct outdated or erroneous knowledge in large language models (LLMs) without the need for costly retraining. Lifelong model editing is the most challenging task that caters to the continuous editing requirements of LLMs. Prior works primarily focus on single or batch editing; nevertheless, these methods fall short in lifelong editing scenarios due to catastrophic knowledge forgetting and the degradation of model performance. Although retrieval-based methods alleviate these issues, they are impeded by slow and cumbersome processes of integrating the retrieved knowledge into the model. In this work, we introduce RECIPE, a RetriEval-augmented Continuous

Prompt IEarning method, to boost editing efficacy and inference efficiency in lifelong learning. RECIPE first converts knowledge statements into short and informative continuous prompts, prefixed to the LLM's input query embedding, to efficiently refine the response grounded on the knowledge. It further integrates the Knowledge Sentinel (KS) that acts as an intermediary to calculate a dynamic threshold, determining whether the retrieval repository contains relevant knowledge. Our retriever and prompt encoder are jointly trained to achieve editing properties, i.e., reliability, generality, and locality. In our experiments, RECIPE is assessed extensively across multiple LLMs and editing datasets, where it achieves superior editing performance. RECIPE also demonstrates its capability to maintain the overall performance of LLMs alongside showcasing fast editing and inference speed.

(Nov 12): 17:45:18:45 (Evening) - Gather

Bi-DCSpell: A Bi-directional Detector-Corrector Interactive Framework for Chinese Spelling Check

Haining Wu, Hanqing Zhang, Richeng xuan, Dawei Song

Chinese Spelling Check (CSC) aims to detect and correct potentially misspelled characters in Chinese sentences. Naturally, it involves the detection and correction subtasks, which interact with each other dynamically. Such interactions are bi-directional, i.e., the detection result would help reduce the risk of over-correction and under-correction while the knowledge learnt from correction would help prevent false detection. Current CSC approaches are of two types: correction-only or single-directional detection-to-correction interactive frameworks. Nonetheless, they overlook the bi-directional interactions between detection and correction. This paper aims to fill the gap by proposing a Bi-directional-Detector-Corrector framework for CSC (Bi-DCSpell). Notably, Bi-DCSpell contains separate detection and correction encoders, followed by a novel interactive learning module facilitating bi-directional feature interactions between detection and correction to improve each other's representation learning. Extensive experimental results demonstrate a robust correction performance of Bi-DCSpell on widely used benchmarking datasets while possessing a satisfactory detection ability.

(Nov 12): 17:45:18:45 (Evening) - Gather

RealVul: Can We Detect Vulnerabilities in Web Applications with LLM?

Di Cao, Yong Liu, Xiwei Shang

The latest advancements in large language models (LLMs) have sparked interest in their potential for software vulnerability detection. However, there is currently a lack of research specifically focused on vulnerabilities in the PHP language, and challenges in data sampling and processing persist, hindering the model's ability to effectively capture the characteristics of specific vulnerabilities. In this paper, we present RealVul, the first LLM-based framework designed for PHP vulnerability detection, addressing these issues. By improving code sampling methods and employing normalization techniques, we can isolate potential vulnerability triggers while streamlining the code and eliminating unnecessary semantic information, enabling the model to better understand and learn from the generated vulnerability samples. We also address the issue of insufficient PHP vulnerability samples by improving data synthesis methods. To evaluate RealVul's performance, we conduct an extensive analysis using five distinct code LLMs on vulnerability data from 180 PHP projects. The results demonstrate a significant improvement in both effectiveness and generalization compared to existing methods, effectively boosting the vulnerability detection capabilities of these models.

(Nov 12): 17:45:18:45 (Evening) - Gather

EXPLORA: Efficient Exemplar Subset Selection for Complex Reasoning

Kiran Purohit, Venkatesh V. Raghuvaran Devalla, Krishna Mohan Yerragorla, Sourangshu Bhattacharya, Avishhek Anand

Answering reasoning-based complex questions over text and hybrid sources, including tables, is a challenging task. Recent advances in large language models (LLMs) have enabled in-context learning (ICL), allowing LLMs to acquire proficiency in a specific task using only a few demonstration samples (exemplars). A critical challenge in ICL is the selection of optimal exemplars, which can be either task-specific (static) or test-example-specific (dynamic). Static exemplars provide faster inference times and increased robustness across a distribution of test examples. In this paper, we propose an algorithm for static exemplar subset selection for complex reasoning tasks. We introduce EXPLORA, a novel exploration method designed to estimate the parameters of the scoring function, which evaluates exemplar subsets without incorporating confidence information. EXPLORA significantly reduces the number of LLM calls to 11% of those required by state-of-the-art methods and achieves a substantial performance improvement of 12.24%. We open-source our code and data (<https://github.com/kiranpurohit/EXPLORA>).

Question Answering

(Nov 12): 17:45:18:45 (Evening) - Room: Gather

(Nov 12): 17:45:18:45 (Evening) - Gather

Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering

Chang Zong, Yuchen Yan, Weinming Lu, Jian Shao, Yongfeng Huang, Heng Chang, Yueting Zhuang

Recent progress with LLM-based agents has shown promising results across various tasks. However, their use in answering questions from knowledge bases remains largely unexplored. Implementing a KBQA system using traditional methods is challenging due to the shortage of task-specific training data and the complexity of creating task-focused model structures. In this paper, we present Triad, a unified framework that utilizes an LLM-based agent with multiple roles for KBQA tasks. The agent is assigned three roles to tackle different KBQA subtasks: agent as a generalist for mastering various subtasks, as a decision maker for the selection of candidates, and as an advisor for answering questions with knowledge. Our KBQA framework is executed in four phases, involving the collaboration of the agent's multiple roles. We evaluated the performance of our framework using three benchmark datasets, and the results show that our framework outperforms state-of-the-art systems on the LC-QuAD and YAGO-QA benchmarks, yielding F1 scores of 11.8% and 20.7%, respectively.

(Nov 12): 17:45:18:45 (Evening) - Gather

Adaption-of-Thought: Learning Question Difficulty Improves Large Language Models for Reasoning

Mai Xu, Yongqi Li, Ke Sun, Tieyun Qian

Large language models (LLMs) have shown excellent capability for solving reasoning problems. Existing approaches do not differentiate the question difficulty when designing prompting methods for them. Clearly, a simple method cannot elicit sufficient knowledge from LLMs to answer a hard question. Meanwhile, a sophisticated one will force the LLM to generate redundant or even inaccurate intermediate steps toward a simple question. Consequently, the performance of existing methods fluctuates among various questions. In this work, we propose Adaption-of-Thought (AdoT), an adaptive method to improve LLMs for the reasoning problem, which first measures the question difficulty and then tailors demonstration set construction and difficulty-adapted retrieval strategies for the adaptive demonstration construction. Experimental results on three reasoning tasks prove the superiority of our proposed method, showing an absolute improvement of up to 5.5% on arithmetic reasoning, 7.4% on symbolic reasoning, and 2.3% on commonsense reasoning. Our codes and implementation details are available at: <https://github.com/NLPGM/AdoT>

(Nov 12): 17:4518:45 (Evening) - Gather

Question-guided Knowledge Graph Re-scoring and Injection for Knowledge Graph Question Answering**Yu Zhang, Kehai Chen, Xuefeng Bai, zhao kang, Quanjiang Guo, Min Zhang**

Knowledge graph question answering (KGQA) involves answering natural language questions by leveraging structured information stored in a knowledge graph. Typically, KGQA initially retrieve a targeted subgraph from a large-scale knowledge graph, which serves as the basis for reasoning models to address queries. However, the retrieved subgraph inevitably brings distraction information for knowledge utilization, impeding the model's ability to perform accurate reasoning. To address this issue, we propose a Question-guided Knowledge Graph Re-scoring method (Q-KGR) to eliminate noisy pathways for the input question, thereby focusing specifically on pertinent factual knowledge. Moreover, we introduce Knowformer, a parameter-efficient method for injecting the re-scored knowledge graph into large language models to enhance their ability to perform factual reasoning. Extensive experiments on multiple KGQA benchmarks demonstrate the superiority of our method over existing systems.

(Nov 12): 17:4518:45 (Evening) - Gather

Advancing Process Verification for Large Language Models via Tree-Based Preference Learning**Mingqian He, Yonghang Shen, Wenqi Zhang, Zegi Tan, Weiming Lu**

Large Language Models (LLMs) have demonstrated remarkable potential in handling complex reasoning tasks by generating step-by-step rationales. Some methods have proven effective in boosting accuracy by introducing extra verifiers to assess these paths. However, existing verifiers, typically trained on binary-labeled reasoning paths, fail to fully utilize the relative merits of intermediate steps, thereby limiting the effectiveness of the feedback provided. To overcome this limitation, we propose Tree-based Preference Learning Verifier (Tree-PLV), a novel approach that constructs reasoning trees via a best-first search algorithm and collects step-level paired data for preference training. Compared to traditional binary classification, step-level preferences more finely capture the nuances between reasoning steps, allowing for a more precise evaluation of the complete reasoning path. We empirically evaluate Tree-PLV across a range of arithmetic and commonsense reasoning tasks, where it significantly outperforms existing benchmarks. For instance, Tree-PLV achieved substantial performance gains over the Mistral-7B self-consistency baseline on GSM8K (67.55% vs. 82.79%), MATH (17.00% vs. 26.80%), CSQA (68.14% vs. 72.97%), and StrategyQA (82.86% vs. 83.25%). Additionally, our study explores the appropriate granularity for applying preference learning, revealing that step-level guidance provides feedback that better aligns with the evaluation of the reasoning process.

(Nov 12): 17:4518:45 (Evening) - Gather

Position Paper: Creative Problem Solving in Large Language and Vision Models – What Would it Take?**Lakshmi Nair, Evana Gizzì, Jívko Sinapov**

We advocate for a strong integration of Computational Creativity (CC) with research in large language and vision models (LLVMs) to address a key limitation of these models, i.e., creative problem solving. We present preliminary experiments showing how CC principles can be applied to address this limitation. Our goal is to foster discussions on creative problem solving in LLVMs and CC at prestigious ML venues.

Resources and Evaluation

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Towards Probing Speech-Specific Risks in Large Multimodal Models: A Taxonomy, Benchmark, and Insights**Hao Yang, Lizhen Qu, Ehsan Shareghi, Reza Haf**

Large Multimodal Models (LMMs) have achieved great success recently, demonstrating a strong capability to understand multimodal information and to interact with human users. Despite the progress made, the challenge of detecting high-risk interactions in multimodal settings, and in particular in speech modality, remains largely unexplored. Conventional research on risk for speech modality primarily emphasises the content (e.g., what is captured as transcription). However, in speech-based interactions, paralinguistic cues in audio can significantly alter the intended meaning behind utterances. In this work, we propose a speech-specific risk taxonomy, covering 8 risk categories under hostility (malicious sarcasm and threats), malicious imitation (age, gender, ethnicity), and stereotypical biases (age, gender, ethnicity). Based on the taxonomy, we create a small-scale dataset for evaluating current LMMs capability in detecting these categories of risk. We observe even the latest models remain ineffective to detect various paralinguistic-specific risks in speech (e.g., Gemini 1.5 Pro is performing only slightly above random baseline). Warning: this paper contains biased and offensive examples.

(Nov 12): 17:4518:45 (Evening) - Gather

ABSEval: An Agent-based Framework for Script Evaluation**Sirui Liang, Baoli Zhang, Jun Zhao, Kang Liu**

Recent research indicates that large language models (LLMs) possess a certain degree of script planning capability. However, there is still a lack of focused work on evaluating scripts generated by LLMs. The evaluation of scripts poses challenges due to their logical structure, sequential organization, adherence to commonsense constraints, and open-endedness. In this work, We introduced a novel script evaluation dataset, MICS-Script, consisting of more than 1,500 script evaluation tasks and steps, and developed an agent-based script evaluation framework, ABSEval, to collaboratively evaluate scripts generated by LLMs. Our experiments demonstrate that ABSEval provides superior accuracy and relevance, aligning closely with human evaluation. We evaluated the script planning capabilities of 15 mainstream LLMs and provided a detailed analysis. Furthermore, we observed phenomena like the key factor influencing the script planning ability of LLM is not parameter size and suggested improvements for evaluating open-ended questions.

(Nov 12): 17:4518:45 (Evening) - Gather

AfrILInstruct: Instruction Tuning of African Languages for Diverse Tasks**Kosei Uemura, Alex Pejovic, Mahe Chen, Chika Maduabuchi, Yifei Sun, En-Shiu Annie Lee**

Large language models (LLMs) for African languages perform worse compared to their performance in high-resource languages. To address this issue, we introduce AfrILInstruct, which specializes in instruction-tuning of multiple African languages covering various tasks. We trained the LLaMa-2-7B using continual pretraining and instruction fine-tuning, which demonstrates superior performance across multiple tasks. Our mixed task evaluation shows that our model outperforms GPT-3.5-Turbo and other baseline models of similar size. Our contributions fill a critical gap of LLM performance between high-resource and African languages.

(Nov 12): 17:4518:45 (Evening) - Gather

EU DisInfoTest: A Benchmark for Evaluating Language Models' Ability to Detect Disinformation Narratives **Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Jahna Otterbacher, Adam Wierzbicki**

As narratives shape public opinion and influence societal actions, distinguishing between truthful and misleading narratives has become a

significant challenge. To address this, we introduce the EU DisinfoTest, a novel benchmark designed to evaluate the efficacy of Language Models in identifying disinformation narratives. Developed through a Human-in-the-Loop methodology and grounded from EU DisinfoLab, the EU DisinfoTest comprises more than 1,300 narratives. Our benchmark includes persuasive elements under Logos, Pathos, and Ethos rhetorical dimensions. We assessed state-of-the-art LLMs, including the newly released GPT-4o, on their capability to perform zero-shot classification of disinformation narratives versus credible narratives. Our findings reveal that LLMs tend to regard narratives with authoritative appeals as trustworthy, while those with emotional appeals are frequently incorrectly classified as disinformative. These findings highlight the challenges LLMs face in nuanced content interpretation and suggest the need for tailored adjustments in LLM training to better handle diverse narrative structures.

(Nov 12): 17:45:18:45 (Evening) - Gather

Walla-LLM: Enhancing Amharic-LLaMA by Integrating Task-Specific and Generative Datasets

Israel Abebe Azime, Atnafu Lambobo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Youas Chanie, Waleign Tewabe Sewinetie, Seid Muhie Yimam

Large language models (LLMs) have received a lot of attention in natural language processing (NLP) research because of their exceptional performance in understanding and generating human languages. However, low-resource languages are left behind due to the unavailability of resources. In this work, we focus on enhancing the LLaMA-2-Amharic model by integrating task-specific and generative datasets to improve language model performance for Amharic. We compile an Amharic instruction fine-tuning dataset and fine-tuned LLaMA-2-Amharic model. The fine-tuned model shows promising results in different NLP tasks. We also explore the effectiveness of translated instruction datasets compared to the dataset we created. Our dataset creation pipeline, along with instruction datasets, trained models, and evaluation outputs, is made publicly available to encourage research in language-specific models.

(Nov 12): 17:45:18:45 (Evening) - Gather

Understanding Faithfulness and Reasoning of Large Language Models on Plain Biomedical Summaries

Biaoan Fang, Xiang Dai, Sarvnaz Karimi

Generating plain biomedical summaries with Large Language Models (LLMs) can enhance the accessibility of biomedical knowledge to the public. However, how faithful the generated summaries are remains an open yet critical question. To address this, we propose FaReBio, a benchmark dataset with expert-annotated Faithfulness and Reasoning on plain Biomedical Summaries. This dataset consists of 175 plain summaries (445 sentences) generated by seven different LLMs, paired with source articles. Using our dataset, we identify the performance gap of LLMs in generating faithful plain biomedical summaries and observe a negative correlation between abstractiveness and faithfulness. We also show that current faithfulness evaluation metrics do not work well in the biomedical domain and confirm the over-confident tendency of LLMs as faithfulness evaluators. To better understand the faithfulness judgements, we further benchmark LLMs in retrieving supporting evidence and show the gap of LLMs in reasoning faithfulness evaluation at different abstractiveness levels. Going beyond the binary faithfulness labels, coupled with the annotation of supporting sentences, our dataset could further contribute to the understanding of faithfulness evaluation and reasoning.

(Nov 12): 17:45:18:45 (Evening) - Gather

Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments

Han Zhou, Xingchen Wan, Yinghong Liu, Nigel Collier, Ivan Vuli, Anna Korhonen

Large language models (LLMs) have shown promising abilities as cost-effective and reference-free evaluators for assessing language generation quality. In particular, pairwise LLM evaluators, which compare two generated texts and determine the preferred one, have been employed in a wide range of applications. However, LLMs exhibit preference biases and worrying sensitivity to prompt designs. In this work, we first reveal that the predictive preference of LLMs can be highly brittle and skewed, even with semantically equivalent instructions. We find that fairer predictive preferences from LLMs consistently lead to judgments that are better aligned with humans. Motivated by this phenomenon, we propose an automatic Zero-shot Evaluation-oriented Prompt Optimization framework, ZEPO, which aims to produce fairer preference decisions and improve the alignment of LLM evaluators with human judgments. To this end, we propose a zero-shot learning objective based on the preference decision fairness. ZEPO demonstrates substantial performance improvements over state-of-the-art LLM evaluators, without requiring labeled data, on representative meta-evaluation benchmarks. Our findings underscore the critical correlation between preference fairness and human alignment, positioning ZEPO as an efficient prompt optimizer for bridging the gap between LLM evaluators and human judgments.

(Nov 12): 17:45:18:45 (Evening) - Gather

ESC-Eval: Evaluating Emotion Support Conversations in Large Language Models

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jian Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Tang, Yanghua Xu, Yingchun Wang

Emotion Support Conversation (ESC) is a crucial application, which aims to reduce human stress, offer emotional guidance, and ultimately enhance human mental and physical well-being. With the advancement of Large Language Models (LLMs), many researchers have employed LLMs as the ESC models. However, the evaluation of these LLM-based ESCs remains uncertain. In detail, we first re-organize 2,801 role-playing cards from seven existing datasets to define the roles of the role-playing agent. Second, we train a specific role-playing model called ESC-Role which behaves more like a confused person than GPT-4. Third, through ESC-Role and organized role cards, we systematically conduct experiments using 14 LLMs as the ESC models, including general AI-assistant LLMs (e.g., ChatGPT) and ESC-oriented LLMs (e.g., EXTES-Llama). We conduct comprehensive human annotations on interactive multi-turn dialogues of different ESC models. The results show that ESC-oriented LLMs exhibit superior ESC abilities compared to general AI-assistant LLMs, but there is still a gap behind human performance. Moreover, to automate the scoring process for future ESC models, we developed ESC-RANK, which trained on the annotated data, achieving a scoring performance surpassing 35 points of GPT-4.

Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas

(Nov 12): 17:45:18:45 (Evening) - Room: Gather

(Nov 12): 17:45:18:45 (Evening) - Gather

Automatically Generated Definitions and their utility for Modeling Word Meaning

Francesco Periti, David Alfter, Nina Tahmasebi

Modeling lexical semantics is a challenging task, often suffering from interpretability pitfalls. In this paper, we delve into the generation of dictionary-like sense definitions and explore their utility for modeling word meaning. We fine-tuned two Llama models and include an existing T5-based model in our evaluation. Firstly, we evaluate the quality of the generated definitions on existing English benchmarks, setting new state-of-the-art results for the Definition Generation task. Next, we explore the use of definitions generated by our models as

intermediate representations subsequently encoded as sentence embeddings. We evaluate this approach on lexical semantics tasks such as the Word-in-Context, Word Sense Induction, and Lexical Semantic Change, setting new state-of-the-art results in all three tasks when compared to unsupervised baselines.

Sentiment Analysis, Stylistic Analysis, and Argument Mining

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

PclGPT: A Large Language Model for Patronizing and Condescending Language Detection

Hongbo Wang, LiMingDa, Junyu Lu, Hebin Xia, Liang Yang, Bo Xu, RuiZhu Liu, Hongfei Lin

Disclaimer: Samples in this paper may be harmful and cause discomfort! Patronizing and condescending language (PCL) is a form of speech directed at vulnerable groups. As an essential branch of toxic language, this type of language exacerbates conflicts and confrontations among Internet communities and detrimentally impacts disadvantaged groups. Traditional pre-trained language models (PLMs) perform poorly in detecting PCL due to its implicit toxicity traits like hypocrisy and false sympathy. With the rise of large language models (LLMs), we can harness their rich emotional semantics to establish a paradigm for exploring implicit toxicity. In this paper, we introduce PclGPT, a comprehensive LLM benchmark designed specifically for PCL. We collect, annotate, and integrate the Pcl-PT/SFT dataset, and then develop a bilingual PclGPT-EN/CN model group through a comprehensive pre-training and supervised fine-tuning staircase process to facilitate implicit toxic detection. Group detection results and fine-grained detection from PclGPT and other models reveal significant variations in the degree of bias in PCL towards different vulnerable groups, necessitating increased societal attention to protect them.

(Nov 12): 17:4518:45 (Evening) - Gather

An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs

Manuj Malik, Jing Jiang, Kian Ming A. Chai

There are recent efforts to personalize large language models (LLMs) by assigning them specific personas. This paper explores the writing styles of such persona-assigned LLMs across different socio-demographic groups based on age, profession, location, and political affiliations, using three widely-used LLMs. Leveraging an existing style embedding model that produces detailed style attributes and latent Dirichlet allocation (LDA) for broad style analysis, we measure style differences using Kullback-Leibler divergence to compare LLM-generated and human-written texts. We find significant style differences among personas. This analysis emphasizes the need to consider socio-demographic factors in language modeling to accurately capture diverse writing styles used for communications. The findings also reveal the strengths and limitations of personalized LLMs, their potential uses, and the importance of addressing biases in their design.

(Nov 12): 17:4518:45 (Evening) - Gather

DetectiveNN: Imitating Human Emotional Reasoning with a Recall-Detect-Predict Framework for Emotion Recognition in Conversations

Simin Hong, Jun Sun, Taihao Li

Emotion Recognition in conversations (ERC) involves an internal cognitive process that interprets emotional cues by using a collection of past emotional experiences. However, many existing methods struggle to decipher emotional cues in dialogues since they are insufficient in understanding the rich historical emotional context. In this work, we introduce an innovative Detective Network (DetectiveNN), a novel model that is grounded in the cognitive theory of emotion and utilizes a "recall-detect-predict" framework to imitate human emotional reasoning. This process begins by 'recalling' past interactions of a specific speaker to collect emotional cues. It then 'detects' relevant emotional patterns by interpreting these cues in the context of the ongoing conversation. Finally, it 'predicts' the speaker's current emotional state. Tested on three benchmark datasets, our approach significantly outperforms existing methods. This highlights the advantages of incorporating cognitive factors into deep learning for ERC, enhancing task efficacy and prediction accuracy.

Special Theme: Efficiency in Model Algorithms, Training, and Inference

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Predicting Rewards Alongside Tokens: Non-disruptive Parameter Insertion for Efficient Inference Intervention in Large Language Model

Chenhan Yuan, Fei Huang, Ru Peng, Keming Lu, Bowen Yu, Chang Zhou, Jingren Zhou

Transformer-based large language models (LLMs) exhibit limitations such as generating unsafe responses, unreliable reasoning, etc. Existing inference intervention approaches attempt to mitigate these issues by finetuning additional models to produce calibration signals (such as rewards) that guide the LLM's decoding process. However, this solution introduces substantial time and space overhead due to the separate models required. This work proposes Non-disruptive parameters insertion (Otter), inserting extra parameters into the transformer architecture to predict calibration signals along with the original LLM output. Otter offers state-of-the-art performance on multiple demanding tasks while saving up to 86.5% extra space and 98.5% extra time. Furthermore, Otter seamlessly integrates with existing inference engines, requiring only a one-line code change, and the original model response remains accessible after the parameter insertion.

(Nov 12): 17:4518:45 (Evening) - Gather

Nash CoT: Multi-Path Inference with Preference Equilibrium

Ziqi Zhang, Cuxiang Wang, Xiao Xiong, Yue Zhang, Donglin Wang

Chain of thought (CoT) is a reasoning framework that can enhance the performance of large language models (LLMs) on complex inference tasks. In particular, among various studies related to CoT, multi-path inference stands out as a simple yet effective improvement. However, there is no optimal setting for the number of inference paths. Therefore, we have to increase the number of inference paths to obtain better results, which in turn increases the inference cost. To address this limitation, we can utilize question-related role templates to guide LLMs into relevant roles, thereby increasing the possibility of correct inferences for each path and further reducing dependence on the number of inference paths while improving reasoning accuracy. However, placing LLMs into specific roles may reduce their reasoning diversity and performance on a few tasks where role dependence is low. To alleviate the excessive immersion of the LLM into a specific role, we propose Nash CoT by constructing a competitive system on each path that balances the generation from role-specific LLMs^{*} and the general LLMs' generation, thereby ensuring both effective role adoption and diversity in LLM generation further maintaining the performance of multi-path

inference while reducing the requirement of the number of inference paths. We evaluate Nash CoT across various inference tasks, including Arabic Reasoning, Commonsense Question Answering, and Symbolic Inference, achieving results that are comparable to or better than those of multi-path CoT with the equal number of inference paths.

Speech Processing and Spoken Language Understanding

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Exploring the Potential of Multimodal LLM with Knowledge-Intensive Multimodal ASR

Minghan Wang, Yuxia Wang, Thuy-Trang Vu, Ehsan Shareghi, Reza Haf

Recent advancements in multimodal large language models (MLLMs) have made significant progress in integrating information across various modalities, yet real-world applications in educational and scientific domains remain challenging. This paper introduces the Multimodal Scientific ASR (MS-ASR) task, which focuses on transcribing scientific conference videos by leveraging visual information from slides to enhance the accuracy of technical terminologies. Realized that traditional metrics like WER fall short in assessing performance accurately, prompting the proposal of severity-aware WER (SWER) that considers the content type and severity of ASR errors. We propose the Scientific Vision Augmented ASR (SciVASR) framework as a baseline method, enabling MLLMs to improve transcript quality through post-editing. Evaluations of state-of-the-art MLLMs, including GPT-4o, show a 45% improvement over speech-only baselines, highlighting the importance of multimodal information integration.

Syntax: Tagging, Chunking and Parsing

(Nov 12): 17:4518:45 (Evening) - Room: Gather

(Nov 12): 17:4518:45 (Evening) - Gather

Representation Alignment and Adversarial Networks for Cross-lingual Dependency Parsing

Ying Li, Jianjian Liu, Zhengtao Yu, Shengxiang Guo, Yuxin Huang, Cunli Mao

With the strong representational capabilities of pre-trained language models, dependency parsing in resource-rich languages has seen significant advancements. However, the parsing accuracy drops sharply when the model is transferred to low-resource language due to distribution shifts. To alleviate this issue, we propose a representation alignment and adversarial model to filter out useful knowledge from rich-resource language and ignore useless ones. Our proposed model consists of two components, i.e., an alignment network in the input layer for selecting useful language-specific features and an adversarial network in the encoder layer for augmenting the language-invariant contextualized features. Experiments on the benchmark datasets show that our proposed model outperforms RoBERTa-enhanced strong baseline models by 1.37 LAS and 1.34 UAS. Detailed analysis shows that both alignment and adversarial networks are equally important in alleviating the distribution shifts problem and can complement each other. In addition, the comparative experiments demonstrate that both the alignment and adversarial networks can substantially facilitate extracting and utilizing relevant target language features, thereby increasing the adaptation capability of our proposed model.

Virtual Poster Session 2 - (Nov 13): 7:458:45 (Morning)

Computational Social Science and Cultural Analytics

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, Hao Wang

This paper explores the open research problem of understanding the social behaviors of LLM-based agents. Using Avalon as a testbed, we employ system prompts to guide LLM agents in gameplay. While previous studies have touched on gameplay with LLM agents, research on their social behaviors is lacking. We propose a novel framework, tailored for Avalon, features a multi-agent system facilitating efficient communication and interaction. We evaluate its performance based on game success and analyze LLM agents' social behaviors. Results affirm the framework's effectiveness in creating adaptive agents and suggest LLM-based agents' potential in navigating dynamic social interactions. By examining collaboration and confrontation behaviors, we offer insights into this field's research and applications.

(Nov 13): 7:458:45 (Morning) - Gather

Multimodal Clickbait Detection by De-confounding Biases Using Causal Representation Inference

Jianxing Yu, Shiqi Wang, Han Yin, Zhenlong Sun, Ruobing Xie, Bo Zhang, Yanghui Rao

This paper focuses on detecting clickbait posts on the Web. These posts often use eye-catching disinformation in mixed modalities to mislead users to click for profit. That affects the user experience and thus would be blocked by content provider. To escape detection, malicious creators use tricks to add some irrelevant non-bait content into bait posts, dressing them up as legal to fool the detector. This content often has biased relations with non-bait labels, yet traditional detectors tend to make predictions based on simple co-occurrence rather than grasping inherent factors that lead to malicious behavior. This spurious bias would easily cause misjudgments. To address this problem, we propose a new debiased method based on causal inference. We first employ a set of features in multiple modalities to characterize the posts. Considering these features are often mixed up with unknown biases, we then disentangle three kinds of latent factors from them, including the invariant factor that indicates intrinsic bait intention; the causal factor which reflects deceptive patterns in a certain scenario, and non-causal noise. By eliminating the noise that causes bias, we can use invariant and causal factors to build a robust model with good generalization ability. Experiments on three popular datasets show the effectiveness of our approach.

(Nov 13): 7:45:45 (Morning) - Gather

M3Hop-CoT: Misogynous Meme Identification with Multimodal Multi-hop Chain-of-Thought**Gitanjali Kumari, Kirtan Jain, Asif Ekbal**

In recent years, there has been a significant rise in the phenomenon of hate against women on social media platforms, particularly through the use of misogynous memes. These memes often target women with subtle and obscure cues, making their detection a challenging task for automated systems. Recently, Large Language Models (LLMs) have shown promising results in reasoning using Chain-of-Thought (CoT) prompting to generate the intermediate reasoning chains as the rationale to facilitate multimodal tasks, but often neglect cultural diversity and key aspects like emotion and contextual knowledge hidden in the visual modalities. To address this gap, we introduce a **M**ultimodal **M**isogynous **M**ulti-hop CoT (M3Hop-CoT) framework for **M**isogynous meme identification, combining a CLIP-based classifier and a multimodal CoT module with entity-object-relationship integration. M3Hop-CoT employs a three-step multimodal prompting principle to induce emotions, target awareness, and contextual knowledge for meme analysis. Our empirical evaluation, including both qualitative and quantitative analysis, validates the efficacy of the M3Hop-CoT framework on the SemEval-2022 Task 5 (**MAMI task**) dataset, highlighting its strong performance in the macro-F1 score. Furthermore, we evaluate the model's generalizability by evaluating it on various benchmark meme datasets, offering a thorough insight into the effectiveness of our approach across different datasets. Codes are available at this link: https://github.com/Gitanjali1801/LLM_CoT

(Nov 13): 7:45:45 (Morning) - Gather

Ukrainian Resilience: A Dataset for Detection of Help-Seeking Signals Amidst the Chaos of War**MSVPJ Sathvik, Abhilash Dowpati, Sreyansh Sethi**

We propose a novel dataset "Ukrainian Resilience" that brings together a collection of social media posts in the Ukrainian language for the detection of help-seeking posts in the Russia-Ukraine war. It is designed to help us analyze and categorize subtle signals in these posts that indicate people are asking for help during times of war. We are using advanced language processing and machine learning techniques to pick up on the nuances of language that show distress or urgency. The dataset is the binary classification of the social media posts that required help and did not require help in the war. The dataset could significantly improve humanitarian efforts, allowing for quicker and more targeted help for those facing the challenges of war. Moreover, the baseline models are implemented and GPT 3.5 achieved an accuracy of 81.15%.

(Nov 13): 7:45:45 (Morning) - Gather

On Fake News Detection with LLM Enhanced Semantics Mining**Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, Hao Fan**

Large language models (LLMs) have emerged as valuable tools for enhancing textual features in various text-related tasks. Despite their superiority in capturing the lexical semantics between tokens for text analysis, our preliminary study on two popular LLMs, i.e., ChatGPT and Llama2, showcases that simply applying the news embeddings from LLMs is ineffective for fake news detection. Such embeddings only encapsulate the language styles between tokens. Meanwhile, the high-level semantics among named entities and topics, which reveal the deviating patterns of fake news, have been ignored. Therefore, we propose a topic model together with a set of specially designed prompts to extract topics and real entities from LLMs and model the relations among news, entities, and topics as a heterogeneous graph to facilitate investigating news semantics. We then propose a Generalized Page-Rank model and a consistent learning criteria for mining the local and global semantics centered on each news piece through the adaptive propagation of features across the graph. Our model shows superior performance on five benchmark datasets over seven baseline methods and the efficacy of the key ingredients has been thoroughly validated.

(Nov 13): 7:45:45 (Morning) - Gather

Message Passing on Semantic-Anchor-Graphs for Fine-grained Emotion Representation Learning and Classification**Pinyi Zhang, Jingyang Chen, Junchen Shen, Zijie Zhai, Ping Li, Jie Zhang, Kai Zhang**

Emotion classification has wide applications in education, robotics, virtual reality, etc. However, identifying subtle differences between fine-grained emotion categories remains challenging. Current methods typically aggregate numerous token embeddings of a sentence into a single vector, which, while being an efficient compressor, may not fully capture complex semantic and temporal distributions. To solve this problem, we propose SEAMantic ANchor Graph Neural Networks (SEAN-GNN) for fine-grained emotion classification. It learns a group of representative, multi-faceted semantic anchors in the token embedding space: using these anchors as a global reference, any sentence can be projected onto them to form a "semantic-anchor graph", with node attributes and edge weights quantifying the semantic and temporal information respectively. The graph structure is well aligned across sentences and, importantly, allows for generating comprehensive emotion representations regarding K different anchors. Message passing on this graph can further integrate and refine the learned features. Empirically, SEAN-GNN can generate meaningful semantic anchors and discriminative graph patterns for different emotion, with promising classification results on 6 popular benchmark datasets against state-of-the-arts.

Demo

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

SparkRA: A Retrieval-Augmented Knowledge Service System Based on Spark Large Language Model**Baoxin Wang, Bo Wang, Dayong Wu, Guoping Hu, Honghong Zhao, Jiaqi Li, Li Qian, Rui Zhang, Shijin Wang, Siyuan Xue, Yanjie Yang, Zhijun Chang, Zhixiong Zhang**

Large language models (LLMs) have shown remarkable achievements across various language tasks. To enhance the performance of LLMs in scientific literature services, we developed the scientific literature LLM (SciLit-LLM) through pre-training and supervised fine-tuning on scientific literature, building upon the iFLYTEK Spark LLM. Furthermore, we present a knowledge service system Spark Research Assistant (SparkRA) based on our SciLit-LLM. SparkRA is accessible online and provides three primary functions: literature investigation, paper reading, and academic writing. As of July 30, 2024, SparkRA has garnered over 50,000 registered users, with a total usage count exceeding 1.3 million.

(Nov 13): 7:45:45 (Morning) - Gather

Generative Dictionary: Improving Language Learner Understanding with Contextual Definitions**Hai-Lun Tu, Jason S. Chang, Kai-Wen Tuan**

We introduce GenerativeDictionary, a novel dictionary system that generates word sense interpretations based on the given context. Our approach involves transforming context sentences to highlight the meaning of target words within their specific context. The method involves automatically transforming context sentences into sequences of low-dimensional vector token representations, automatically processing the input embeddings through multiple layers of transformers, and automatically generate the word senses based on the latent representations

derived from the context. At runtime, context sentences with target words are processed through a transformer model that outputs the relevant word senses. Blind evaluations on a combined set of dictionary example sentences and generated sentences based on given word senses demonstrate that our method is comparable to traditional word sense disambiguation (WSD) methods. By framing WSD as a generative problem, GenerativeDictionary delivers more precise and contextually appropriate word senses, enhancing the effectiveness of language learning tools.

(Nov 13): 7:45:45 (Morning) - Gather

WalledEval: A Comprehensive Safety Evaluation Toolkit for Large Language Models

Dar Win Liew, Hugo Maximus Lim, I-Shiang Lee, Koh Jia Hng, Le Qi Yau, Prannaya Gupta, Rishabh Bharadwaj, Soujanya Poria, Low Hao Han, Rajat Bharadwaj, Teoh Yu Xin

WalledEval is a comprehensive AI safety testing toolkit designed to evaluate large language models (LLMs). It accommodates a diverse range of models, including both open-weight and API-based ones, and features over 35 safety benchmarks covering areas such as multilingual safety, exaggerated safety, and prompt injections. The framework supports both LLM and judge benchmarking, and incorporates custom mutations to test safety against various text-style mutations such as future tense and paraphrasing. Additionally, WalledEval introduces WalledGuard, a new, small, and performant content moderation tool, and SGXSTest, a benchmark for assessing exaggerated safety in cultural contexts. We make WalledEval publicly available at <https://github.com/walledai/walledeval> with a demonstration video at <https://youtu.be/50Zy97kj1MA>.

(Nov 13): 7:45:45 (Morning) - Gather

RAGLAB: A Modular and Research-Oriented Unified Framework for Retrieval-Augmented Generation

Shikun Zhang, Shuyun Tang, Wei Ye, Wenyuan Xu, Xinfeng Li, Xinyu Dai, Xuanwang Zhang, Yidong Wang, Yue Zhang, Yun-Ze Song, Zhen Wu, Qingsong Wen, ZHENGRAN ZENG

Large Language Models (LLMs) demonstrate human-level capabilities in dialogue, reasoning, and knowledge retention. However, even the most advanced LLMs face challenges such as hallucinations and real-time updating of their knowledge. Current research addresses this bottleneck by equipping LLMs with external knowledge, a technique known as Retrieval Augmented Generation (RAG). However, two key issues constrained the development of RAG. First, there is a growing lack of comprehensive and fair comparisons between novel RAG algorithms. Second, open-source tools such as LlamaIndex and LangChain employ high-level abstractions, which results in a lack of transparency and limits the ability to develop novel algorithms and evaluation metrics. To close this gap, we introduce RAGLAB, a modular and research-oriented open-source library. RAGLAB reproduces 6 existing algorithms and provides a comprehensive ecosystem for investigating RAG algorithms. Leveraging RAGLAB, we conduct a fair comparison of 6 RAG algorithms across 10 benchmarks. With RAGLAB, researchers can efficiently compare the performance of various algorithms and develop novel algorithms.

(Nov 13): 7:45:45 (Morning) - Gather

Sailor: Open Language Models for South-East Asia

Guangtao Zeng, Jia Guo, Jiahui Zhou, Longxu Dou, Min Lin, Qian Liu, Xin Mao, Wei Lu, Jin ziqi

We present Sailor, a family of open language models ranging from 0.5B to 14B parameters, tailored for South-East Asian (SEA) languages. From Qwen1.5, Sailor models accept 200B to 400B tokens during continual pre-training, primarily covering the languages of English, Chinese, Vietnamese, Thai, Indonesian, Malay, and Lao. The training leverages several techniques, including BPE dropout for improving the model robustness, aggressive data cleaning and deduplication, and small proxy models to optimize the data mixture. Experimental results on four typical tasks indicate that Sailor models demonstrate strong performance across different benchmarks, including commonsense reasoning, question answering, reading comprehension and examination. We share our insights to spark a wider interest in developing large language models for multilingual use cases.

(Nov 13): 7:45:45 (Morning) - Gather

DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning

Alexander Popov, Anastasia Voznyuk, Anna Korzanova, Dmitry Karpov, Fedor Ignatov, Maksim Savkin, Vasily Kononov

We present DeepPavlov 1.0, an open-source framework for using Natural Language Processing (NLP) models by leveraging transfer learning techniques. DeepPavlov 1.0 is created for modular and configuration-driven development of state-of-the-art NLP models and supports a wide range of NLP model applications. DeepPavlov 1.0 is designed for practitioners with limited knowledge of NLP/ML. DeepPavlov is based on PyTorch and supports HuggingFace transformers. DeepPavlov is publicly released under the Apache 2.0 license and provides access to an online demo.

(Nov 13): 7:45:45 (Morning) - Gather

Medico: Towards Hallucination Detection and Correction with Multi-source Evidence Fusion

Baotian Hu, Jifang Wang, Jindi Yu, Min Zhang, Xinpeng Zhao, Yibin Chen, zhenyu liu, dongfang li

As we all know, hallucinations prevail in Large Language Models (LLMs), where the generated content is coherent but factually incorrect, which inflicts a heavy blow on the widespread application of LLMs. Previous studies have shown that LLMs could confidently state non-existent facts rather than answering "I don't know". Therefore, it is necessary to resort to external knowledge to detect and correct the hallucinated content. Since manual detection and correction of factual errors is labor-intensive, developing an automatic end-to-end hallucination-checking approach is indeed a needful thing. To this end, we present Medico, a Multi-source evidence fusion enhanced hallucination detection and correction framework. It fuses diverse evidence from multiple sources, detects whether the generated content contains factual errors, provides the rationale behind the judgment, and iteratively revises the hallucinated content. Experimental results on evidence retrieval (0.964 HR@5, 0.908 MRR@5), hallucination detection (0.927-0.951 F1), and hallucination correction (0.973-0.979 approval rate) manifest the great potential of Medico. A video demo of Medico can be found at <https://youtu.be/RtsO6CSesBI>.

(Nov 13): 7:45:45 (Morning) - Gather

OpenOmni: A Collaborative Open Source Tool for Building Future-Ready Multimodal Conversational Agents

Qiang Sun, Sirui Li, Wei Liu, Wenxiao Zhang, Yuanyi Luo

Multimodal conversational agents are highly desirable because they offer natural and human-like interaction. However, there is a lack of comprehensive end-to-end solutions to support collaborative development and benchmarking. While proprietary systems like GPT-4o and Gemini demonstrating impressive integration of audio, video, and text with response times of 200-250ms, challenges remain in balancing latency, accuracy, cost, and data privacy. To better understand and quantify these issues, we developed **OpenOmni**, an open-source, end-to-end pipeline benchmarking tool that integrates advanced technologies such as Speech-to-Text, Emotion Detection, Retrieval Augmented Generation, Large Language Models, along with the ability to integrate customized models. OpenOmni supports local and cloud deployment, ensuring data privacy and supporting latency and accuracy benchmarking. This flexible framework allows researchers to customize the pipeline, focusing on real bottlenecks and facilitating rapid proof-of-concept development. OpenOmni can significantly enhance applications like indoor assistance for visually impaired individuals, advancing human-computer interaction. Our demonstration video is available <https://www.youtube.com/watch?v=zaSiT3clWqY>, demo is available via <https://openomni.ai4wa.com>, code is available via <https://github.com/AI4WA/OpenOmniFramework>.

(Nov 13): 7:458:45 (Morning) - Gather

CAVA: A Tool for Cultural Alignment Visualization & Analysis

Cheng Charles Ma, Daphne Ippolito, Nevan Giuliani, Prakruthi Pradeep

It is well-known that language models are biased; they have patchy knowledge of countries and cultures that are poorly represented in their training data. We introduce CAVA, a visualization tool for identifying and analyzing country-specific biases in language models. Our tool allows users to identify whether a language model successfully captures the perspectives of people of different nationalities. The tool supports analysis of both longform and multiple-choice models responses and comparisons between models. Our open-source code easily allows users to upload any country-based language model generations they wish to analyze. To showcase CAVA's efficacy, we present a case study analyzing how several popular language models answer survey questions from the World Values Survey.

(Nov 13): 7:458:45 (Morning) - Gather

OpenResearcher: Unleashing AI for Accelerated Scientific Research

Binjie Wang, Dongyu Ru, Jifan Lin, Lin Qiu, Pengfei Liu, Qingkai Min, Renjie Pan, Shichao Sun, Wenjie Li, Xuefeng Li, Yang Xu, Yun Luo, Yuxiang Zheng, Zizhao Zhang, Jiayang Cheng, Wang Ywen

The rapid growth of scientific literature imposes significant challenges for researchers endeavoring to stay updated with the latest advancements in their fields and delve into new areas. We introduce OpenResearcher, an innovative platform that leverages Artificial Intelligence (AI) techniques to accelerate the research process by answering diverse questions from researchers. OpenResearcher is built based on Retrieval-Augmented Generation (RAG) to integrate Large Language Models (LLMs) with up-to-date, domain-specific knowledge. Moreover, we develop various tools for OpenResearcher to understand researchers' queries, search from the scientific literature, filter retrieved information, provide accurate and comprehensive answers, and self-refine these answers. OpenResearcher can flexibly use these tools to balance efficiency and effectiveness. As a result, OpenResearcher enables researchers to save time and increase their potential to discover new insights and drive scientific breakthroughs. Demo, video, and code are available at: <https://github.com/GAIR-NLP/OpenResearcher>.

(Nov 13): 7:458:45 (Morning) - Gather

OpenT2T: An Open-Source Toolkit for Table-to-Text Generation

Arman Cohan, Haowei Zhang, Limyong Nan, Lyuhwan Chen, Pengcheng Wang, Shengyun Si, Yilun Zhao

Table data is pervasive in various industries, and its comprehension and manipulation demand significant time and effort for users seeking to extract relevant information. Consequently, an increasing number of studies have been directed towards table-to-text generation tasks. However, most existing methods are benchmarked solely on a limited number of datasets with varying configurations, leading to a lack of unified, standardized, fair, and comprehensive comparison between methods. This paper presents OpenT2T, the first open-source toolkit for table-to-text generation, designed to reproduce existing table pre-training models for performance comparison and expedite the development of new models. We have implemented and compared 7 fine-tuned models as well as 44 large language models under zero- and few-shot settings on 9 table-to-text generation datasets, covering data insight generation, table summarization, and free-form table question answering. Additionally, we maintain a public leaderboard to provide insights into how to choose appropriate table-to-text generation systems for real-world scenarios.

(Nov 13): 7:458:45 (Morning) - Gather

Xinference: Making Large Model Serving Easy

Feng Zhang, Lingfeng Xiong, Weizheng Lu, Xuev Qin, Yueguo Chen

The proliferation of open-source large models necessitates dedicated tools for deployment and accessibility. To mitigate the complexities of model serving, we develop Xinference, an open-source library designed to simplify the deployment and management of large models. Xinference effectively simplifies deployment complexities for users by (a) preventing users from writing code and providing built-in support for various models and OpenAI-compatible APIs; (b) enabling full model serving lifecycle management; (c) guaranteeing efficient and scalable inference and achieving high throughput and low latency. In comparative experiments with similar products like BentoML and Ray Serve, Xinference outperforms these tools and offers superior ease of use. Xinference is available at <https://github.com/xorbitsai/inference>.

(Nov 13): 7:458:45 (Morning) - Gather

Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts

Ika Karolina Idara, Lucky Susanto, Musa Izzanardi Wijanarko, Prasetya Anugrah Pratama, Derry Wijaya

Hate speech propagated online is a complex issue, deeply influenced by the cultural, historical, and societal contexts of both the perpetrator and the target. Consequently, the development of a universally robust hate speech classifier for diverse social media texts remains a challenging and unsolved task. The lack of mechanisms to track the spread and severity of hate speech further complicates the formulation of effective solutions. In response to this, to monitor hate speech in Indonesia during the recent 2024 presidential election, we have employed advanced Natural Language Processing (NLP) technologies to create an improved hate speech classifier tailored for a narrower subset of texts; specifically, texts that target minority groups that have historically been the targets of hate speech in Indonesia. Our focus is on texts that mention these seven vulnerable minority groups in Indonesia: Shia, Ahmadiyah, Christian, LGBTQ+, Chinese, Jewish, and people with disabilities. The insights gained from our dashboard have assisted stakeholders in devising more effective strategies to counteract hate speech. Notably, our dashboard has persuaded the General Election Supervisory Body in Indonesia (BAWASLU) to collaborate with our institution and the Alliance of Independent Journalists (AJI) to monitor social media hate speech in vulnerable areas in the country known for hate speech dissemination or hate-related violence in the upcoming Indonesian regional elections. This dashboard is available online at <https://www.aji.or.id/>

(Nov 13): 7:458:45 (Morning) - Gather

Instruction-Driven Game-Development and Game-Play: A Case Study for Poker

Hongqin Wu, Xingyuan Liu, Yan Wang, hai zhao

We present Instruction-Driven Game Engine (IDGE) system, a game system that democratizes game development by enabling a large language model (LLM) to autonomously generate game-play processes from free-form game descriptions. By interpreting natural language instructions, IDGE significantly lowers the barrier for creating custom games. We model IDGE learning as a Next State Prediction task, ensuring precise computation of game states to maintain game-play integrity. Utilizing a curriculum learning approach, we enhance the engine's stability and diversity. Our initial focus on Poker demonstrates the engine's capability to support various poker variants and novel user-defined games, laying the foundation for future advancements in instruction-driven game creation.

(Nov 13): 7:458:45 (Morning) - Gather

Evalverse: Unified and Accessible Library for Large Language Model Evaluation

Chanjun Park, Dahyun Kim, Jihoo Kim, Wonho Song, Yungi Kim, Yunsu Kim

This paper introduces Evalverse, a novel library that streamlines the evaluation of Large Language Models (LLMs) by unifying disparate evaluation tools into a single, user-friendly framework. Evalverse enables individuals with limited knowledge of artificial intelligence to easily request LLM evaluations and receive detailed reports, facilitated by an integration with communication platforms like Slack. Thus, Evalverse serves as a powerful tool for the comprehensive assessment of LLMs, offering both researchers and practitioners a centralized and

easily accessible evaluation framework. Finally, we also provide a demo video for Evalverse, showcasing its capabilities and implementation in a two-minute format.

(Nov 13): 7:45:45 (Morning) - Gather

TransAgents: Build Your Translation Company with Language Agents

Jiahao Xu, Longyue Wang, Minghao Wu

Multi-agent systems empowered by large language models (LLMs) have demonstrated remarkable capabilities in a wide range of downstream applications. In this work, we introduce TransAgents, a novel multi-agent translation system inspired by human translation companies. TransAgents employs specialized agents (Senior Editor, Junior Editor, Translator, Localization Specialist, and Proofreader) to collaboratively produce translations that are accurate, culturally sensitive, and of high quality. Our system is flexible, allowing users to configure their translation company based on specific needs, and universal, with empirical evidence showing superior performance across various domains compared to state-of-the-art methods. Additionally, TransAgents features a user-friendly interface and offers translations at a cost approximately $80 \times$ cheaper than professional human translation services. Evaluations on literary, legal, and financial test sets demonstrate that TransAgents produces translations preferred by human evaluators, even surpassing human-written references in literary contexts. Our live demo website is available at <https://www.transagents.ai/>. Our demonstration video is available at <https://www.youtube.com/watch?v=p7jAtF-WKc>.

Dialogue and Interactive Systems

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Class Name Guided Out-of-Scope Intent Classification

Chandan Gautam, Sethupathy Parameswaran, Aditya Kane, Yuan Fang, Savitha Ramasamy, Suresh Sundaram, Sunil Kumar Sahu, Xiaoli Li
The paper introduces Semantics of Class Label-based Unsupervised Out of Scope Intent Detection (SCOOS), a novel method aimed at enhancing out-of-scope (OOS) intent classification in task-oriented dialogue systems. Unlike prior approaches that rely solely on in-domain (ID) data features, SCOOS leverages semantic cues embedded in class labels to improve classification accuracy. The method entails forming a compact feature space centered around the semantics of class labels by minimizing losses between ID features and class names. SCOOS achieves this by creating a compact feature space centered around class label semantics, achieved through minimizing losses between in-domain (ID) features and class names. This involves training two spherical variational autoencoders concurrently to learn a shared latent space between ID features and class names, aligning ID feature data based on the corresponding classes in the latent space, and training a classifier for $(m + 1)$ -class classification using only ID samples, where the $(m + 1)^{th}$ class represents OOS samples. Extensive evaluation of three datasets demonstrates that SCOOS outperforms existing methods not only for OOS intent detection but also for ID intent classification. Additionally, an ablation study is conducted to analyze the impact of different components of SCOOS, and we also presented the visualization of the latent space representation providing insights into the influence of semantic information from class labels. The code will be made publicly accessible once the paper is accepted.

(Nov 13): 7:45:45 (Morning) - Gather

Strength Lies in Differences! Towards Effective Non-collaborative Dialogues via Tailored Strategy Planning

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, zujie wen, Wenqiang Lei, Tat-Seng Chua

We investigate non-collaborative dialogue agents, which are expected to engage in strategic conversations with diverse users, for securing a mutual agreement that leans favorably towards the systems objectives. This poses two main challenges for existing dialogue agents: 1) The inability to integrate user-specific characteristics into the strategic planning, and 2) The difficulty of training strategic planners that can be generalized to diverse users. To address these challenges, we propose TRIP to enhance the capability in tailored strategic planning, incorporating a user-aware strategic planning module and a population-based training paradigm. Through experiments on benchmark non-collaborative dialogue tasks, we demonstrate the effectiveness of TRIP in catering to diverse users.

(Nov 13): 7:45:45 (Morning) - Gather

Relevance Is a Guiding Light: Relevance-aware Adaptive Learning for End-to-end Task-oriented Dialogue System

Zhangpeng Chen, Zhihong Zhu, Wanshi Xu, Xianwei Zhuang, Yuexian Zou

Retrieving accurate domain knowledge and providing helpful information are crucial in developing an effective end-to-end task-oriented dialogue system (E2ETOD). The field has witnessed numerous methods following the retrieve-then-generate paradigm and training their systems on one specific domain. However, existing approaches still suffer from the Distractive Attributes Problem (DAP): struggling to deal with false but similar knowledge (hard negative entities), which is even more intractable when countless pieces of knowledge from different domains are blended in a real-world scenario. To alleviate DAP, we propose the Relevance-aware Adaptive Learning (ReAL) method, a two-stage training framework that eliminates hard negatives step-by-step and aligns retrieval with generation. In the first stage, we introduce a top-k adaptive contrastive loss and utilize the divergence-driven feedback from the frozen generator to pre-train the retriever. In the second stage, we propose using the metric score distribution as an anchor to align retrieval with generation. Thorough experiments on three benchmark datasets demonstrate ReAL's superiority over existing methods, with extensive analysis validating its strong capabilities of overcoming in- and cross-domain distractions.

(Nov 13): 7:45:45 (Morning) - Gather

Incomplete Utterance Rewriting with Editing Operation Guidance and Utterance Augmentation

Zhiyu Cao, PEIFENG LI, Yaxin FAN, Qiaoming Zhu

Although existing fashionable generation methods on Incomplete Utterance Rewriting (IUR) can generate coherent utterances, they often result in the inclusion of irrelevant and redundant tokens in rewritten utterances due to their inability to focus on critical tokens in dialogue context. Furthermore, the limited size of the training datasets also contributes to the insufficient training of the IUR model. To address the first issue, we propose a multi-task learning framework EO-IUR (Editing Operation-guided Incomplete Utterance Rewriting) that introduces the editing operation labels generated by sequence labeling module to guide generation model to focus on critical tokens. Furthermore, we introduce a token-level heterogeneous graph to represent dialogues. To address the second issue, we propose a two-dimensional utterance augmentation strategy, namely editing operation-based incomplete utterance augmentation and LLM-based historical utterance augmentation. The experimental results on three datasets demonstrate that our EO-IUR outperforms previous state-of-the-art (SOTA) baselines in both open-domain and task-oriented dialogue.

(Nov 13): 7:45:45 (Morning) - Gather

MORPHEUS: Modeling Role from Personalized Dialogue History by Exploring and Utilizing Latent Space

Yihong Tang, Bo Wang, Dongming Zhao, Jinxiaojia, Zhangjijun, Ruifang He, Yuxian Hou

Personalized Dialogue Generation (PDG) aims to create coherent responses according to roles or personas. Traditional PDG relies on external role data, which can be scarce and raise privacy concerns. Approaches address these issues by extracting role information from dialogue history, which often fail to generically model roles in continuous space. To overcome these limitations, we introduce a novel framework MORPHEUS through a three-stage training process. Specifically, we create a persona codebook to represent roles in latent space compactly, and this codebook is used to construct a posterior distribution of role information. This method enables the model to generalize across roles, allowing the generation of personalized dialogues even for unseen roles. Experiments on both Chinese and English datasets demonstrate that MORPHEUS enhances the extraction of role information, and improves response generation without external role data. Additionally, MORPHEUS can be considered an efficient fine-tuning for large language models.

(Nov 13): 7:458:45 (Morning) - Gather

Enhancing AI Assisted Writing with One-Shot Implicit Negative Feedback

Benjamin Towle, Ke Zhou

AI-mediated communication enables users to communicate more quickly and efficiently. Various systems have been proposed such as smart reply and AI-assisted writing. Yet, the heterogeneity of the forms of inputs and architectures often renders it challenging to combine insights from user behaviour in one system to improve performance in another. In this work, we consider the case where the user does not select any of the suggested replies from a smart reply system, and how this can be used as one-shot implicit negative feedback to enhance the accuracy of an AI writing model. We introduce Nifty, an approach that uses classifier guidance to controllably integrate implicit user feedback into the text generation process. Empirically, we find up to 34% improvement in Rouge-L, 89% improvement in generating the correct intent, and an 86% win-rate according to human evaluators compared to a vanilla AI writing system on the MultiWOZ and Schema-Guided Dialog datasets. The code is available at <https://github.com/BenjaminTowle/NIFTY>.

(Nov 13): 7:458:45 (Morning) - Gather

Game on Tree: Visual Hallucination Mitigation via Coarse-to-Fine View Tree and Game Theory

Xianwei Zhuang, Zhihong Zhu, Zhanpeng Chen, Yuxin Xie, Liming Liang, Yuxian Zou

Large Vision-Language Models (LVLMs) may produce outputs that are unfaithful to reality, also known as visual hallucinations (VH), which hinders their application in multimodal understanding and decision-making. In this work, we introduce a novel plug-and-play train-free decoding algorithm named Game and Tree-based Hallucination Mitigation (GTHM), designed for mitigating VH. GTHM is inspired by empirical observations that the fuzziness of multi-granularity view perception exacerbates VH. Based on this, GTHM leverages visual information to construct a coarse-to-fine visual view tree (CFTree) that organizes visual objects, attributes, and relationships in a hierarchical manner. Additionally, we innovatively model the optimal visual-token matching process on the CFTree as the cooperative game. Specifically, we define the Tree-based Shapley Value (TSV) for each visual view on the CFTree to assess its significant contribution to the overall visual understanding, thereby determining the optimal visual granularity. Subsequently, we utilize the TSV as guidance to implement adaptive weight contrastive decoding to achieve vision-aware decoding. Extensive experiments on four popular benchmarks confirm the effectiveness of our GTHM in alleviating VH across different LVLM families without additional training or post-processing. Our code is published at <https://github.com/-mengchuang123/GTHM>.

(Nov 13): 7:458:45 (Morning) - Gather

Large Language Model-based Human-Agent Collaboration for Complex Task Solving

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, Ji-Rong Wen

In recent developments within the research community, the integration of Large Language Models (LLMs) in creating fully autonomous agents has garnered significant interest. Despite this, LLM-based agents frequently demonstrate notable shortcomings in adjusting to dynamic environments and fully grasping human needs. In this work, we introduce the problem of LLM-based human-agent collaboration for complex task-solving, exploring their synergistic potential. To tackle the problem, we propose a Reinforcement Learning-based Human-Agent Collaboration method, ReHAC, which trains a policy model designed to determine the most opportune stages for human intervention within the task-solving process. We conduct experiments under real and simulated human-agent collaboration scenarios. Experimental results demonstrate that the synergistic efforts of humans and LLM-based agents significantly improve performance in complex tasks, primarily through well-planned, limited human intervention. Datasets and code are available at: <https://github.com/XueyangFeng/ReHAC/>.

(Nov 13): 7:458:45 (Morning) - Gather

BC-Prover: Backward Chaining Prover for Formal Theorem Proving

Dushan He, Jiahui Zhang, Jianzhu Bao, Fangquan Lin, Cheng Yang, Bing Qin, Rui Feng Xu, Wotao Yin

Despite the remarkable progress made by large language models in mathematical reasoning, interactive theorem proving in formal logic still remains a prominent challenge. Previous methods resort to neural models for proofstep generation and search. However, they suffer from exploring possible proofsteps empirically in a large search space. Moreover, they directly use a less rigorous informal proof for proofstep generation, neglecting the incomplete reasoning within. In this paper, we propose BC-Prover, a backward chaining framework guided by pseudo steps. Specifically, BC-Prover prioritizes pseudo steps to proofstep generation. The pseudo steps boost the proof construction in two aspects: (1) Backward Chaining that decomposes the proof into sub-goals for goal-oriented exploration. (2) Step Planning that makes a fine-grained planning to bridge the gap between informal and formal proofs. Experiments on the miniF2F benchmark show significant performance gains by our framework over the state-of-the-art approaches. Our framework is also compatible with existing provers and further improves their performance with the backward chaining technique.

(Nov 13): 7:458:45 (Morning) - Gather

Thoughts to Target: Enhance Planning for Target-driven Conversation

Zhonghua Zheng, Lizi Liao, Yang Deng, Ee-Peng Lim, Minlie Huang, Lijiang Nie

In conversational AI, large-scale models excel in various tasks but struggle with target-driven conversation planning. Current methods, such as chain-of-thought reasoning and tree-search policy learning techniques, either neglect plan rationality or require extensive human simulation procedures. Addressing this, we propose a novel two-stage framework, named EnPL, to improve the LLMs' capability in planning conversations towards designated targets, including (1) distilling natural language plans from target-driven conversation corpus and (2) generating new plans with demonstration-guided in-context learning. Specifically, we first propose a filter approach to distill a high-quality plan dataset, ConvPlan (Resources of this paper can be found at <https://github.com/pandazh2020/ConvPlan>). With the aid of corresponding conversational data and support from relevant knowledge bases, we validate the quality and rationality of these plans. Then, these plans are leveraged to help guide LLMs to further plan for new targets. Empirical results demonstrate that our method significantly improves the planning ability of LLMs, especially in target-driven conversations. Furthermore, EnPL is demonstrated to be quite effective in collecting target-driven conversation datasets and enhancing response generation, paving the way for constructing extensive target-driven conversational models.

(Nov 13): 7:458:45 (Morning) - Gather

Rescue Conversations from Dead-ends: Efficient Exploration for Task-oriented Dialogue Policy Optimization

Yangyang Zhao, Mehdi Dastani, Jinchuan Long, Zhenyu Wang, Shihai Wang

Training a task-oriented dialogue policy using deep reinforcement learning is promising but requires extensive environment exploration. The amount of wasted invalid exploration makes policy learning inefficient. In this paper, we define and argue that dead-end states are important reasons for invalid exploration. When a conversation enters a dead-end state, regardless of the actions taken afterward, it will continue in a dead-end trajectory until the agent reaches a termination state or maximum turn. We propose a Dead-end Detection and Resurrection (DDR) method that detects dead-end states in an efficient manner and provides a rescue action to guide and correct the exploration direction. To prevent dialogue policies from repeating errors, DDR also performs dialogue data augmentation by adding relevant experiences that include dead-end states and penalties into the experience pool. We first validate the dead-end detection reliability and then demonstrate the effectiveness and generality of the method across various domains through experiments on four public dialogue datasets.

Discourse and Pragmatics

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Improving Multi-party Dialogue Generation via Topic and Rhetorical Coherence

Yaxin Fan, Peifeng Li, Qiaoming Zhu

Previous studies on multi-party dialogue generation predominantly concentrated on modeling the reply-to structure of dialogue histories, always overlooking the coherence between generated responses and target utterances. To address this issue, we propose a Reinforcement Learning approach emphasizing both Topic and Rhetorical Coherence (RL-TRC). In particular, the topic- and rhetorical-coherence tasks are designed to enhance the model's perception of coherence with the target utterance. Subsequently, an agent is employed to learn a coherence policy, which guides the generation of responses that are topically and rhetorically aligned with the target utterance. Furthermore, three discourse-aware rewards are developed to assess the coherence between the generated response and the target utterance, with the objective of optimizing the policy. The experimental results and in-depth analyses on two popular datasets demonstrate that our RL-TRC significantly outperforms the state-of-the-art baselines, particularly in generating responses that are more coherent with the target utterances.

(Nov 13): 7:45:45 (Morning) - Gather

Language Models in Dialogue: Conversational Maxims for Human-AI Interactions

Erik Michling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorowski, John T. Richards

Modern language models, while sophisticated, exhibit some inherent shortcomings, particularly in conversational settings. We claim that many of the observed shortcomings can be attributed to violation of one or more conversational principles. By drawing upon extensive research from both the social sciences and AI communities, we propose a set of maxims – quantity, quality, relevance, manner, benevolence, and transparency – for describing effective human-AI conversation. We first justify the applicability of the first four maxims (from Grice) in the context of human-AI interactions. We then argue that two new maxims, benevolence (concerning the generation of, and engagement with, harmful content) and transparency (concerning recognition of one's knowledge boundaries, operational constraints, and intents), are necessary for addressing behavior unique to modern human-AI interactions. We evaluate the degree to which various language models are able to understand these maxims and find that models possess an internal prioritization of principles that can significantly impact accurate interpretability of the maxims.

Ethics, Bias, and Fairness

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Mitigate Extrinsic Social Bias in Pre-trained Language Models via Continuous Prompts Adjustment

Yiwei Dai, Hengrui Gu, Ying Wang, Xin Wang

Although pre-trained language models (PLMs) have been widely used in natural language understandings (NLU), they are still exposed to fairness issues. Most existing extrinsic debiasing methods rely on manually curated word lists for each sensitive groups to modify training data or to add regular constraints. However, these word lists are often limited by length and scope, resulting in the degradation performance of extrinsic bias mitigation. To address the aforementioned issues, we propose a **P**continuous **P**rompts, **A**djustment **D**ebiasing method (CPAD), which generates continuous token lists from the entire vocabulary space and uses them to bridge the gap between outputs and targets in fairness learning process. Specifically, CPAD encapsulates fine-tuning objective and debiasing objectives into several independent prompts. To avoid the limitation of manual word lists, in fairness learning phase, we extract outputs from the entire vocabulary space via fine-tuned PLM. Then, we aggregate the outputs from the same sensitive group as continuous token lists to map the output into protected attribute labels. Finally, after we learn the debiasing prompts in the perspective of adversarial learning, we improve fairness by adjusting continuous prompts at model inference time. Through extensive experiments on three NLU tasks, we evaluate the debiasing performance from the perspectives of group fairness and fairness through unawareness. The experimental results show that CPAD outperforms all baselines in term of single and two-attributes debiasing performance.

(Nov 13): 7:45:45 (Morning) - Gather

Distract Large Language Models for Automatic Jailbreak Attack

Zeguan Xiao, Yan Yang, Guanhua Chen, Yun Chen

Extensive efforts have been made before the public release of Large language models (LLMs) to align their behaviors with human values. However, even meticulously aligned LLMs remain vulnerable to malicious manipulations such as jailbreaking, leading to unintended behaviors. In this work, we propose a novel black-box jailbreak framework for automated red teaming of LLMs. We designed malicious content concealing and memory reframing with an iterative optimization algorithm to jailbreak LLMs, motivated by the research about the distractibility and over-confidence phenomenon of LLMs. Extensive experiments of jailbreaking both open-source and proprietary LLMs demonstrate the superiority of our framework in terms of effectiveness, scalability and transferability. We also evaluate the effectiveness of existing jailbreak defense methods against our attack and highlight the crucial need to develop more effective and practical defense strategies.

(Nov 13): 7:45:45 (Morning) - Gather

Knowledge-Centric Hallucination Detection

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, Zheng Zhang

Large Language Models (LLMs) have shown impressive capabilities but also a concerning tendency to hallucinate. This paper presents RefChecker, a framework that introduces claim-triplets to represent claims in LLM responses, aiming to detect fine-grained hallucinations. In RefChecker, an extractor generates claim-triplets from a response, which are then evaluated by a checker against a reference. We delineate three task settings: Zero, Noisy and Accurate Context, to reflect various real-world use cases. We curated a benchmark spanning various NLP tasks and annotated 11k claim-triplets from 2.1k responses by seven LLMs. RefChecker supports both proprietary and open-source models as the extractor and checker. Experiments demonstrate that claim-triplets enable superior hallucination detection, compared to other granularities such as response, sentence and sub-sentence level claims. RefChecker outperforms prior methods by 18.2 to 27.2 points on our benchmark and the checking results of RefChecker are strongly aligned with human judgments.

(Nov 13): 7:45:45 (Morning) - Gather

BaitAttack: Alleviating Intention Shift in Jailbreak Attacks via Adaptive Bait Crafting

Rui Pu, Chaozhuo Li, Rui Ha, Litian Zhang, Lirong Qiu, Xi Zhang

Jailbreak attacks enable malicious queries to evade detection by LLMs. Existing attacks focus on meticulously constructing prompts to disguise harmful intentions. However, the incorporation of sophisticated disguising prompts may incur the challenge of "intention shift". Intention shift occurs when the additional semantics within the prompt distract the LLMs, causing the responses to deviate significantly from the original harmful intentions. In this paper, we propose a novel component, "bait", to alleviate the effects of intention shift. Bait comprises an initial response to the harmful query, prompting LLMs to rectify or supplement the knowledge within the bait. By furnishing rich semantics relevant to the query, the bait helps LLMs focus on the original intention. To conceal the harmful content within the bait, we further propose a novel attack paradigm, BaitAttack. BaitAttack adaptively generates necessary components to persuade targeted LLMs that they are engaging with a legitimate inquiry in a safe context. Our proposal is evaluated on a popular dataset, demonstrating state-of-the-art attack performance and an exceptional capability for mitigating intention shift. The implementation of BaitAttack is accessible at: <https://anonymous.4open.science/r/BaitAttack-D1F5>.

(Nov 13): 7:45:45 (Morning) - Gather

CMD: a framework for Context-aware Model self-DeToxicification

Zecheng Tang, Keyan Zhou, Juntao Li, Yuyang Ding, Pinzheng Wang, Yan Bowen, Renjie Hua, Min Zhang

Text detoxification aims to minimize the risk of language models producing toxic content. Existing detoxification methods of directly constraining the model output or further training the model on the non-toxic corpus fail to achieve a decent balance between detoxification effectiveness and generation quality. This issue stems from the neglect of constraint imposed by the context since language models are designed to generate output that closely matches the context while detoxification methods endeavor to ensure the safety of the output even if it semantically deviates from the context. In view of this, we introduce a Context-aware Model self-DeToxicification (CMD) framework that pays attention to both the context and the detoxification process, i.e., first detoxifying the context and then making the language model generate along the safe context. Specifically, CMD framework involves two phases: utilizing language models to synthesize data and applying these data for training. We also introduce a toxic contrastive loss that encourages the model generation away from the negative toxic samples. Experiments on various LLMs have verified the effectiveness of our MSD framework, which can yield the best performance compared to baselines.

(Nov 13): 7:45:45 (Morning) - Gather

Universal Vulnerabilities in Large Language Models: Backdoor Attacks for In-context Learning

Shuai Zhao, Meihuiji Jia, Anh Tuan Luu, Fengjun Pan, Jimming Wan

In-context learning, a paradigm bridging the gap between pre-training and fine-tuning, has demonstrated high efficacy in several NLP tasks, especially in few-shot settings. Despite being widely applied, in-context learning is vulnerable to malicious attacks. In this work, we raise security concerns regarding this paradigm. Our studies demonstrate that an attacker can manipulate the behavior of large language models by poisoning the demonstration context, without the need for fine-tuning the model. Specifically, we design a new backdoor attack method, named ICLAttack, to target large language models based on in-context learning. Our method encompasses two types of attacks: poisoning demonstration examples and poisoning demonstration prompts, which can make models behave in alignment with predefined intentions. ICLAttack does not require additional fine-tuning to implant a backdoor, thus preserving the model's generality. Furthermore, the poisoned examples are correctly labeled, enhancing the natural stealth of our attack method. Extensive experimental results across several language models, ranging in size from 1.3B to 180B parameters, demonstrate the effectiveness of our attack method, exemplified by a high average attack success rate of 95.0% across the three datasets on OPT models.

(Nov 13): 7:45:45 (Morning) - Gather

How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Yongbin Li

Large language models (LLMs) rely on safety alignment to avoid responding to malicious user inputs. Unfortunately, jailbreak can circumvent safety guardrails, resulting in LLMs generating harmful content and raising concerns about LLM safety. Due to language models with intensive parameters often regarded as black boxes, the mechanisms of alignment and jailbreak are challenging to elucidate. In this paper, we employ weak classifiers to explain LLM safety through the intermediate hidden states. We first confirm that LLMs learn ethical concepts during pre-training rather than alignment and can identify malicious and normal inputs in the early layers. Alignment actually associates the early concepts with emotion guesses in the middle layers and then refines them to the specific reject tokens for safe generations. Jailbreak disturbs the transformation of early unethical classification into negative emotions. We conduct experiments on models from 7B to 70B across various model families to prove our conclusion. Overall, our paper indicates the intrinsic mechanism of LLM safety and how jailbreaks circumvent safety guardrails, offering a new perspective on LLM safety and reducing concerns.

(Nov 13): 7:45:45 (Morning) - Gather

Unlabeled Debiasing in Downstream Tasks via Class-wise Low Variance Regularization

Shahed Masoudian, Markus Froehmann, Navid Rekabsaz, Markus Schedl

Language models frequently inherit societal biases from their training data. Numerous techniques have been proposed to mitigate these biases during both the pre-training and fine-tuning stages. However, fine-tuning a pre-trained debiased language model on a downstream task can reintroduce biases into the model. Additionally, existing debiasing methods for downstream tasks either (i) require labels of protected attributes (e.g., age, race, or political views) that are often not available or (ii) rely on indicators of bias, which restricts their applicability to gender debiasing since they rely on gender-specific words. To address this, we introduce a novel debiasing regularization technique based on the class-wise variance of embeddings. Crucially, our method does not require attribute labels and targets any attribute, thus addressing the shortcomings of existing debiasing methods. Our experiments on encoder language models and three datasets demonstrate that our method outperforms existing strong debiasing baselines that rely on target attribute labels while maintaining performance on the target task.

(Nov 13): 7:45:45 (Morning) - Gather

Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, Qi Liu

Large Language Models (LLMs) have shown remarkable capabilities in various natural language processing tasks. However, LLMs may rely on dataset biases as shortcuts for prediction, which can significantly impair their robustness and generalization capabilities. This paper presents Shortcut Suite, a comprehensive test suite designed to evaluate the impact of shortcuts on LLMs' performance, incorporating six shortcut types, five evaluation metrics, and four prompting strategies. Our extensive experiments yield several key findings: 1) LLMs demonstrate varying reliance on shortcuts for downstream tasks, which significantly impairs their performance. 2) Larger LLMs are more likely to utilize shortcuts under zero-shot and few-shot in-context learning prompts. 3) Chain-of-thought prompting notably reduces shortcut reliance and outperforms other prompting strategies, while few-shot prompts generally underperform compared to zero-shot prompts. 4) LLMs often exhibit overconfidence in their predictions, especially when dealing with datasets that contain shortcuts. 5) LLMs generally have a lower explanation quality in shortcut-laden datasets, with errors falling into three types: distraction, disguised comprehension, and logical fallacy. Our findings offer new insights for evaluating robustness and generalization in LLMs and suggest potential directions for mitigating the reliance on shortcuts.

Generation

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

SEER: Self-Aligned Evidence Extraction for Retrieval-Augmented Generation

Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, Min Zhang

Recent studies in Retrieval-Augmented Generation (RAG) have investigated extracting evidence from retrieved passages to reduce computational costs and enhance the final RAG performance, yet it remains challenging. Existing methods heavily rely on heuristic-based augmentation, encountering several issues: (1) Poor generalization due to hand-crafted context filtering; (2) Semantics deficiency due to rule-based context chunking; (3) Skewed length due to sentence-wise filter learning. To address these issues, we propose a model-based evidence extraction learning framework, SEER, optimizing a vanilla model as an evidence extractor with desired properties through self-aligned learning. Extensive experiments show that our method largely improves the final RAG performance, enhances the faithfulness, helpfulness, and conciseness of the extracted evidence, and reduces the evidence length by 9.25 times. The code will be available at <https://github.com/HITsz-TMG/SEER>.

(Nov 13): 7:45:45 (Morning) - Gather

Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism

Lang Cao

Large language models (LLMs) have demonstrated impressive language understanding and generation capabilities, enabling them to answer a wide range of questions across various domains. However, these models are not flawless and often produce responses that contain errors or misinformation. These inaccuracies, commonly referred to as hallucinations, render LLMs unreliable and even unusable in many scenarios. In this paper, our focus is on mitigating the issue of hallucination in LLMs, particularly in the context of question-answering. Instead of attempting to answer all questions, we explore a refusal mechanism that instructs LLMs to refuse to answer challenging questions in order to avoid errors. We then propose a simple yet effective solution called Learn to Refuse (L2R), which incorporates the refusal mechanism to enable LLMs to recognize and refuse to answer questions that they find difficult to address. To achieve this, we utilize a structured knowledge base to represent all the LLM's understanding of the world, enabling it to provide traceable gold knowledge. This knowledge base is separate from the LLM and initially empty. It can be filled with validated knowledge and progressively expanded. When an LLM encounters questions outside its domain, the system recognizes its knowledge scope and determines whether it can answer the question independently. Additionally, we introduce a method for automatically and efficiently expanding the knowledge base of LLMs. Through qualitative and quantitative analysis, we demonstrate that our approach enhances the controllability and reliability of LLMs.

(Nov 13): 7:45:45 (Morning) - Gather

CltruS: Chunked Instruction-aware State Eviction for Long Sequence Modeling

Yu Bai, Xiyuan Zou, Heyan Huang, Sanxing Chen, Marc-Antoine Rondeau, Yang Gao, Jackie CK Cheung

Long sequence modeling has gained broad interest as large language models (LLMs) continue to advance. Recent research has identified that a large portion of hidden states within the key-value caches of Transformer models can be discarded (also termed evicted) without affecting the perplexity performance in generating long sequences. However, we show that these methods, despite preserving perplexity performance, often drop information that is important for solving downstream tasks, a problem which we call information neglect. To address this issue, we introduce Chunked Instruction-aware State Eviction (CltruS), a novel modeling technique that integrates the attention preferences useful for a downstream task into the eviction process of hidden states. In addition, we design a method for chunked sequence processing to further improve efficiency. Our training-free method exhibits superior performance on long sequence comprehension and retrieval tasks over several strong baselines under the same memory budget, while preserving language modeling perplexity. The code and data have been released at <https://github.com/ybai-nlp/CltruS>.

(Nov 13): 7:45:45 (Morning) - Gather

MirrorStories: Reflecting Diversity through Personalized Narrative Generation with Large Language Models

Sarfaroz Yunusov, Hamza Sidiq, Ali Emami

This study explores the effectiveness of Large Language Models (LLMs) in creating personalized "mirror stories" that reflect and resonate with individual readers' identities, addressing the significant lack of diversity in literature. We present MirrorStories, a corpus of 1,500 personalized short stories generated by integrating elements such as name, gender, age, ethnicity, reader interest, and story moral. We demonstrate that LLMs can effectively incorporate diverse identity elements into narratives, with human evaluators identifying personalized elements in the stories with high accuracy. Through a comprehensive evaluation involving 26 diverse human judges, we compare the effectiveness of MirrorStories against generic narratives. We find that personalized LLM-generated stories not only outscore generic human-written and LLM-generated ones across all metrics of engagement (with average ratings of 4.22 versus 3.37 on a 5-point scale), but also achieve higher textual diversity while preserving the intended moral. We also provide analyses that include bias assessments and a study on the potential for integrating images into personalized stories.

(Nov 13): 7:45:45 (Morning) - Gather

AdaSwitch: Adaptive Switching between Small and Large Agents for Effective Cloud-Local Collaborative Learning

Hao Sun, Jiayi Wu, Hengyi Cai, Xiaochi Wei, Yue Feng, Bo Wang, Shuaiqiang Wang, Yan Zhang, Dawei Yin

Recent advancements in large language models (LLMs) have been remarkable. Users face a choice between using cloud-based LLMs for generation quality and deploying local-based LLMs for lower computational cost. The former option is typically costly and inefficient, while

the latter usually fails to deliver satisfactory performance for reasoning steps requiring deliberate thought processes. In this work, we propose a novel LLM utilization paradigm that facilitates the collaborative operation of large cloud-based LLMs and smaller local-deployed LLMs. Our framework comprises two primary modules: the local agent instantiated with a relatively smaller LLM, handling less complex reasoning steps, and the cloud agent equipped with a larger LLM, managing more intricate reasoning steps. This collaborative processing is enabled through an adaptive mechanism where the local agent introspectively identifies errors and proactively seeks assistance from the cloud agent, thereby effectively integrating the strengths of both locally-deployed and cloud-based LLMs, resulting in significant enhancements in task completion performance and efficiency. We evaluate AdaSwitch across 7 benchmarks, ranging from mathematical reasoning and complex question answering, using various types of LLMs to instantiate the local and cloud agents. The empirical results show that AdaSwitch effectively improves the performance of the local agent, and sometimes achieves competitive results compared to the cloud agent while utilizing much less computational overhead.

(Nov 13): 7:45:45 (Morning) - Gather

Towards Verifiable Text Generation with Evolving Memory and Self-Reflection

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, Dawei Yin

Despite the remarkable ability of large language models (LLMs) in language comprehension and generation, they often suffer from producing factually incorrect information, also known as hallucination. A promising solution to this issue is verifiable text generation, which prompts LLMs to generate content with citations for accuracy verification. However, verifiable text generation is non-trivial due to the focus-shifting phenomenon, the intricate reasoning needed to align the claim with correct citations, and the dilemma between the precision and breadth of retrieved documents. In this paper, we present VTG, an innovative framework for Verifiable Text Generation with evolving memory and self-reflection. VTG introduces evolving long short-term memory to retain both valuable documents and recent documents. A two-tier verifier equipped with an evidence finder is proposed to rethink and reflect on the relationship between the claim and citations. Furthermore, active retrieval and diverse query generation are utilized to enhance both the precision and breadth of the retrieved documents. We conduct extensive experiments on five datasets across three knowledge-intensive tasks and the results reveal that VTG significantly outperforms baselines.

(Nov 13): 7:45:45 (Morning) - Gather

Curriculum Consistency Learning for Conditional Sentence Generation

Liangxin Liu, Xuebo Liu, Lian Lian, shengjun cheng, Jun Rao, Tengfei Yu, Hexuan Deng, Min Zhang

Consistency learning (CL) has proven to be a valuable technique for improving the robustness of models in conditional sentence generation (CSG) tasks by ensuring stable predictions across various input data forms. However, models augmented with CL often face challenges in optimizing consistency features, which can detract from their efficiency and effectiveness. To address these challenges, we introduce Curriculum Consistency Learning (CCL), a novel strategy that guides models to learn consistency in alignment with their current capacity to differentiate between features. CCL is designed around the inherent aspects of CL-related losses, promoting task independence and simplifying implementation. Implemented across four representative CSG tasks, including instruction tuning (IT) for large language models and machine translation (MT) in three modalities (text, speech, and vision), CCL demonstrates marked improvements. Specifically, it delivers +2.0 average accuracy point improvement compared with vanilla IT and an average increase of +0.7 in COMET scores over traditional CL methods in MT tasks. Our comprehensive analysis further indicates that models utilizing CCL are particularly adept at managing complex instances, showcasing the effectiveness and efficiency of CCL in improving CSG models. Code and scripts are available at <https://github.com/xinxinxing/Curriculum-Consistency-Learning>.

(Nov 13): 7:45:45 (Morning) - Gather

Amateur-free Contrastive Decoding via Cognitive Layers Skipping

Wenhao Zhu, Sizhe Liu, Shujian Huang, Shuaijie She, Chris Wendl, Jiajun Chen

Decoding by contrasting layers (DoLa), is designed to improve the generation quality of large language models (LLMs) by contrasting the prediction probabilities between the early exit output (amateur logits) and the final output (expert logits).However, we find that this approach does not work well on non-English tasks.Inspired by previous interpretability work on language transition during the model's forward pass, we discover that this issue arises from a language mismatch between early exit output and final output.In this work, we propose an improved contrastive decoding algorithm that is effective for diverse languages beyond English.To obtain more helpful amateur logits, we devise two strategies to skip a set of bottom, language-agnostic layers based on our preliminary analysis.Experimental results on multilingual reasoning benchmarks demonstrate that our proposed method outperforms previous contrastive decoding baselines and substantially improves LLM's chain-of-thought reasoning accuracy across 11 languages.

(Nov 13): 7:45:45 (Morning) - Gather

Leveraging Large Language Models for NLG Evaluation: Advances and Challenges

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, Shuai Ma

In the rapidly evolving domain of Natural Language Generation (NLG) evaluation, introducing Large Language Models (LLMs) has opened new avenues for assessing generated content quality, e.g., coherence, creativity, and context relevance. This paper aims to provide a thorough overview of leveraging LLMs for NLG evaluation, a burgeoning area that lacks a systematic analysis. We propose a coherent taxonomy for organizing existing LLM-based evaluation metrics, offering a structured framework to understand and compare these methods. Our detailed exploration includes critically assessing various LLM-based methodologies, as well as comparing their strengths and limitations in evaluating NLG outputs. By discussing unresolved challenges, including bias, robustness, domain-specificity, and unified evaluation, this paper seeks to offer insights to researchers and advocate for fairer and more advanced NLG evaluation techniques.

(Nov 13): 7:45:45 (Morning) - Gather

Improving Diversity of Commonsense Generation by Large Language Models via In-Context Learning

Tianhui Zhang, Bei Peng, Danushka Bollegala

Generative Commonsense Reasoning (GCR) requires a model to reason about a situation using commonsense knowledge, while generating coherent sentences. Although the quality of the generated sentences is crucial, the diversity of the generation is equally important because it reflects the model's ability to use a range of commonsense knowledge facts. Large Language Models (LLMs) have shown proficiency in enhancing the generation quality across various tasks through in-context learning (ICL) using given examples without the need for any fine-tuning. However, the diversity aspect in LLM outputs has not been systematically studied before. To address this, we propose a simple method that diversifies the LLM generations, while preserving their quality. Experimental results on three benchmark GCR datasets show that our method achieves an ideal balance between the quality and diversity. Moreover, the sentences generated by our proposed method can be used as training data to improve diversity in existing commonsense generators.

(Nov 13): 7:45:45 (Morning) - Gather

Self-supervised Preference Optimization: Enhance Your Language Model with Preference Degree Awareness

Jian Li, Haojing Huang, Yujia Zhang, Pengfei Xu, Xi Chen, Rui Song, Lida Shi, Jingwen Wang, Hao Xu

Recently, there has been significant interest in replacing the reward model in Reinforcement Learning with Human Feedback (RLHF) methods for Large Language Models (LLMs), such as Direct Preference Optimization (DPO) and its variants. These approaches commonly use a

binary cross-entropy mechanism on pairwise samples, i.e., minimizing and maximizing the loss based on preferred or dis-preferred responses, respectively. However, while this training strategy omits the reward model, it also overlooks the varying preference degrees within different responses. We hypothesize that this is a key factor hindering LLMs from sufficiently understanding human preferences. To address this problem, we propose a novel Self-supervised Preference Optimization (SPO) framework, which constructs a self-supervised preference degree loss combined with the alignment loss, thereby helping LLMs improve their ability to understand the degree of preference. Extensive experiments are conducted on two widely used datasets of different tasks. The results demonstrate that SPO can be seamlessly integrated with existing preference optimization methods and significantly boost their performance to achieve state-of-the-art performance. We also conduct detailed analyses to offer comprehensive insights into SPO, which verifies its effectiveness. The code is available at <https://github.com/lilian16/SPO>.

(Nov 13): 7:45:45 (Morning) - Gather

Inference-Time Language Model Alignment via Integrated Value Guidance

Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, Yu Qiao

Large language models are typically fine-tuned to align with human preferences, but tuning large models is computationally intensive and complex. In this work, we introduce **Integrated Value Guidance (IVG)**, a method that uses implicit and explicit value functions to guide language model decoding at token and chunk-level respectively, efficiently aligning large language models purely at inference time. This approach circumvents the complexities of direct fine-tuning and outperforms traditional methods. Empirically, we demonstrate the versatility of IVG across various tasks. In controlled sentiment generation and summarization tasks, our method significantly improves the alignment of large models using inference-time guidance from **gpt2**-based value functions. Moreover, in a more challenging instruction-following benchmark AlpacaEval 2.0, we show that both specifically tuned and off-the-shelf value functions greatly improve the length-controlled win rates of large models against gpt-4-turbo (e.g., 19.51 % → 26.51% for **Mistral-7B-Instruct-v0.2** and 25.58 % → 33.75% for **Mistral-8x7B-Instruct-v0.1** with Tulu guidance).

(Nov 13): 7:45:45 (Morning) - Gather

Tree of Problems: Improving structured problem solving with compositionality

Armel Randy Zebaze, Benoît Sagot, Rachel Bawden

Large Language Models (LLMs) have demonstrated remarkable performance across multipletasks through in-context learning. For complex reasoning tasks that require step-by-step thinking, Chain-of-Thought (CoT) prompting has given impressive results, especially when combined with self-consistency. Nonetheless, some tasks remain particularly difficult for LLMs to solve. Tree of Thoughts (ToT) and Graph of Thoughts (GoT) emerged as alternatives, dividing the complex problem into paths of subproblems. In this paper, we propose Tree of Problems (ToP), a simpler version of ToT, which we hypothesise can work better for complex tasks that can be divided into identical subtasks. Our empirical results show that our approach outperforms ToT and GoT, and in addition performs better than CoT on complex reasoning tasks. All code for this paper will be made available.

(Nov 13): 7:45:45 (Morning) - Gather

Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, Le Sun

The rapid development of large language models has led to the widespread adoption of Retrieval-Augmented Generation (RAG), which integrates external knowledge to alleviate knowledge bottlenecks and mitigate hallucinations. However, the existing RAG paradigm inevitably suffers from the impact of flawed information introduced during the retrieval phase, thereby diminishing the reliability and correctness of the generated outcomes. In this paper, we propose Credibility-aware Generation (CAG), a universally applicable framework designed to mitigate the impact of flawed information in RAG. At its core, CAG aims to equip models with the ability to discern and process information based on its credibility. To this end, we propose an innovative data transformation framework that generates data based on credibility, thereby effectively endowing models with the capability of CAG. Furthermore, to accurately evaluate the models' capabilities of CAG, we construct a comprehensive benchmark covering three critical real-world scenarios. Experimental results demonstrate that our model can effectively understand and employ credibility for generation, significantly outperform other models with retrieval augmentation, and exhibit robustness despite the increasing noise in the context.

(Nov 13): 7:45:45 (Morning) - Gather

Extending Context Window of Large Language Models from a Distributional Perspective

Yingsheng Wu, Yuxuan Gu, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin

Scaling the rotary position embedding (RoPE) has become a common method for extending the context window of RoPE-based large language models (LLMs). However, existing scaling methods often rely on empirical approaches and lack a profound understanding of the internal distribution within RoPE, resulting in suboptimal performance in extending the context window length. In this paper, we propose to optimize the context window extending task from the view of rotary angle distribution. Specifically, we first estimate the distribution of the rotary angles within the model and analyze the extent to which length extension perturbs this distribution. Then, we present a novel extension strategy that minimizes the disturbance between rotary angle distributions to maintain consistency with the pre-training phase, enhancing the model's capability to generalize to longer sequences. Experimental results compared to the strong baseline methods demonstrate that our approach reduces by up to 72% of the distributional disturbance when extending LLaMA2's context window to 8k, and reduces by up to 32% when extending to 16k. On the LongBench-E benchmark, our method achieves an average improvement of up to 4.33% over existing state-of-the-art methods. Furthermore, Our method maintains the model's performance on the Hugging Face Open LLM benchmark after context window extension, with only an average performance fluctuation ranging from -0.12 to +0.22.

Industry

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

QUSI: Question-guided Insights Generation for Automated Exploratory Data Analysis

Abhijit Manatkar, Ashlesha Akella, Parthiv Gupta, Krishnasuri Narayanan

Discovering meaningful insights from a large dataset, known as Exploratory Data Analysis (EDA), is a challenging task that requires thorough exploration and analysis of the data. Automated Data Exploration (ADE) systems use goal-oriented methods with Large Language Models and Reinforcement Learning towards full automation. However, these methods require human intervention to anticipate goals that may limit insight extraction, while fully automated systems demand significant computational resources and retraining for new datasets. We introduce QUSI, a fully automated EDA system that operates in two stages: insight generation (ISGen) driven by question generation (QUGen). The QUGen module generates questions in iterations, refining them from previous iterations to enhance coverage without human intervention or manually curated examples. The ISGen module analyzes data to produce multiple relevant insights in response to each question, requiring no

prior training and enabling QUIS to adapt to new datasets.

(Nov 13): 7:458:45 (Morning) - Gather

News Risk Alerting System (NRAS): A Data-Driven LLM Approach to Proactive Credit Risk Monitoring

Ashish Upadhyay, Xenia Skotti, Adil Nygaard, Lauren Hinkle, Joe Halliwel, Ian C Brown, Glen Noronha

Credit risk monitoring is an essential process for financial institutions to evaluate the creditworthiness of borrowing entities and minimize potential losses. Traditionally, this involves periodic and best-efforts assessment of news about companies to identify events which impact on credit risk. This is time-consuming, and can delay response to critical developments. The News Risk Alerting System (NRAS) proactively identifies credit-relevant news related to clients and alerts the relevant Credit Officer (CO). This production system has been deployed for nearly three years and has alerted COs to over 2700 credit-relevant events with an estimated precision of 77%.

(Nov 13): 7:458:45 (Morning) - Gather

LARA: Linguistic-Adaptive Retrieval-Augmentation for Multi-Turn Intent Classification

Junhua Liu, Yong Keat Tan, Bin Fu, Kwan Hui Lim

Multi-turn intent classification is notably challenging due to the complexity and evolving nature of conversational contexts. This paper introduces LARA, a Linguistic-Adaptive Retrieval-Augmentation framework to enhance accuracy in multi-turn classification tasks across six languages, accommodating numerous intents in chatbot interactions. LARA combines a fine-tuned smaller model with a retrieval-augmented mechanism, integrated within the architecture of LLMs. The integration allows LARA to dynamically utilize past dialogues and relevant intents, thereby improving the understanding of the context. Furthermore, our adaptive retrieval techniques bolster the cross-lingual capabilities of LLMs without extensive retraining and fine-tuning. Comprehensive experiments demonstrate that LARA achieves state-of-the-art performance on multi-turn intent classification tasks, enhancing the average accuracy by 3.67% from state-of-the-art single-turn intent classifiers.

(Nov 13): 7:458:45 (Morning) - Gather

TPTU-v2: Boosting Task Planning and Tool Usage of Large Language Model-based Agents in Real-world Industry Systems

Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shi Shiwei, Du Guo Qing, Xiaorui Hu, Hangyu Mao, Ziyue Li, Xingyu Zeng, Rui Jia, Xueqian Wang

Large Language Models (LLMs) have demonstrated proficiency in addressing tasks that necessitate a combination of task planning and the usage of external tools, such as weather and calculator APIs. However, real-world industrial systems present prevalent challenges in task planning and tool usage: numerous APIs in the real system make it intricate to invoke the appropriate one, while the inherent limitations of LLMs pose challenges in orchestrating an accurate sub-task sequence and API-calling order. This paper introduces a comprehensive framework aimed at enhancing the Task Planning and Tool Usage (TPTU) abilities of LLM-based agents in industry. Our framework comprises three key components designed to address these challenges: (1) the API Retriever selects the most pertinent APIs among the extensive API set; (2) the Demo Selector retrieves demonstrations related to hard-to-distinguish APIs, which is further used for in-context learning to boost the final performance; (3) LLM Finetuner tunes a base LLM to enhance its capability for task planning and API calling. We validate our methods using a real-world industry system and an open-sourced academic dataset, demonstrating the efficacy of each individual component as well as the integrated framework. The anonymous code is here.

(Nov 13): 7:458:45 (Morning) - Gather

Detecting Ambiguous Utterances in an Intelligent Assistant

Satoshi Akasaki, Manabu Sassano

In intelligent assistants that perform both chatting and tasks through dialogue, like Siri and Alexa, users often make ambiguous utterances such as "I'm hungry" or "I have a headache," which can be interpreted as either chat or task intents. Naively determining these intents can lead to mismatched responses, spoiling the user experience. Therefore, it is desirable to determine the ambiguity of user utterances. We created a dataset from an actual intelligent assistant via crowdsourcing and analyzed tendencies of ambiguous utterances. Using this labeled data of chat, task, and ambiguous intents, we developed a supervised intent classification model. To detect ambiguous utterances robustly, we propose feeding sentence embeddings developed from microblogs and search logs with a self-attention mechanism. Experiments showed that our model outperformed two baselines, including a strong LLM-based one. We will release the dataset upon acceptance to support future research.

(Nov 13): 7:458:45 (Morning) - Gather

FanLoRA: Fantastic LoRAs and Where to Find Them in Large Language Model Fine-tuning

Wei Zhu, Yi Ge, Xing Tian, Yi Yin, Congrui Yin, Aaron Xuxiang Tian

Full-parameter fine-tuning is computationally prohibitive for large language models (LLMs), making parameter-efficient fine-tuning (PEFT) methods like low-rank adaptation (LoRA) increasingly popular. However, LoRA and its existing variants introduce significant latency in multi-tenant settings, hindering their applications in the industry. To address this issue, we propose the Fantastic LoRA (FanLoRA) framework, which consists of four steps: (a) adding LoRA modules to all the Transformer linear weights and fine-tuning on a large-scale instruction tuning dataset. (b) The importance of each module is then assessed using a novel importance scoring method. (c) only the most critical modules per layer are retained, resulting in the FanLoRA setting. (d) The FanLoRA setting is applied to fine-tune various downstream tasks. Our extensive experiments demonstrate that: (a) FanLoRA outperforms existing PEFT baselines across a wide collection of tasks with comparable tunable parameters. (b) FanLoRA significantly reduces the inference latency of LoRA, making it valuable for further broadening the applications of LLMs in the industry.⁹

(Nov 13): 7:458:45 (Morning) - Gather

Sample Design Engineering: An Empirical Study on Designing Better Fine-Tuning Samples for Information Extraction with LLMs

Biyang Guo, He Wang, WenyiLiu Xiao, Hong Chen, Zhuxin Lee, Songqiao Han, Haizhang Huang

Large language models (LLMs) have achieved significant leadership in many NLP tasks, but aligning structured output with generative models in information extraction (IE) tasks remains a challenge. Prompt Engineering (PE) is renowned for improving IE performance through prompt modifications. However, the realm of the sample design for downstream fine-tuning, crucial for task-specific LLM adaptation, is largely unexplored. This paper introduces **Sample Design Engineering** (SDE), a methodical approach to enhancing LLMs' post-tuning performance on IE tasks by refining input, output, and reasoning designs. Through extensive ID and OOD experiments across six LLMs, we first assess the impact of various design options on IE performance, revealing several intriguing patterns. Based on these insights, we then propose an integrated SDE strategy and validate its consistent superiority over heuristic sample designs on three complex IE tasks with four additional LLMs, demonstrating the generality of our method. Additionally, analyses of LLMs' inherent prompt/output perplexity, zero-shot, and ICL abilities illustrate that good PE strategies may not always translate to good SDE strategies.

(Nov 13): 7:458:45 (Morning) - Gather

Breaking the Hourglass Phenomenon of Residual Quantization: Enhancing the Upper Bound of Generative Retrieval

⁹Due to the company's policy, codes will be open-sourced to facilitate future research.

Zhirui Kuai, Zuxu Chen, Wang Binbin, Dadong Miao, Xusong Chen, Jiaxing Wang, Lin Liu, Songlin Wang, Guoyu Tang, Li Kuang, Yuxing Han, Mingwei Li, Huijun Wang, Jingwei Zhuo

Generative retrieval (GR) has emerged as a transformative paradigm in search and recommender systems, leveraging numeric-based identifier representations to enhance efficiency and generalization. Notably, methods like TIGER, which employ Residual Quantization-based Semantic Identifiers (RQ-SID), have shown significant promise in e-commerce scenarios by effectively managing item IDs. However, a critical issue termed the "Hourglass" phenomenon, occurs in RQ-SID, where intermediate codebook tokens become overly concentrated, hindering the full utilization of generative retrieval methods. This paper analyses and addresses this problem by identifying data sparsity and long-tailed distribution as the primary causes. Through comprehensive experiments and detailed ablation studies, we analyze the impact of these factors on codebook utilization and data distribution. Our findings reveal that the "Hourglass" phenomenon substantially impacts the performance of RQ-SID in generative retrieval. We propose effective solutions to mitigate this issue, thereby significantly enhancing the effectiveness of generative retrieval in real-world E-commerce applications.

(Nov 13): 7:45:45 (Morning) - Gather

Improving Few-Shot Cross-Domain Named Entity Recognition by Instruction Tuning a Word-Embedding based Retrieval Augmented Large Language Model

Subhadip Nandi, Neeraj Agrawal

Few-Shot Cross-Domain NER is the process of leveraging knowledge from data-rich source domains to perform entity recognition on data-scarce target domains. Most previous state-of-the-art (SOTA) approaches use pre-trained language models (PLMs) for cross-domain NER. However, these models are often domain specific. To successfully use these models for new target domains, we need to modify either the model architecture or perform model fine-tuning using data from the new domains. Both of these result in the creation of entirely new NER models for each target domain which is infeasible for practical scenarios. Recently, several works have attempted to use LLMs to solve Few-Shot Cross-Domain NER. However, most of these are either too expensive for practical purposes or struggle to follow LLM prompt instructions. In this paper, we propose IF-WRANER (Instruction Finetuned Word-embedding based Retrieval Augmented large language model for Named Entity Recognition), a retrieval augmented LLM, finetuned for the NER task. By virtue of the regularization techniques used during LLM finetuning and the adoption of word-level embedding over sentence-level embedding during the retrieval of in-prompt examples, IF-WRANER is able to outperform previous SOTA Few-Shot Cross-Domain NER approaches. We have demonstrated the effectiveness of our model by benchmarking its performance on the open source CrossNER dataset, on which it shows more than 2% F1 score improvement over the previous SOTA model. We have deployed the model for multiple customer care domains of an enterprise. Accurate entity prediction through IF-WRANER helps direct customers to automated workflows for the domains, thereby reducing escalations to human agents by almost 15% and leading to millions of dollars in yearly savings for the company.

(Nov 13): 7:45:45 (Morning) - Gather

PARA: Parameter-Efficient Fine-tuning with Prompt-Aware Representation Adjustment

Wei Zhu, Zeguan Liu, Ming Tan, Yi Zhao, Aaron Xuxiang Tian

Despite the presence of many competitive parameter-efficient fine-tuning (PEFT) methods like LoRA, the industry still needs a PEFT method that is efficient under the single-backbone multi-tenant setting in industrial applications while performing competitively in the downstream tasks. In this work, we propose a novel yet simple PEFT method, Prompt Aware Representation Adjustment (PARA). We propose installing a lightweight vector generator at each Transformer layer to generate vectors conditioned on the input prompts that will modify the hidden representations. We have conducted experiments on various tasks, and the experimental results demonstrate that: (a) our PARA method can outperform the recent PEFT baselines with comparable tunable parameters. (b) Our PARA method is more efficient than LoRA under the single-backbone multi-tenant setting, showing great potential for the industry.

(Nov 13): 7:45:45 (Morning) - Gather

ULMR: Unlearning Large Language Models via Negative Response and Model Parameter Average

Shaojie Shi, Xiaoyu Tan, Xihé Qiu, Chao Qu, Kexin Nie, Yuan Cheng, Wei Chu, Xu Yinghui, Yuan Qi

In recent years, large language models (LLMs) have attracted significant interest from the research community due to their broad applicability in many language-oriented tasks, and are now widely used in numerous areas of production and daily life. One source of the powerful capabilities of LLMs is the massive scale of their pre-training dataset. However, these pre-training datasets contain many outdated, harmful, and personally sensitive information, which inevitably becomes memorized by LLM during the pre-training process. Eliminating this undesirable data is crucial for ensuring the model's safety and enhancing the user experience. However, the cost of extensively cleaning the pre-training dataset and retraining the model from scratch is very high. In this work, we propose ULMR, a unlearning framework for LLMs, which first uses carefully designed prompts to rewrite the instructions in the specified dataset, and generate corresponding negative responses. Subsequently, to ensure that the model does not excessively deviate post-training, we perform model parameter averaging to preserve the performance of the original LLM. We conducted experiments on two public datasets, TOFU and RWKU, demonstrating that our method can effectively forget specified information while retaining the capabilities of the original LLM.

(Nov 13): 7:45:45 (Morning) - Gather

ProConSuL: Project Context for Code Summarization with LLMs

Vadim Lomshakov, Andrey Podivilov, Sergey Savin, Oleg Baryshnikov, Alena Lisevych, Sergey Nikolenko

We propose Project Context for Code Summarization with LLMs (ProConSuL), a new framework to provide a large language model (LLM) with precise information about the code structure from program analysis methods such as a compiler or IDE language services and use task decomposition derived from the code structure. ProConSuL builds a call graph to provide the context from callees and uses a two-phase training method (SFT + preference alignment) to train the model to use the project context. We also provide a new evaluation benchmark for C/C++ functions and a set of proxy metrics. Experimental results demonstrate that ProConSuL allows to significantly improve code summaries and reduce the number of hallucinations compared to the base model (CodeLlama-7B-instruct). We make our code and dataset available at <https://github.com/TypingCat13/ProConSuL>.

(Nov 13): 7:45:45 (Morning) - Gather

Building an Efficient Multilingual Non-Profit IR System for the Islamic Domain Leveraging Multiprocessing Design in Rust

Vera Pavlova, Mohammed Makhlof

The widespread use of large language models (LLMs) has dramatically improved many applications of Natural Language Processing (NLP), including Information Retrieval (IR). However, domains that are not driven by commercial interest often lag behind in benefiting from AI-powered solutions. One such area is religious and heritage corpora. Alongside similar domains, Islamic literature holds significant cultural value and is regularly utilized by scholars and the general public. Navigating this extensive amount of text is challenging, and there is currently no unified resource that allows for easy searching of this data using advanced AI tools. This work focuses on the development of a multilingual non-profit IR system for the Islamic domain. This process brings a few major challenges, such as preparing multilingual domain-specific corpora when data is limited in certain languages, deploying a model on resource-constrained devices, and enabling fast search on a limited budget. By employing methods like continued pre-training for domain adaptation and language reduction to decrease model size, a lightweight multilingual retrieval model was prepared, demonstrating superior performance compared to larger models pre-trained on general

domain data. Furthermore, evaluating the proposed architecture that utilizes Rust Language capabilities shows the possibility of implementing efficient semantic search in a low-resource setting.

(Nov 13): 7:458:45 (Morning) - Gather

Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow

Tian Guo, Emmanuel Haupmann

Large language models (LLMs) and related fine-tuning techniques have demonstrated superior performance on various language understanding and generation tasks. This paper explores fine-tuning LLMs for stock return forecasting with financial newsflow. In quantitative investing, return forecasting is fundamental for subsequent tasks like stock picking, portfolio optimization, etc. We formulate the model consisting of the text representation and forecasting modules. We propose to compare the encoder-only and decoder-only LLMs, considering they generate text representations in distinct ways. How these different representations affect the forecasting performance is still an open question. Meanwhile, we present two simple methods of integrating LLMs' token-level representations into the forecasting module. The experiments on real news and investment universes show that encoder and decoder LLMs perform comparably for predicting the upside stocks, and decoder LLMs are more capable of capturing the downside risks; in the small investment universe with less data for fine-tuning, large decoder LLMs' forecasts lead to better-performing portfolios than encoder LLMs; the return prediction-based portfolios outperform sentiment-based portfolios.

(Nov 13): 7:458:45 (Morning) - Gather

Language, OCR, Form Independent (LOFI) pipeline for Industrial Document Information Extraction

Chang Oh Yoon, Wonbeen Lee, Seokhwan Jang, Kyuwon Choi, Minsung Jung, Daewoo Choi

This paper presents LOFI (Language, OCR, Form Independent), a pipeline for Document Information Extraction (DIE) in Low-Resource Language (LRL) business documents. LOFI pipeline solves language, OCR, and form dependencies through flexible language model integration, a token-level box split algorithm, and the SPADE decoder. Experiments on Korean and Japanese documents demonstrate high performance without additional pre-training. The pipeline's effectiveness is validated through real-world applications in insurance and tax-free declaration services, advancing DIE capabilities for diverse languages and document types in industrial settings.

(Nov 13): 7:458:45 (Morning) - Gather

Knowledge-augmented Financial Market Analysis and Report Generation

Yuemin Chen, Feijian Wu, Jingwei Wang, Hao Qian, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Meng Wang

Crafting a convincing financial market analysis report necessitates a wealth of market information and the expertise of financial analysts, posing a highly challenging task. While large language models (LLMs) have enabled the automated generation of financial market analysis texts, they still face issues such as hallucinations, errors in financial knowledge, and insufficient capability to reason about complex financial problems, which limits the quality of the generation. To tackle these shortcomings, we propose a novel task and a retrieval-augmented framework grounded in a financial knowledge graph (FKG). The proposed framework is compatible with commonly used instruction-tuning methods. Experiments demonstrate that our framework, coupled with a small-scale language model fine-tuned with instructions, can significantly enhance the logical consistency and quality of the generated analysis texts, outperforming both large-scale language models and other retrieval-augmented baselines.

(Nov 13): 7:458:45 (Morning) - Gather

mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, Min Zhang

We present systematic efforts in building long-context multilingual text representation model (TRM) and reranker from scratch for text retrieval. We first introduce a text encoder (base size) enhanced with RoPE and unpadding, pre-trained in a native 8192-token context (longer than 512 of previous multilingual encoders). Then we construct a hybrid TRM and a cross-encoder reranker by contrastive learning. Evaluations show that our text encoder outperforms the same-sized previous state-of-the-art XLM-R. Meanwhile, our TRM and reranker match the performance of large-sized state-of-the-art BGE-M3 models and achieve better results on long-context retrieval benchmarks. Further analysis demonstrate that our proposed models exhibit higher efficiency during both training and inference. We believe their efficiency and effectiveness could benefit various researches and industrial applications.

(Nov 13): 7:458:45 (Morning) - Gather

The Program Testing Ability of Large Language Models for Code

Weimin Xiong, Yiwu Guo, Hao Chen

Recent development of large language models (LLMs) for code like CodeX and CodeT5+ shows promise in achieving code intelligence. Their ability of synthesizing program targeting a pre-defined algorithmic coding task has been intensively tested and verified on datasets including HumanEval and MBPP. Yet, evaluation of these LLMs from more perspectives (than just program synthesis) is also anticipated, considering their broad scope of applications. In this paper, we explore their ability of automatic test cases generation. We show intriguing observations and reveal how the quality of their generated test cases can be improved. Following recent work which uses generated test cases to enhance program synthesis, we further leverage our findings in improving the quality of the synthesized programs and show +11.77% and +4.22% higher code pass rates on HumanEval+ comparing with the GPT-3.5-turbo baseline and the recent state-of-the-art, respectively.

(Nov 13): 7:458:45 (Morning) - Gather

CharacterGLM: Customizing Social Characters with Large Language Models

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, Jiaming Yang, Xiaoxiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, Minlie Huang

Character-based dialogue (CharacterDial) has become essential in the industry (e.g., CharacterAI), enabling users to freely customize social characters for social interactions. However, the generalizability and adaptability across various conversational scenarios inherent in customizing social characters still lack public industrial solutions. To address these challenges, by dissecting well-rounded social characters composed of both inherent social profiles and external social behaviors, we manually collect a large-scale Chinese corpus featuring characters with diverse categories and behaviors, and develop CharacterGLM models alongside well-designed refinement methods. Extensive experiments show that CharacterGLM outperforms most popular open- and closed-source LLMs and performs comparably to GPT-4. We will release our data and models for local development and deployment.

(Nov 13): 7:458:45 (Morning) - Gather

FuxiTranyu: A Multilingual Large Language Model Trained with Balanced Data

Haoran Sun, Renren Jin, Shaoyang Xu, Leiyu Pan, Supryadi, Menglong Cui, Jiangcun Du, Yikun Lei, Lei Yang, Ling Shi, Juesi Xiao, Shaolin Zhu, Deyi Xiong

Large language models (LLMs) have demonstrated prowess in a wide range of tasks. However, many LLMs exhibit significant performance discrepancies between high-resource and low-resource languages, often observed in existing multilingual models, which tend to have limited focus on enhancing their multilingual capabilities. To address this gap, we introduce FuxiTranyu, an open-sourced multilingual LLM

designed to address the research community's need for balanced and high-performing multilingual models. FuxiTanyu-8B, the foundational model, contains 8 billion parameters and is trained from scratch on a meticulously balanced dataset of 600 billion tokens spanning 43 natural languages and 16 programming languages. Beyond releasing the foundational model, we also introduce FuxiTanyu-8B-Instruct, fine-tuned on a diverse multilingual instruction dataset, and FuxiTanyu-8B-Chat, further refined with DPO on a preference dataset for enhanced conversational ability. Extensive evaluation on multilingual benchmarks demonstrate FuxiTanyu's superior performance compared to existing multilingual LLMs like BLOOM-7B and PolyLM-13B.

(Nov 13): 7:45:45 (Morning) - Gather

Intent Detection in the Age of LLMs

Gaurav Arora, Shreya Jain, Srijana Merugu

Intent detection is a critical component of task-oriented dialogue systems (TODS) which enables the identification of suitable actions to address user utterances at each dialog turn. Traditional approaches relied on computationally efficient supervised sentence transformer encoder models, which require substantial training data and struggle with out-of-scope (OOS) detection. The emergence of generative large language models (LLMs) with intrinsic world knowledge presents new opportunities to address these challenges. In this work, we adapt SOTA LLMs using adaptive in-context learning and chain-of-thought prompting for intent detection, and compare their performance with contrastively fine-tuned sentence transformer (SetFit) models to highlight prediction quality and latency tradeoff. We propose a hybrid system using uncertainty-based routing strategy to combine the two approaches that along with negative data augmentation results in achieving the best of both worlds (i.e. within 2% of native LLM accuracy with 50% less latency). To better understand LLM OOS detection capabilities, we perform controlled experiments revealing that this capability is significantly influenced by the scope of intent labels and the size of the label space. We also introduce a two-step approach utilizing internal LLM representations, demonstrating empirical gains in OOS detection accuracy and F1-score by >5% for the MISTRAL-7B model.

(Nov 13): 7:45:45 (Morning) - Gather

Aegis: An Advanced LLM-Based Multi-Agent for Intelligent Functional Safety Engineering

Lu Shi, Bin Qi, Jianru Luo, Yang Zhang, Zhanzhao Liang, Zhaowei Gao, Wenke Deng, Lin Sun

Functional safety is a critical aspect of automotive engineering, encompassing all phases of a vehicle's lifecycle, including design, development, production, operation, and decommissioning. This domain involves highly knowledge-intensive tasks. This paper introduces Aegis: An Advanced LLM-Based Multi-Agent for Intelligent Functional Safety Engineering. Aegis is specifically designed to support complex functional safety tasks within the automotive sector. It is tailored to perform Hazard Analysis and Risk Assessment (HARA), document Functional Safety Requirements (FSR), and plan test cases for Automatic Emergency Braking (AEB) systems. The most advanced version, Aegis-Max, leverages Retrieval-Augmented Generation (RAG) and reflective mechanisms to enhance its capability in managing complex, knowledge-intensive tasks. Additionally, targeted prompt refinement by professional functional safety practitioners can significantly optimize Aegis' performance in the functional safety domain. This paper demonstrates the potential of Aegis to improve the efficiency and effectiveness of functional safety processes in automotive engineering.

(Nov 13): 7:45:45 (Morning) - Gather

Efficient Answer Retrieval System (EARS): Combining Local DB Search and Web Search for Generative QA

Nikita Krayko, Ivan Sidorov, Fedor Laputin, Daria Galimzianova, Vasily Konovalov

In this work we propose an efficient production-ready factoid question answering (QA) system that combines a local knowledge base search and generative context-based QA. To assess the quality of the generated content, we devise metrics for both manual and automatic evaluation of the answers to questions. A distinctive feature of our system is the Ranker component, which ranks potential answers according to their relevance. This enhances the quality of local knowledge base retrieval by 23%. Another crucial aspect of the system is the LLM, which utilizes contextual information from the internet to formulate responses. It boosts the utility of voice-based answers by 94%. **EARS** is language-agnostic and the approach can be implemented for any data domain.

(Nov 13): 7:45:45 (Morning) - Gather

Mixture of Diverse Size Experts

Manxi Sun, Wei Liu, Jian Luan, Pengzhi Gao, Bin Wang

The Sparsely-Activated Mixture-of-Experts (MoE) architecture has gained popularity for scaling large language models (LLMs) due to the sub-linearly increasing computational costs. Despite its success, most of the current structure designs face the challenge that the experts share the same size such that tokens have no chance to choose the experts with the most appropriate size to generate the next token. To mitigate this defect, we propose Mixture of Diverse Size Experts (MoDSE), a new MoE architecture with designed layers where experts have different sizes. Analysis on difficult token generation tasks shows that experts with different sizes give better predictions, and the routing path of the experts tends to be stable after a period of training. The diversity of experts' size will lead to load unbalancing. To tackle this limitation, we introduce an expert-pair allocation strategy to distribute the workload evenly across the GPUs. Comprehensive evaluations across multiple benchmarks demonstrate the effectiveness of MoDSE, surpassing existing MoEs by adaptively assigning the parameter budget to experts while maintaining the same total parameter size and number of experts.

(Nov 13): 7:45:45 (Morning) - Gather

Course-Correction: Safety Alignment Using Synthetic Preferences

Rongwu Xu, Yishuo Cai, Zhenhong Zhou, Renjie Gu, Haiqin Weng, Liu Yan, Tianwei Zhang, Wei Xu, Han Qiu

The risk of harmful contents generated by large language models (LLMs) becomes a critical concern. This paper systematically evaluates and enhances LLMs' capability to perform course-correction, i.e., the model can steer away from generating harmful content autonomously. First, we introduce the C²-Eval benchmark for quantitative assessment and analyze 10 popular LLMs, revealing varying proficiency of current safety-tuned LLMs in course-correction. To improve, we propose fine-tuning LLMs with preference learning, emphasizing the preference for timely course-correction. Using an automated pipeline, we create C²-Syn, a synthetic C²-Syn with 750K pairwise preferences, to teach models the concept of timely course-correction through data-driven learning. Experiments on LLAMA2-CHAT 7B and Qwen2 7B show that our method effectively enhances course-correction skills without affecting general performance. Additionally, it effectively improves LLMs' safety, particularly in resisting jailbreak attacks.

(Nov 13): 7:45:45 (Morning) - Gather

GOVERN: Gradient Orientation Vote Ensemble for Multi-Teacher Reinforced Distillation

Wenjie Zhou, Zhenxin Ding, Xiaodong Zhang, Haibo Shi, Junfeng Wang, Dawei Yin

Pre-trained language models have become an integral component of question-answering systems, achieving remarkable performance. For practical deployment, it is critical to carry out knowledge distillation to preserve high performance under computational constraints. In this paper, we address a key question: given the importance of unsupervised distillation for student performance, how does one effectively ensemble knowledge from multiple teachers at this stage without the guidance of labels? We propose a novel algorithm, GOVERN, to tackle this issue. GOVERN has demonstrated significant improvements in both offline and online experiments. The proposed algorithm has been

successfully deployed in a real-world commercial question-answering system.

(Nov 13): 7:458:45 (Morning) - Gather

Scaling Parameter-Constrained Language Models with Quality Data

Ernie Chang, Matteo Paltenghi, Yang Li, Pin-Jie Lin, Changsheng Zhao, Patrick Huber, Zechun Liu, Rastislav Rabatin, Yangyang Shi, Vikas Chandra

Scaling laws in language modeling traditionally quantify training loss as a function of dataset size and model parameters, providing compute-optimal estimates but often neglecting the impact of data quality on model generalization. In this paper, we extend the conventional understanding of scaling law by offering a microscopic view of data quality within the original formulation – *effective training tokens* – which we posit to be a critical determinant of performance for parameter-constrained language models. Specifically, we formulate the proposed term of effective training tokens to be a combination of two readily-computed indicators of text: (i) text diversity and (ii) syntheticity as measured by a teacher model. We pretrained over 200 models of 25M to 1.5B parameters on a diverse set of sampled, synthetic data, and estimated the constants that relate text quality, model size, training tokens, and eight reasoning task accuracy scores. We demonstrated the estimated constants yield +0.83 Pearson correlation with true accuracies, and analyze it in scenarios involving widely-used data techniques such as data sampling and synthesis which aim to improve data quality.

(Nov 13): 7:458:45 (Morning) - Gather

DL-QAT: Weight-Decomposed Low-Rank Quantization-Aware Training for Large Language Models

Wenjing Ke, Zhe Li, Dong Li, Lu Tian, Emad Barsoom

Improving the efficiency of inference in Large Language Models (LLMs) is a critical area of research. Post-training Quantization (PTQ) is a popular technique, but it often faces challenges at low-bit levels, particularly in downstream tasks. Quantization-aware Training (QAT) can alleviate this problem, but it requires significantly more computational resources. To tackle this, we introduced Weight-Decomposed Low-Rank Quantization-Aware Training (DL-QAT), which merges the advantages of QAT while training only less than 1% of the total parameters. Specifically, we introduce a group-specific quantization magnitude to adjust the overall scale of each quantization group. Within each quantization group, we use LoRA matrices to update the weight size and direction in the quantization space. We validated the effectiveness of our method on the LLaMA and LLaMA2 model families. The results show significant improvements over our baseline method across different quantization granularities. For instance, for LLaMA-7B, our approach outperforms the previous state-of-the-art method by 4.2% in MMLU on 3-bit LLaMA-7B. Additionally, our quantization results on pre-trained models also surpass previous QAT methods, demonstrating the superior performance and efficiency of our approach.

(Nov 13): 7:458:45 (Morning) - Gather

Hybrid-RACA: Hybrid Retrieval-Augmented Composition Assistance for Real-time Text Prediction

Menglin Xia, Xuchao Zhang, Camille Couturier, Guojing Zheng, Saravanan Rajmohan, Victor Rühle

Large language models (LLMs) enhanced with retrieval augmentation has shown great performance in many applications. However, the computational demands for these models pose a challenge when applying them to real-time tasks, such as composition assistance. To address this, we propose Hybrid Retrieval-Augmented Composition Assistance (Hybrid-RACA), a novel system for real-time text prediction that efficiently combines a cloud-based LLM with a smaller client-side model through retrieval augmented memory. This integration enables the client model to generate better responses, benefiting from the LLM's capabilities and cloud-based data. Meanwhile, via a novel asynchronous memory update mechanism, the client model can deliver real-time completions to user inputs without the need to wait for responses from the cloud. Our experiments on five datasets demonstrate that Hybrid-RACA offers strong performance while maintaining low latency.

(Nov 13): 7:458:45 (Morning) - Gather

LLMC: Benchmarking Large Language Model Quantization with a Versatile Compression Toolkit

Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, Xianglong Liu

Recent advancements in large language models (LLMs) are propelling us toward artificial general intelligence with their remarkable emergent abilities and reasoning capabilities. However, the substantial computational and memory requirements limit the widespread adoption. Quantization, a key compression technique, can effectively mitigate these demands by compressing and accelerating LLMs, albeit with potential risks to accuracy. Numerous studies have aimed to minimize the accuracy loss associated with quantization. However, their quantization configurations vary from each other and cannot be fairly compared. In this paper, we present LLMC, a plug-and-play compression toolkit, to fairly and systematically explore the impact of quantization. LLMC integrates dozens of algorithms, models, and hardwares, offering high extensibility from integer to floating-point quantization, from LLM to vision-language (VLM) model, from fixed-bit to mixed precision, and from quantization to sparsification. Powered by this versatile toolkit, our benchmark covers three key aspects: calibration data, algorithms (three strategies), and data formats, providing novel insights and detailed analyses for further research and practical guidance for users. Our toolkit is available at <https://github.com/anonymous-emnlp123/llmc>.

(Nov 13): 7:458:45 (Morning) - Gather

Context Matters: Pushing the Boundaries of Open-Ended Answer Generation with Graph-Structured Knowledge Context

Sonmath Banerjee, Amrit Sahoo, Sayan Layek, Avik Dutta, Rima Hazra, Animesh Mukherjee

This paper introduces a novel framework that combines graph-driven context retrieval in conjunction to knowledge graphs based enhancement, honing the proficiency of LLMs, especially in domain specific community question answering platforms like AskUbuntu, Unix, and ServerFault. We conduct experiments on various LLMs with different parameter sizes to evaluate their ability to ground knowledge and determine factual accuracy in answers to open-ended questions. Our methodology GraphContextGen consistently outperforms dominant text-based retrieval systems, demonstrating its robustness and adaptability to a larger number of use cases. This advancement highlights the importance of pairing context rich data retrieval with LLMs, offering a renewed approach to knowledge sourcing and generation in AI systems. We also show that, due to rich contextual data retrieval, the crucial entities, along with the generated answer, remain factually coherent with the gold answer. We shall release the source code and datasets upon acceptance.

(Nov 13): 7:458:45 (Morning) - Gather

Pretraining and Finetuning Language Models on Geospatial Networks for Accurate Address Matching

Saket Maheshwary, Arpan Paul, Saurabh Sohney

We propose a novel framework for pretraining and fine-tuning language models with the goal of determining whether two addresses represent the same physical building. For delivery and logistics, improving address matching positively impacts geocoding, route planning, and delivery time estimations, leading to an efficient and reliable delivery experience. We propose to view a list of addresses as an address graph and curate inputs for language models by placing geospatially linked addresses in the same context. Our approach jointly integrates concepts from graph theory and weak supervision with address text and geospatial semantics. This integration enables us to generate informative and diverse address pairs, facilitating pretraining and fine-tuning in a self-supervised manner. Experiments and ablation studies on manually curated datasets and comparisons with state-of-the-art techniques demonstrate the efficacy of our proposed approach. We achieve a 24.49% improvement in recall while maintaining 95% precision on average, in comparison to the current baseline across multiple geographies. Further, we demonstrate the impact of improving address matching on geocode learning. We performed offline evaluations and launched online

A/B experiments which show that our proposed approach improves delivery precision by 14.68% and reduces delivery defects by 8.79% on average across geographies.

Information Extraction

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Advancing Event Causality Identification via Heuristic Semantic Dependency Inquiry Network

Haoran Li, Qiang Gao, Hongmei Wu, Li Huang

Event Causality Identification (ECI) focuses on extracting causal relations between events in texts. Existing methods for ECI primarily rely on causal features and external knowledge. However, these approaches fall short in two dimensions: (1) causal features between events in a text often lack explicit clues, and (2) external knowledge may introduce bias, while specific problems require tailored analyses. To address these issues, we propose SemDI - a simple and effective Semantic Dependency Inquiry Network for ECI. SemDI captures semantic dependencies within the context using a unified encoder. Then, it utilizes a Cloze Analyzer to generate a fill-in token based on comprehensive context understanding. Finally, this fill-in token is used to inquire about the causal relation between two events. Extensive experiments demonstrate the effectiveness of SemDI, surpassing state-of-the-art methods on three widely used benchmarks. Code is available at <https://github.com/hrlics/SemDI>.

(Nov 13): 7:45:45 (Morning) - Gather

Cross-domain NER with Generated Task-Oriented Knowledge: An Empirical Study from Information Density Perspective

Zhihao Zhang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, Guodong Zhou

Cross-domain Named Entity Recognition (CDNER) is crucial for Knowledge Graph (KG) construction and natural language processing (NLP), enabling learning from source to target domains with limited data. Previous studies often rely on manually collected entity-relevant sentences from the web or attempt to bridge the gap between tokens and entity labels across domains. These approaches are time-consuming and inefficient, as these data are often weakly correlated with the target task and require extensive pre-training. To address these issues, we propose automatically generating task-oriented knowledge (GTOK) using large language models (LLMs), focusing on the reasoning process of entity extraction. Then, we employ task-oriented pre-training (TOPD) to facilitate domain adaptation. Additionally, current cross-domain NER methods often lack explicit explanations for their effectiveness. Therefore, we introduce the concept of information density to better evaluate the model's effectiveness before performing entity recognition. We conduct systematic experiments and analyses to demonstrate the effectiveness of our proposed approach and the validity of using information density for model evaluation.

(Nov 13): 7:45:45 (Morning) - Gather

NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoît Crabillé, Etienne P Bernard

Large Language Models (LLMs) have shown impressive abilities in data annotation, opening the way for new approaches to solve classic NLP problems. In this paper, we show how to use LLMs to create NuNER, a compact language representation model specialized in the Named Entity Recognition (NER) task. NuNER can be fine-tuned to solve downstream NER problems in a data-efficient way, outperforming similar-sized foundation models in the few-shot regime and competing with much larger LLMs. We find that the size and entity-type diversity of the pre-training dataset are key to achieving good performance. We view NuNER as a member of the broader family of task-specific foundation models, recently unlocked by LLMs. NuNER and NuNERs dataset are open-sourced with MIT License.

(Nov 13): 7:45:45 (Morning) - Gather

Multi-Level Cross-Modal Alignment for Speech Relation Extraction

Liang Zhang, Zhen Yang, Biao Fu, Ziyao Lu, Liangying Shao, Shiyu Liu, Fandong Meng, Jie Zhou, Xiaoli Wang, Jinsong Su

Speech Relation Extraction (SpeechRE) aims to extract relation triplets from speech data. However, existing studies usually use synthetic speech to train and evaluate SpeechRE models, hindering the further development of SpeechRE due to the disparity between synthetic and real speech. Meanwhile, the modality gap issue, unexplored in SpeechRE, limits the performance of existing models. In this paper, we construct two real SpeechRE datasets to facilitate subsequent researches and propose a Multi-level Cross-modal Alignment Model (MCAM) for SpeechRE. Our model consists of three components: 1) a speech encoder, extracting speech features from the input speech; 2) an alignment adapter, mapping these speech features into a suitable semantic space for the text decoder; and 3) a text decoder, autoregressively generating relation triplets based on the speech features. During training, we first additionally introduce a text encoder to serve as a semantic bridge between the speech encoder and the text decoder, and then train the alignment adapter to align the output features of speech and text encoders at multiple levels. In this way, we can effectively train the alignment adapter to bridge the modality gap between the speech encoder and the text decoder. Experimental results and in-depth analysis on our datasets strongly demonstrate the efficacy of our method.

(Nov 13): 7:45:45 (Morning) - Gather

RoCEL: Advancing Table Entity Linking through Distinctive Row and Column Contexts

Yuanzheng Wang, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Xueqi Cheng

Table entity linking (TEL) aims to map entity mentions in the table to their corresponding entities in a knowledge base (KB). The core of this task is to leverage structured contexts, specifically row and column contexts, to enhance the semantics of mentions in entity disambiguation. Most entity linking (EL) methods primarily focus on understanding sequential text contexts, making it difficult to adapt to the row and column structure of tables. Additionally, existing methods for TEL indiscriminately mix row and column contexts together, overlooking their semantic differences. In this paper, we explicitly distinguish the modeling of row and column contexts, and propose a method called RoCEL to capture their distinct semantics. Specifically, for row contexts in tables, we take the attention mechanism to learn the implicit relational dependencies between each cell and the mention. For column contexts in tables, we employ a set-wise encoder to learn the categorical information about the group of mentions. At last, we merge both contexts to obtain the final mention embedding for link prediction. Experiments on four benchmarks show that our approach outperforms the state-of-the-art (SOTA) baseline by about 1.5% on the in-domain dataset, and by 3.7% on average across three out-of-domain datasets.

(Nov 13): 7:45:45 (Morning) - Gather

Efficient Overshadowed Entity Disambiguation by Mitigating Shortcut Learning

Panuthep Tasawong, Peerat Limkachotai, Potsavee Manakul, Kan Udomcharoenchaikit, Ekapol Chuangsawanich, Sarana Nutanong

Entity disambiguation (ED) is crucial in natural language processing (NLP) for tasks such as question-answering and information extraction. A major challenge in ED is handling overshadowed entities—uncommon entities sharing mention surfaces with common entities. The current approach to enhance performance on these entities involves reasoning over facts in a knowledge base (KB), increasing computational over-

head during inference. We argue that the ED performance on overshadowed entities can be enhanced during training by addressing shortcut learning, which does not add computational overhead at inference. We propose a simple yet effective debiasing technique to prevent models from shortcut learning during training. Experiments on a range of ED datasets show that our method achieves state-of-the-art performance without compromising inference speed. Our findings suggest a new research direction for improving entity disambiguation via shortcut learning mitigation.

(Nov 13): 7:458:45 (Morning) - Gather

SRF: Enhancing Document-Level Relation Extraction with a Novel Secondary Reasoning Framework

Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, Yongxue Wu

Document-level Relation Extraction (DocRE) aims to extract relations between entity pairs in a document and poses many challenges as it involves multiple mentions of entities and cross-sentence inference. However, several aspects that are important for DocRE have not been considered and explored. Existing work ignores bidirectional mention interaction when generating relational features for entity pairs. Also, sophisticated neural networks are typically designed for cross-sentence evidence extraction to further enhance DocRE. More interestingly, we reveal a noteworthy finding: If a model has predicted a relation between an entity and other entities, this relation information may help infer and predict more relations between the entity's adjacent entities and these other entities. Nonetheless, none of existing methods leverage secondary reasoning to exploit results of relation prediction. To this end, we propose a novel Secondary Reasoning Framework (SRF) for DocRE. In SRF, we initially propose a DocRE model that incorporates bidirectional mention fusion and a simple yet effective evidence extraction module (incurring only an additional learnable parameter overhead) for relation prediction. Further, for the first time, we elaborately design and propose a novel secondary reasoning method to discover more relations by exploring the results of the first relation prediction. Extensive experiments show that SRF achieves SOTA performance and our secondary reasoning method is both effective and general when integrated into existing models.

(Nov 13): 7:458:45 (Morning) - Gather

Unleashing the Power of Large Language Models in Zero-shot Relation Extraction via Self-Prompting

Siyi Liu, Yang Li, Jiang Li, Shan Yang, Yunshi Lan

Recent research in zero-shot Relation Extraction (RE) has focused on using Large Language Models (LLMs) due to their impressive zero-shot capabilities. However, current methods often perform suboptimally, mainly due to a lack of detailed, context-specific prompts needed for understanding various sentences and relations. To address this, we introduce the Self-Prompting framework, a novel method designed to fully harness the embedded RE knowledge within LLMs. Specifically, our framework employs a three-stage diversity approach to prompt LLMs, generating multiple synthetic samples that encapsulate specific relations from scratch. These generated samples act as in-context learning samples, offering explicit and context-specific guidance to efficiently prompt LLMs for RE. Experimental evaluations on benchmark datasets show our approach outperforms existing LLM-based zero-shot RE methods. Additionally, our experiments confirm the effectiveness of our generation pipeline in producing high-quality synthetic data that enhances performance.

Information Retrieval and Text Mining

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

Bridging Cultures in the Kitchen: A Framework and Benchmark for Cross-Cultural Recipe Retrieval

Tianyi Hu, Maria Mastro, Daniel Hershcovich

The cross-cultural adaptation of recipes is an important application of identifying and bridging cultural differences in language. The challenge lies in retaining the essence of the original recipe while also aligning with the writing and dietary habits of the target culture. Information Retrieval (IR) offers a way to address the challenge because it retrieves results from the culinary practices of the target culture while maintaining relevance to the original recipe. We introduce a novel task about cross-cultural recipe retrieval and present a unique Chinese-English cross-cultural recipe retrieval benchmark. Our benchmark is manually annotated under limited resource, utilizing various retrieval models to generate a pool of candidate results for manual annotation. The dataset provides retrieval samples that are culturally adapted but textually diverse, presenting greater challenges. We propose CARROT, a plug-and-play cultural-aware recipe information retrieval framework that incorporates cultural-aware query rewriting and re-ranking methods and evaluate it both on our benchmark and intuitive human judgments. The results show that our framework significantly enhances the preservation of the original recipe and its cultural appropriateness for the target culture. We believe these insights will significantly contribute to future research on cultural adaptation.

(Nov 13): 7:458:45 (Morning) - Gather

Optimizing Code Retrieval: High-Quality and Scalable Dataset Annotation through Large Language Models

Rui Li, Qi Liu, Liyang He, Zheng Zhang, Hao Zhang, Shengyu Ye, Junyu Lu, Zhenya Huang

Code retrieval aims to identify code from extensive codebases that semantically aligns with a given query code snippet. Collecting a broad and high-quality set of query and code pairs is crucial to the success of this task. However, existing data collection methods struggle to effectively balance scalability and annotation quality. In this paper, we first analyze the factors influencing the quality of function annotations generated by Large Language Models (LLMs). We find that the invocation of intra-repository functions and third-party APIs plays a significant role. Building on this insight, we propose a novel annotation method that enhances the annotation context by incorporating the content of functions called within the repository and information on third-party API functionalities. Additionally, we integrate LLMs with a novel sorting method to address the multi-level function call relationships within repositories. Furthermore, by applying our proposed method across a range of repositories, we have developed the Query4Code dataset. The quality of this synthesized dataset is validated through both model training and human evaluation, demonstrating high-quality annotations. Moreover, cost analysis confirms the scalability of our annotation method.

(Nov 13): 7:458:45 (Morning) - Gather

Dual-Phase Accelerated Prompt Optimization

Muchen Yang, Moxin Li, Yongle Li, Zijun Chen, Chongming Gao, Junji Zhang, Yangyang Li, Fulai Feng

Gradient-free prompt optimization methods have made significant strides in enhancing the performance of closed-source Large Language Model (LLMs) across a wide range of tasks. However, existing approaches make light of the importance of high-quality prompt initialization and the identification of effective optimization directions, thus resulting in substantial optimization steps to obtain satisfactory performance. In this light, we aim to accelerate prompt optimization process to tackle the challenge of low convergence rate. We propose a dual-phase approach which starts with generating high-quality initial prompts by adopting a well-designed meta-instruction to delve into task-specific information, and iteratively optimize the prompts at the sentence level, leveraging previous tuning experience to expand prompt candidates and accept effective ones. Extensive experiments on eight datasets demonstrate the effectiveness of our proposed method, achieving a consistent accuracy gain over baselines with less than five optimization steps.

(Nov 13): 7:45:45 (Morning) - Gather

An LLM-Enabled Knowledge Elicitation and Retrieval Framework for Zero-Shot Cross-Lingual Stance Identification

Ruike Zhang, Yuan Penghui Wei, Daniel Dajun Zeng, Wenji Mao

Stance detection aims to identify the attitudes toward specific targets from text, which is an important research area in text mining and social media analytics. Existing research is mainly conducted in monolingual setting on English datasets. To tackle the data scarcity problem in low-resource languages, cross-lingual stance detection (CLSD) transfers the knowledge from high-resource (source) language to low-resource (target) language. The CLSD task is the most challenging in zero-shot setting when no training data is available in target language, and transferring stance-relevant knowledge learned from high-resource language to bridge the language gap is the key for improving the performance of zero-shot CLSD. In this paper, we leverage the capability of large language model (LLM) for stance knowledge acquisition, and propose KEAR, a knowledge elicitation and retrieval framework. The knowledge elicitation module in KEAR first derives different types of stance knowledge from LLM's reasoning process. Then, the knowledge retrieval module in KEAR matches the target language input to the most relevant stance knowledge for enhancing text representations. Experiments on multilingual datasets show the effectiveness of KEAR compared with competitive baselines as well as the CLSD approaches trained with labeled data in target language.

(Nov 13): 7:45:45 (Morning) - Gather

Decoding Matters: Addressing Amplification Bias and Homogeneity Issue in Recommendations for Large Language Models

Kegen Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, Fuli Feng

Adapting Large Language Models (LLMs) for recommendation requires careful consideration of the decoding process, given the inherent differences between generating items and natural language. Existing approaches often directly apply LLMs' original decoding methods. However, we find these methods encounter significant challenges: 1) amplification bias where standard length normalization inflates scores for items containing tokens with generation probabilities close to 1 (termed ghost tokens), and 2) homogeneity issue generating multiple similar or repetitive items for a user. To tackle these challenges, we introduce a new decoding approach named Debiasing-Diversifying Decoding (D^3). D^3 disables length normalization for ghost tokens to alleviate amplification bias, and it incorporates a text-free assistant model to encourage tokens less frequently generated by LLMs for counteracting recommendation homogeneity. Extensive experiments on real-world datasets demonstrate the method's effectiveness in enhancing accuracy and diversity.

(Nov 13): 7:45:45 (Morning) - Gather

Enhancing Legal Case Retrieval via Scaling High-quality Synthetic Query-Candidate Pairs

Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huijin Chen, Zhiyuan Liu, Maosong Sun

Legal case retrieval (LCR) aims to provide similar cases as references for a given fact description. This task is crucial for promoting consistent judgments in similar cases, effectively enhancing judicial fairness and improving work efficiency for judges. However, existing works face two main challenges for real-world applications: existing works mainly focus on case-to-case retrieval using lengthy queries, which does not match real-world scenarios; and the limited data scale, with current datasets containing only hundreds of queries, is insufficient to satisfy the training requirements of existing data-hungry neural models. To address these issues, we introduce an automated method to construct synthetic query-candidate pairs and build the largest LCR dataset to date, LEAD, which is hundreds of times larger than existing datasets. This data construction method can provide ample training signals for LCR models. Experimental results demonstrate that model training with our constructed data can achieve state-of-the-art results on two widely-used LCR benchmarks. Besides, the construction method can also be applied to civil cases and achieve promising results. The data and codes can be found in <https://github.com/thunlp/LEAD>.

Interpretability and Analysis of Models for NLP

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Embedding and Gradient Say Wrong: A White-Box Method for Hallucination Detection

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, Junbo Zhao

In recent years, large language models (LLMs) have achieved remarkable success in the field of natural language generation. Compared to previous small-scale models, they are capable of generating fluent output based on the provided prefix or prompt. However, one critical challenge — the "hallucination" problem — remains to be resolved. Generally, the community refers to the undetected hallucination scenario where the LLMs generate text unrelated to the input text or facts. In this study, we intend to model the distributional distance between the regular conditional output and the unconditional output, which is generated without a given input text. Based upon Taylor Expansion for this distance at the output probability space, our approach manages to leverage the embedding and first-order gradient information. The resulting approach is plug-and-play that can be easily adapted to any autoregressive LLM. On the hallucination benchmarks HADES and other datasets, our approach achieves state-of-the-art performance.

(Nov 13): 7:45:45 (Morning) - Gather

Unveiling the Lexical Sensitivity of LLMs: Combinatorial Optimization for Prompt Enhancement

Pengwei Zhan, Zhen Xu, Qian Tan, Jie Song, Ru Xie

Large language models (LLMs) demonstrate exceptional instruct-following ability to complete various downstream tasks. Although this impressive ability makes LLMs flexible task solvers, their performance in solving tasks also heavily relies on instructions. In this paper, we reveal that LLMs are over-sensitive to lexical variations in task instructions, even when the variations are imperceptible to humans. By providing models with neighborhood instructions, which are closely situated in the latent representation space and differ by only one semantically similar word, the performance on downstream tasks can be vastly different. Following this property, we propose a black-box Combinatorial Optimization framework for Prompt Lexical Enhancement (COPE). COPE performs iterative lexical optimization according to the feedback from a batch of proxy tasks, using a search strategy related to word influence. Experiments show that even widely-used human-crafted prompts for current benchmarks suffer from the lexical sensitivity of models, and COPE recovers the declined model ability in both instruction-following and solving downstream tasks.

(Nov 13): 7:45:45 (Morning) - Gather

Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method

Weichao Zhang, Ruqiang Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, Xueqi Cheng

As the scale of training corpora for large language models (LLMs) grows, model developers become increasingly reluctant to disclose details on their data. This lack of transparency poses challenges to scientific evaluation and ethical deployment. Recently, pretraining data detection approaches, which infer whether a given text was part of an LLM's training data through black-box access, have been explored. The Min-K% Prob method, which has achieved state-of-the-art results, assumes that a non-training example tends to contain a few outlier words with

low token probabilities. However, the effectiveness may be limited as it tends to misclassify non-training texts that contain many common words with high probabilities predicted by LLMs. To address this issue, we introduce a divergence-based calibration method, inspired by the divergence-from-randomness concept, to calibrate token probabilities for pretraining data detection. We compute the cross-entropy (i.e., the divergence) between the token probability distribution and the token frequency distribution to derive a detection score. We have developed a Chinese-language benchmark, PatentMIA, to assess the performance of detection approaches for LLMs on Chinese text. Experimental results on English-language benchmarks and PatentMIA demonstrate that our proposed method significantly outperforms existing methods. Our code and PatentMIA benchmark are available at <https://github.com/zhang-wei-chao/DC-PDD>.

(Nov 13): 7:458:45 (Morning) - Gather

Revealing the Parallel Multilingual Learning within Large Language Models

Yongyu Mu, Peinan Feng, Zhiqian Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, JingBo Zhu

Large language models (LLMs) can handle multilingual and cross-lingual text within a single input; however, previous works leveraging multilingualism in LLMs primarily focus on using English as the pivot language to enhance language understanding and reasoning. Given that multiple languages are a compensation for the losses caused by a single language's limitations, it's a natural next step to enrich the models learning context through the integration of the original input with its multiple translations. In this paper, we start by revealing that LLMs learn from parallel multilingual input (PMI). Our comprehensive evaluation shows that PMI enhances the model's comprehension of the input, achieving superior performance than conventional in-context learning (ICL). Furthermore, to explore how multilingual processing affects prediction, we examine the activated neurons in LLMs. Surprisingly, involving more languages in the input activates fewer neurons, leading to more focused and effective neural activation patterns. Also, this neural reaction coincidentally mirrors the neuroscience insight about synaptic pruning, highlighting a similarity between artificial and biological 'brains'.

(Nov 13): 7:458:45 (Morning) - Gather

MARE: Multi-Aspect Rationale Extractor on Unsupervised Rationale Extraction

Han Jiang, Junwen Duan, Zhe Qiu, Jianxin Wang

Unsupervised rationale extraction aims to extract text snippets to support model predictions without explicit rationale annotation. Researchers have made many efforts to solve this task. Previous works often encode each aspect independently, which may limit their ability to capture meaningful internal correlations between aspects. While there has been significant work on mitigating spurious correlations, our approach focuses on leveraging the beneficial internal correlations to improve multi-aspect rationale extraction. In this paper, we propose a Multi-Aspect Rationale Extractor (MARE) to explain and predict multiple aspects simultaneously. Concretely, we propose a Multi-Aspect Multi-Head Attention (MAMHA) mechanism based on hard deletion to encode multiple text chunks simultaneously. Furthermore, multiple special tokens are prepended in front of the text with each corresponding to one certain aspect. Finally, multi-task training is deployed to reduce the training overhead. Experimental results on two unsupervised rationale extraction benchmarks show that MARE achieves state-of-the-art performance. Ablation studies further demonstrate the effectiveness of our method. Our codes have been available at <https://github.com/CSU-NLP-Group/MARE>.

(Nov 13): 7:458:45 (Morning) - Gather

Leveraging Estimated Transferability Over Human Intuition for Model Selection in Text Ranking

Jun Bai, Zhiufan Chen, Zhenzi Li, Hanhua Hong, Jianfei Zhang, Chen Li, Chenghua Lin, Wenge Rong

Text ranking has witnessed significant advancements, attributed to the utilization of dual-encoder enhanced by Pre-trained Language Models (PLMs). Given the proliferation of available PLMs, selecting the most effective one for a given dataset has become a non-trivial challenge. As a promising alternative to human intuition and brute-force fine-tuning, Transferability Estimation (TE) has emerged as an effective approach to model selection. However, current TE methods are primarily designed for classification tasks, and their estimated transferability may not align well with the objectives of text ranking. To address this challenge, we propose to compute the expected rank as transferability, explicitly reflecting the model's ranking capability. Furthermore, to mitigate anisotropy and incorporate training dynamics, we adaptively scale isotropic sentence embeddings to yield an accurate expected rank score. Our resulting method, Adaptive Ranking Transferability (AiTran), can effectively capture subtle differences between models. On challenging model selection scenarios across various text ranking datasets, it demonstrates significant improvements over previous classification-oriented TE methods, human intuition, and ChatGPT with minor time consumption.

(Nov 13): 7:458:45 (Morning) - Gather

Enhancing Training Data Attribution for Large Language Models with Fitting Error Consideration

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng

The black-box nature of large language models (LLMs) poses challenges in interpreting results, impacting issues such as data intellectual property protection and hallucination tracing. Training data attribution (TDA) methods are considered effective solutions to address these challenges. Most recent TDA methods rely on influence functions, assuming the model achieves minimized empirical risk. However, achieving this criterion is difficult, and sourcing accuracy can be compromised by fitting errors during model training. In this paper, we introduce a novel TDA method called Debias and Denoise Attribution (DDA), which enhances influence functions by addressing fitting errors. Specifically, the debias strategy seeks to improve the performance of influence functions by eliminating the knowledge bias present in the base model before fine-tuning, while the denoise strategy aims to reduce discrepancies in influence scores arising from varying degrees of fitting during the training process through smoothing techniques. Experimental results demonstrate that our method significantly outperforms existing approaches, achieving an averaged AUC of 91.64%. Moreover, DDA exhibits strong generality and scalability across various sources and different-scale models like LLaMA2, Qwen2, and Mistral.

(Nov 13): 7:458:45 (Morning) - Gather

RepMatch: Quantifying Cross-Instance Similarities in Representation Space

Mohammad Reza Modares, Sina Abbasi, Mohammad Taher Pilehvar

Advances in dataset analysis techniques have enabled more sophisticated approaches to analyzing and characterizing training data instances, often categorizing data based on attributes such as "difficulty". In this work, we introduce RepMatch, a novel method that characterizes data through the lens of similarity. RepMatch quantifies the similarity between subsets of training instances by comparing the knowledge encoded in models trained on them, overcoming the limitations of existing analysis methods that focus solely on individual instances and are restricted to within-dataset analysis. Our framework allows for a broader evaluation, enabling similarity comparisons across arbitrary subsets of instances, supporting both dataset-to-dataset and instance-to-dataset analyses. We validate the effectiveness of RepMatch across multiple NLP tasks, datasets, and models. Through extensive experimentation, we demonstrate that RepMatch can effectively compare datasets, identify more representative subsets of a dataset (that lead to better performance than randomly selected subsets of equivalent size), and uncover heuristics underlying the construction of some challenge datasets.

(Nov 13): 7:458:45 (Morning) - Gather

Knowledge Graph Enhanced Large Language Model Editing

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, Zhumin Chen

Large language models (LLMs) are pivotal in advancing natural language processing (NLP) tasks, yet their efficacy is hampered by inaccuracies and outdated knowledge. Model editing emerges as a promising solution to address these challenges. However, existing editing methods struggle to track and incorporate changes in knowledge associated with edits, which limits the generalization ability of post-edit LLMs in processing edited knowledge. To tackle these problems, we propose a novel model editing method that leverages knowledge graphs for enhancing LLM editing, namely GLAME. Specifically, we first utilize a knowledge graph augmentation module to uncover associated knowledge that has changed due to editing, obtaining its internal representations within LLMs. This approach allows knowledge alterations within LLMs to be reflected through an external graph structure. Subsequently, we design a graph-based knowledge edit module to integrate structured knowledge into the model editing. This ensures that the updated parameters reflect not only the modifications of the edited knowledge but also the changes in other associated knowledge resulting from the editing process. Comprehensive experiments conducted on GPT-J and GPT-2 XL demonstrate that GLAME significantly improves the generalization capabilities of post-edit LLMs in employing edited knowledge.

(Nov 13): 7:45:45 (Morning) - Gather

Transfer Learning for Text Classification via Model Risk Analysis

Yujie Sun, Chuyi Fan, Qun Chen

It has been well recognized that text classification can be satisfactorily performed by Deep Neural Network (DNN) models, provided that there are sufficient in-distribution training data. However, in the presence of distribution drift, a well trained DNN model may not perform well on a new dataset even though class labels are aligned between training and target datasets. To alleviate this limitation, we propose a novel approach based on model risk analysis to adapt a pre-trained DNN model towards a new dataset given only a small set of representative data. We first present a solution of model risk analysis for text classification, which can effectively quantify misprediction risk of a classifier on a dataset. Built upon the existing framework of LearnRisk, the proposed solution, denoted by LearnRisk-TC, first generates interpretable risk features, then constructs a risk model by aggregating these features, and finally trains the risk model on a small set of labeled data. Furthermore, we present a transfer learning solution based on model risk analysis, which can effectively fine-tune a pre-trained model toward a target dataset by minimizing its misprediction risk. We have conducted extensive experiments on real datasets. Our experimental results show that the proposed solution performs considerably better than the existing alternative approaches. By using text classification as a test case, we demonstrate the potential applicability of risk-based transfer learning to various challenging NLP tasks. Our codes are available at <https://github.com/sjycomputer/LRTC>.

(Nov 13): 7:45:45 (Morning) - Gather

Deciphering the Interplay of Parametric and Non-Parametric Memory in RAG Models

Mehrdad Farahani, Richard Johansson

Generative language models often struggle with specialized or less-discussed knowledge. A potential solution is found in Retrieval-Augmented Generation (RAG) models which act like retrieving information before generating responses. In this study, we explore how the ATLAS approach, a RAG model, decides between what it already knows (parametric) and what it retrieves (non-parametric). We use causal mediation analysis and controlled experiments to examine how internal representations influence information processing. Our findings disentangle the effects of parametric knowledge and the retrieved context. They indicate that in cases where the model can choose between both types of information (parametric and non-parametric), it relies more on the context than the parametric knowledge. Furthermore, the analysis investigates the computations involved in how the model uses the information from the context. We find that multiple mechanisms are active within the model and can be detected with mediation analysis: first, the decision of *whether the context is relevant*, and second, how the encoder computes output representations to support copying when relevant.

(Nov 13): 7:45:45 (Morning) - Gather

Exploring Reward Model Strength's Impact on Language Models

Yanyan Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, Xiaoyu Shen

Reinforcement Learning from Human Feedback significantly enhances Natural Language Processing by aligning language models with human expectations. A critical factor in this alignment is the strength of reward models used during training. This study explores whether stronger reward models invariably lead to better language models. In this paper, through experiments on relevance, factuality, and completeness tasks using the QA-FEEDBACK dataset and reward models based on Longformer, we uncover a surprising paradox: language models trained with moderately accurate reward models outperform those guided by highly accurate ones. This challenges the widely held belief that stronger reward models always lead to better language models, and opens up new avenues for future research into the key factors driving model performance and how to choose the most suitable reward models.

(Nov 13): 7:45:45 (Morning) - Gather

On the In-context Generation of Language Models

Zhongtao Jiang, Yuanzhe Zhang, Kun Luo, Xiaowei Yuan, Jun Zhao, Kang Liu

Large language models (LLMs) are found to have the ability of in-context generation (ICG): when they are fed with an in-context prompt concatenating a few somehow similar examples, they can implicitly recognize the pattern of them and then complete the prompt in the same pattern. ICG is curious, since language models are usually not explicitly trained in the same way as the in-context prompt, and the distribution of examples in the prompt differs from that of sequences in the pretrained corpora. This paper provides a systematic study of the ICG ability of language models, covering discussions about its source and influential factors, in the view of both theory and empirical experiments. Concretely, we first propose a plausible latent variable model to model the distribution of the pretrained corpora, and then formalize ICG as a problem of next topic prediction. With this framework, we can prove that the repetition nature of a few topics ensures the ICG ability on them theoretically. Then, we use this controllable pretrained distribution to generate several medium-scale synthetic datasets (token scale: 2.1B-3.9B) and experiment with different settings of Transformer architectures (parameter scale: 4M-234M). Our experimental results further offer insights into how the data and model architectures influence ICG.

(Nov 13): 7:45:45 (Morning) - Gather

Unsupervised Hierarchical Topic Modeling via Anchor Word Clustering and Path Guidance

Jiyuan Liu, Hegang Chen, Chunjiang Zhu, Yanghui Rao

Hierarchical topic models nowadays tend to capture the relationship between words and topics, often ignoring the role of anchor words that guide text generation. For the first time, we detect and add anchor words to the text generation process in an unsupervised way. Firstly, we adopt a clustering algorithm to adaptively detect anchor words that are highly consistent with every topic, which forms the path of topic \rightarrow anchor word. Secondly, we add the causal path of anchor word \rightarrow word to the popular Variational Auto-Encoder (VAE) framework via implicitly using word co-occurrence graphs. We develop the causal path of topic+anchor word \rightarrow higher-layer topic that aids the expression of topic concepts with anchor words to capture a more semantically tight hierarchical topic structure. Finally, we enhance the model's representation of the anchor words through a novel contrastive learning. After jointly training the aforementioned constraint objectives, we can produce more coherent and diverse topics with a better hierarchical structure. Extensive experiments on three datasets show that our model outperforms state-of-the-art methods.

(Nov 13): 7:458:45 (Morning) - Gather

Limited Out-of-Context Knowledge Reasoning in Large Language Models

Peng Hu, Changjiang Gao, Ruiqi Gao, Jiajun Chen, Shujian Huang

Large Language Models (LLMs) possess extensive knowledge and strong capabilities in performing in-context reasoning. However, previous work challenges their out-of-context reasoning ability, i.e., the ability to infer information from their training data, instead of from the context or prompt. This paper focuses on a significant aspect of out-of-context reasoning: Out-of-Context Knowledge Reasoning (OCCR), which is to combine multiple knowledge to infer new knowledge. We designed a synthetic dataset with seven representative OCCR tasks to systematically assess the OCCR capabilities of LLMs. Using this dataset, we evaluated several LLMs and discovered that their proficiency in this aspect is limited, regardless of whether the knowledge is trained in a separate or adjacent training settings. Moreover, training the model to reason with reasoning examples does not result in significant improvement, while training the model to perform explicit knowledge retrieval helps for retrieving attribute knowledge but not the relation knowledge, indicating that the model's limited OCCR capabilities are due to difficulties in knowledge retrieval. Furthermore, we treat cross-lingual knowledge transfer as a distinct form of OCCR, and evaluate this ability. Our results show that the evaluated model also exhibits limited ability in transferring knowledge across languages.

(Nov 13): 7:458:45 (Morning) - Gather

Exploring Intra and Inter-language Consistency in Embeddings with ICA

Rongchi Li, Takeru Matsuda, Hitomi Yanaka

Word embeddings represent words as multidimensional real vectors, facilitating data analysis and processing, but are often challenging to interpret. Independent Component Analysis (ICA) creates clearer semantic axes by identifying independent key features. Previous research has shown ICA's potential to reveal universal semantic axes across languages. However, it lacked verification of the consistency of independent components within and across languages. We investigated the consistency of semantic axes in two ways: both within a single language and across multiple languages. We first probed into intra-language consistency, focusing on the reproducibility of axes by performing ICA multiple times and clustering the outcomes. Then, we statistically examined inter-language consistency by verifying those axes' correspondences using statistical tests. We newly applied statistical methods to establish a robust framework that ensures the reliability and universality of semantic axes.

(Nov 13): 7:458:45 (Morning) - Gather

What Matters in Learning Facts in Language Models? Multifaceted Knowledge Probing with Diverse Multi-Prompt Datasets

Xin Zhao, Naoki Yoshinaga, Daisuke Oba

Language models often struggle with handling factual knowledge, exhibiting factual hallucination issue. This makes it vital to evaluate the models' ability to recall its parametric knowledge about facts. In this study, we introduce a knowledge probing benchmark, BELIEF(ICL), to evaluate the knowledge recall ability of both encoder- and decoder-based pre-trained language models (PLMs) from diverse perspectives. BELIEFs utilize a multi-prompt dataset to evaluate PLMs' accuracy, consistency, and reliability in factual knowledge recall. To enable a more reliable evaluation with BELIEFs, we semi-automatically create MyriadLAMA, which has massively diverse prompts. We validate the effectiveness of BELIEFs in comprehensively evaluating PLMs' knowledge recall ability on diverse PLMs, including recent large language models (LLMs). We then investigate key factors in memorizing and recalling facts in PLMs, such as model size, pretraining strategy and corpora, instruction-tuning process and in-context learning settings. Finally, we reveal the limitation of the prompt-based knowledge probing. The MyriadLAMA is publicized.

Language Modeling

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

AMR-Evol: Adaptive Modular Response Evolution Elicits Better Knowledge Distillation for Large Language Models in Code Generation

Ziyang Luo, Xin Li, Hongzhan Lin, Jing Ma, Lidong Bing

The impressive performance of proprietary LLMs like GPT4 in code generation has led to a trend to replicate these capabilities in open-source models through knowledge distillation (e.g. Code Evol-Instruct). However, these efforts often neglect the crucial aspect of response quality, relying heavily on teacher models for direct response distillation. This paradigm, especially for complex instructions, can degrade the quality of synthesized data, compromising the knowledge distillation process. To this end, our study introduces the Adaptive Modular Response Evolution (AMR-Evol) framework, which employs a two-stage process to refine response distillation. The first stage, modular decomposition, breaks down the direct response into more manageable sub-modules. The second stage, adaptive response evolution, automatically evolves the response with the related function modules. Our experiments with three popular code benchmarks HumanEval, MBPP, and EvalPlus+ tests to the superiority of the AMR-Evol framework over baseline response distillation methods. By comparing with the open-source Code LLMs trained on a similar scale of data, we observed performance enhancements: more than +3.0 points on HumanEval-Plus and +1.0 points on MBPP-Plus, which underscores the effectiveness of our framework. Our codes are available at <https://github.com/ChiYeungLaw/AMR-Evol>.

(Nov 13): 7:458:45 (Morning) - Gather

Towards Tool Use Alignment of Large Language Models

Zhi-Yuan Chen, Shiqi Shen, Guangyao Shen, Gong Zhi, Xu Chen, Yankai Lin

Recently, tool use with LLMs has become one of the primary research topics as it can help LLM generate truthful and helpful responses. Existing studies on tool use with LLMs primarily focus on enhancing the tool-calling ability of LLMs. In practice, like chat assistants, LLMs are also required to align with human values in the context of tool use. Specifically, LLMs should refuse to answer unsafe tool use relevant instructions and insecure tool responses to ensure their reliability and harmlessness. At the same time, LLMs should demonstrate autonomy in tool use to reduce the costs associated with tool calling. To tackle this issue, we first introduce the principle that LLMs should follow in tool use scenarios: H2A. The goal of H2A is to align LLMs with **helpfulness**, **harmlessness**, and **autonomy**. In addition, we propose ToolAlign, a dataset comprising instruction-tuning data and preference data to align LLMs with the H2A principle for tool use. Based on ToolAlign, we develop LLMs by supervised fine-tuning and preference learning, and experimental results demonstrate that the LLMs exhibit remarkable tool-calling capabilities, while also refusing to engage with harmful content, and displaying a high degree of autonomy in tool utilization. The code and datasets are available at: <https://github.com/zhiyuanc2001/ToolAlign>.

(Nov 13): 7:458:45 (Morning) - Gather

Knowledge Verification to Nip Hallucination in the Bud

Fangji Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, Shuming Shi

While large language models (LLMs) have demonstrated exceptional performance across various tasks following human alignment, they

may still generate responses that sound plausible but contradict factual knowledge, a phenomenon known as hallucination. In this paper, we demonstrate the feasibility of mitigating hallucinations by verifying and minimizing the inconsistency between external knowledge present in the alignment data and the intrinsic knowledge embedded within foundation LLMs. Specifically, we propose a novel approach called Knowledge Consistent Alignment (KCA), which employs a well-aligned LLM to automatically formulate assessments based on external knowledge to evaluate the knowledge boundaries of foundation LLMs. To address knowledge inconsistencies in the alignment data, KCA implements several specific strategies to deal with these data instances. We demonstrate the superior efficacy of KCA in reducing hallucinations across six benchmarks, utilizing foundation LLMs of varying backbones and scales. This confirms the effectiveness of mitigating hallucinations by reducing knowledge inconsistency. Our code, model weights, and data are openly accessible at <https://github.com/fanqiwan/KCA>.

(Nov 13): 7:45:45 (Morning) - Gather

Retrieved In-Context Principles from Previous Mistakes

Hao Sun, Yong Jiang, Bo Wang, Yingyan Hou, Yan Zhang, Pengjun Xie, Fei Huang

In-context learning (ICL) has been instrumental in adapting large language models (LLMs) to downstream tasks using correct input-output examples. Recent advances have attempted to improve model performance through principles derived from mistakes, yet these approaches suffer from lack of customization and inadequate error coverage. To address these limitations, we propose Retrieved In-Context Principles (RICP), a novel teacher-student framework. In RICP, the teacher model analyzes mistakes from the student model to generate reasons and insights for preventing similar mistakes. These mistakes are clustered based on their underlying reasons for developing task-level principles, enhancing the error coverage of principles. During inference, the most relevant mistakes for each question are retrieved to create question-level principles, improving the customization of the provided guidance. RICP is orthogonal to existing prompting methods and does not require intervention from the teacher model during inference. Experimental results across seven reasoning benchmarks reveal that RICP effectively enhances performance when applied to various prompting strategies.

(Nov 13): 7:45:45 (Morning) - Gather

KNN-Instruct: Automatic Instruction Construction with K Nearest Neighbor Deduction

Jianshang Kou, Benfeng Xu, Chiwei Zhu, Zhendong Mao

Supervised fine-tuning (SFT) is a critical procedure for aligning large language models. Despite its efficiency, the construction of SFT data often struggles with issues of quality, diversity, and scalability. Many existing methods, inspired by the Self-Instruct framework, typically generate synthetic instructions by prompting aligned proprietary models like ChatGPT. However, such process suffers from stale distribution, resulting in instructions that are merely trivial variations of existing ones. In this paper, we introduce a novel bootstrapping approach termed KNN-Instruct, which incorporates KNN deduction to produce meaningful new instructions by effectively summarizing and learning from similar existing ones. We conduct an economical controlled experiment to preliminarily validate its effectiveness. In the further experiment, we construct a high-quality SFT dataset named KNN-Inst-12k*. Applying the dataset to 7B, we get a MT-Bench score of 7.64, which outperforms all 7B models on the LMSYS leaderboard, including Starling-LM-7B (7.48), OpenChat-3.5 (7.06) and Zephyr-7B-beta (6.53). Our code and data are available at <https://github.com/CrossmodalGroup/KNN-Instruct/>.

(Nov 13): 7:45:45 (Morning) - Gather

InferAligner: Inference-Time Alignment for Harmlessness through Cross-Model Guidance

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, Xipeng Qiu

As large language models (LLMs) rapidly evolve, they are increasingly being customized through fine-tuning to suit the specific needs of various applications. A critical aspect of this advancement is the alignment process, which ensures that these models perform tasks in ways that align with human values and expectations. Current alignment methods, such as direct preference optimization (DPO) and reinforcement learning from human feedback (RLHF), focus primarily on alignment during training phase. However, these methods often involve complex and resource-intensive training processes, posing significant challenge for their implementation. Therefore, we propose **InferAligner**, a simple yet effective method for harmlessness alignment during inference phase. InferAligner decouples harmlessness from helpfulness. During the training phase, it focuses solely on enhancing the target model's capabilities for downstream tasks. In the inference phase, it utilizes safety steering vectors extracted from the aligned model to guide the target model towards harmlessness alignment. Experimental results show that our method can be very effectively applied to domain-specific models in finance, medicine, and mathematics, as well as to multimodal large language models (MLLMs) such as LLaVA. It significantly diminishes the attack success rate (ASR) of both harmful instructions and jailbreak instructions, while maintaining almost unchanged performance in downstream tasks.

(Nov 13): 7:45:45 (Morning) - Gather

Firs Heuristic Then Rational: Dynamic Use of Heuristics in Language Model Reasoning

Yoichi Aoki, Keito Kudo, Tatsuki Kuribayashi, Shusaku Sone, Masaya Taniguchi, Keisuke Sakaguchi, Kentaro Inui

Explicit multi-step reasoning, such as chain-of-thought, is widely adopted in the community to explore the better performance of language models (LMs). We report on the systematic strategy that LMs use in this process. Our controlled experiments reveal that LMs rely more heavily on heuristics, such as lexical overlap, in the earlier stages of reasoning when more steps are required to reach an answer. Conversely, their reliance on heuristics decreases as LMs progress closer to the final answer. This suggests that LMs track only a limited number of future steps and dynamically combine heuristic strategies with rational ones in solving tasks involving multi-step reasoning.

(Nov 13): 7:45:45 (Morning) - Gather

Re-Reading Improves Reasoning in Large Language Models

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, Shuai Ma

To enhance the reasoning capabilities of off-the-shelf Large Language Models (LLMs), we introduce a simple, yet general and effective prompting method, RE2, i.e., Re-Reading the question as input. Unlike most thought-eliciting prompting methods, such as Chain-of-Thought (CoT), which aim to elicit the reasoning process in the output, RE2 shifts the focus to the input by processing questions twice, thereby enhancing the understanding process. Consequently, RE2 demonstrates strong generality and compatibility with most thought-eliciting prompting methods, including CoT. Crucially, RE2 facilitates a "bidirectional" encoding in unidirectional decoder-only LLMs because the first pass could provide global information for the second pass. We begin with a preliminary empirical study as the foundation of RE2, illustrating its potential to enable "bidirectional" attention mechanisms. We then evaluate RE2 on extensive reasoning benchmarks across 14 datasets, spanning 112 experiments, to validate its effectiveness and generality. Our findings indicate that, with the exception of a few scenarios on vanilla ChatGPT, RE2 consistently enhances the reasoning performance of LLMs through a simple re-reading strategy. Further analyses reveal RE2's adaptability, showing how it can be effectively integrated with different LLMs, thought-eliciting prompting, and ensemble strategies.

(Nov 13): 7:45:45 (Morning) - Gather

Zero-Shot Detection of LLM-Generated Text using Token Cohesiveness

Shixuan Ma, Quan Wang

The increasing capability and widespread usage of large language models (LLMs) highlight the desirability of automatic detection of LLM-generated text. Zero-shot detectors, due to their training-free nature, have received considerable attention and notable success. In this paper, we identify a new feature, token cohesiveness, that is useful for zero-shot detection, and we demonstrate that LLM-generated text tends to

exhibit higher token cohesiveness than human-written text. Based on this observation, we devise TOCSIN, a generic dual-channel detection paradigm that uses token cohesiveness as a plug-and-play module to improve existing zero-shot detectors. To calculate token cohesiveness, TOCSIN only requires a few rounds of random token deletion and semantic difference measurement, making it particularly suitable for a practical black-box setting where the source model used for generation is not accessible. Extensive experiments with four state-of-the-art base detectors on various datasets, source models, and evaluation settings demonstrate the effectiveness and generality of the proposed approach. Code available at: <https://github.com/Shixuan-Ma/TOCSIN>.

(Nov 13): 7:458:45 (Morning) - Gather

Safety Arithmetic: A Framework for Test-time Safety Alignment of Language Models by Steering Parameters and Activations

Rima Hazra, Sayan Layek, Somnath Banerjee, Soujanya Poria

Ensuring the safe alignment of large language models (LLMs) with human values is critical as they become integral to applications like translation and question answering. Current alignment methods struggle with dynamic user intentions and complex objectives, making models vulnerable to generating harmful content. We propose Safety Arithmetic, a training-free framework enhancing LLM safety across different scenarios: Base models, Supervised fine-tuned models (SFT), and Edited models. Safety Arithmetic involves Harm Direction Removal to avoid harmful content and Safety Alignment to promote safe responses. Additionally, we present NoInvertEdit, a dataset highlighting edit instances that could compromise model safety if used unintentionally. Our experiments show that Safety Arithmetic significantly improves safety measures, reduces over-safety, and maintains model utility, outperforming existing methods in ensuring safe content generation.

(Nov 13): 7:458:45 (Morning) - Gather

Improving Referring Ability for Biomedical Language Models

Junfeng Jiang, Fei Cheng, Akiko Aizawa

Existing auto-regressive large language models (LLMs) are primarily trained using documents from general domains. In the biomedical domain, continual pre-training is a prevalent method for domain adaptation to inject professional knowledge into powerful LLMs that have been pre-trained in general domains. Previous studies typically conduct standard pre-training by randomly packing multiple documents into a long pre-training sequence. Recently, some existing works suggest that enhancing the relatedness of documents within the same pre-training sequence may be advantageous. However, these studies primarily focus on general domains, which cannot be readily applied in the biomedical domain where the distinction of fine-grained topics is harder. Is it possible to further improve the pre-training for biomedical language models (LMs) using exactly the same corpus? In this paper, we explore an improved approach to continual pre-training, which is a prevalent method for domain adaptation, by utilizing information from the citation network in this challenging scenario. Empirical studies demonstrate that our proposed LinkLM data improves both the intra-sample and inter-sample referring abilities of auto-regressive LMs in the biomedical domain, encouraging more profound consideration of task-specific pre-training sequence design for continual pre-training.

(Nov 13): 7:458:45 (Morning) - Gather

TS-Align: A Teacher-Student Collaborative Framework for Scalable Iterative Finetuning of Large Language Models

Chen Zhang, chengguang tang, Dading Chong, Ke Shi, Guohua Tang, Feng Jiang, Haizhou Li

Mainstream approaches to aligning large language models (LLMs) heavily rely on human preference data, particularly when models require periodic updates. The standard process for iterative alignment of LLMs involves collecting new human feedback for each update. However, the data collection process is costly and challenging to scale. To address this issue, we introduce the "TS-Align" framework, which fine-tunes a policy model using pairwise feedback data automatically mined from its outputs. This automatic mining process is efficiently accomplished through the collaboration between a large-scale teacher model and a small-scale student model. The policy fine-tuning process can be iteratively repeated using on-policy generations within our proposed teacher-student collaborative framework. Through extensive experiments, we demonstrate that our final aligned policy outperforms the base policy model with an average win rate of 69.7% across seven conversational or instruction-following datasets. Furthermore, we show that the ranking capability of the teacher is effectively distilled into the student through our pipeline, resulting in a small-scale yet effective reward model for policy model alignment.

(Nov 13): 7:458:45 (Morning) - Gather

Dual Modalities of Text: Visual and Textual Generative Pre-Training

Yekun Chai, Qingyi Liu, Jingwu Xiao, Shuhuan Wang, Yu Sun, Hua Wu

The integration of visual and textual information represents a promising direction in the advancement of language models. In this paper, we explore the dual modality of language both visual and textual within an autoregressive framework, pre-trained on both document images and texts. Our method employs a multimodal training strategy, utilizing visual data through next token prediction with a regression head and/or textual data through next token prediction with a classification head. We focus on understanding the interaction between these two modalities and their combined impact on model performance. Our extensive evaluation across a wide range of benchmarks shows that incorporating both visual and textual data significantly improves the performance of pixel-based language models. Remarkably, we find that a unidirectional pixel-based model trained solely on visual data can achieve comparable results to state-of-the-art bidirectional models on several language understanding tasks. This work uncovers the untapped potential of integrating visual and textual modalities for more effective language modeling. We release our code, data, and model checkpoints at <https://github.com/ernie-research/pixelgpt>.

(Nov 13): 7:458:45 (Morning) - Gather

LLaMA-MoE: Building Mixture-of-Experts from LLaMA with Continual Pre-Training

Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, Yu Cheng

Mixture-of-Experts (MoE) has gained increasing popularity as a promising framework for scaling up large language models (LLMs). However, training MoE from scratch in a large-scale setting still suffers from data-hungry and instability problems. Motivated by this limit, we investigate building MoE models from existing dense large language models. Specifically, based on the well-known LLaMA-2 7B model, we obtain an MoE model by: (1) Expert Construction, which partitions the parameters of original Feed-Forward Networks (FFNs) into multiple experts; (2) Continual pre-training, which further trains the transformed MoE model and additional gate networks. In this paper, we comprehensively explore different methods for expert construction and various data sampling strategies for continual pre-training. After these stages, our LLaMA-MoE models could maintain language abilities and route the input tokens to specific experts with part of the parameters activated. Empirically, by training 200B tokens, LLaMA-MoE-3.5B models significantly outperform dense models that contain similar activation parameters.

(Nov 13): 7:458:45 (Morning) - Gather

Scaling Laws for Linear Complexity Language Models

Xuyang Shen, Dong Li, Ruitao Leng, Zhen Qin, Weigao Sun, Yiran Zhong

The interest in linear complexity models for large language models is on the rise, although their scaling capacity remains uncertain. In this study, we present the scaling laws for linear complexity language models to establish a foundation for their scalability. Specifically, we examine the scaling behaviors of three efficient linear architectures. These include TNL, a linear attention model with data-independent decay; HGRN2, a linear RNN with data-dependent decay; and cosFormer2, a linear attention model without decay. We also include LLaMA as a baseline architecture for comparison with softmax attention. These models were trained with six variants, ranging from 70M to 7B parameters.

ters on a 300B-token corpus, and evaluated with a total of 1,376 intermediate checkpoints on various downstream tasks. These tasks include validation loss, commonsense reasoning, and information retrieval and generation. The study reveals that existing linear complexity language models exhibit similar scaling capabilities as conventional transformer-based models while also demonstrating superior linguistic proficiency and knowledge retention.

(Nov 13): 7:45:45 (Morning) - Gather

Tokenization Falling Short: The Curse of Tokenization

Yekun Chai, Yewei Fang, Qiwei Peng, Xuhong Li

Language models typically tokenize raw text into sequences of subword identifiers from a predefined vocabulary, a process inherently sensitive to typographical errors, length variations, and largely oblivious to the internal structure of tokens. We term "the curse of tokenization". In this study, we delve into these drawbacks and demonstrate that large language models (LLMs) remain susceptible to these problems. This study systematically investigates these challenges and their impact on LLMs through three critical research questions: (1) complex problem solving, (2) token structure probing, and (3) resilience to typographical variation. Our findings reveal that scaling model parameters can mitigate the issue of tokenization; however, LLMs still suffer from biases induced by typos and other text format variations. Our experiments show that subword regularization such as BPE-dropout can mitigate this issue. We release our evaluation code and data at <https://github.com/FloatAI/TKEval>.

(Nov 13): 7:45:45 (Morning) - Gather

XLLaMA2: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, Fei Yuan

Large Language Models (LLMs) demonstrate remarkable translation capabilities in high-resource language tasks, yet their performance in low-resource languages is hindered by insufficient multilingual data during pre-training. To address this, we conduct extensive multilingual continual pre-training on the LLaMA series models, enabling translation support across more than 100 languages. Through a comprehensive analysis of training strategies, such as vocabulary expansion and data augmentation, we develop LLaMAX. Remarkably, without sacrificing its generalization ability, LLaMAX achieves significantly higher translation performance compared to existing open-source LLMs (by more than 10 spBLEU points) and performs on-par with specialized translation model (M2M-100-12B) on the Flores-101 benchmark. Extensive experiments indicate that LLaMAX can serve as a robust multilingual foundation model. The code¹⁰ and the models¹¹ are publicly available.

(Nov 13): 7:45:45 (Morning) - Gather

On Training Data Influence of GPT Models

Qingyi Liu, Yekun Chai, Shuohuan Wang, Yu Sun, Qiwei Peng, Hua Wu

Amidst the rapid advancements in generative language models, the investigation of how training data shapes the performance of GPT models is still emerging. This paper presents GPTfluence, a novel approach that leverages a featurized simulation to assess the impact of training examples on the training dynamics of GPT models. Our approach not only traces the influence of individual training instances on performance trajectories, such as loss and other key metrics, on targeted test points, but also enables a comprehensive comparison with existing methods across various training scenarios in GPT models, ranging from 14 million to 2.8 billion parameters, across a range of downstream tasks. Contrary to earlier methods that struggle with generalization to new data, GPTfluence introduces a parameterized simulation of training dynamics, demonstrating robust generalization capabilities to unseen training data. This adaptability is evident across both fine-tuning and instruction-tuning scenarios, spanning tasks in natural language understanding and generation. We make our code and data publicly available at <https://github.com/ernie-research/gptfluence>.

Linguistic Theories, Cognitive Modeling and Psycholinguistics

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

CogGPT: Unleashing the Power of Cognitive Dynamics on Large Language Models

Yaojia Lv, Haojie Pan, Zekun Wang, Jiefeng Liang, Yuanxing Liu, Ruiji Fu, Ming Liu, Zhongyuan Wang, Bing Qin

Cognitive dynamics, which refer to the evolution in human cognitive processes, are pivotal to advance human understanding of the world. Recent advancements in large language models (LLMs) highlight their potential for cognitive simulation. However, these LLM-based cognitive studies primarily focus on replicating human cognition in specific contexts, overlooking the inherently dynamic nature of cognition. To bridge this gap, we explore the cognitive dynamics of LLMs and present a corresponding task inspired by longitudinal studies. Toward the task, we develop CogBench, a novel benchmark to assess the cognitive dynamics of LLMs and validate it through participant surveys. We also design two evaluation metrics for CogBench, including Authenticity and Rationality. Recognizing the inherent static nature of LLMs, we further introduce CogGPT for the task, which features an innovative iterative cognitive mechanism to develop lifelong cognitive dynamics. Empirical results demonstrate the superiority of CogGPT over several existing methods, particularly in its ability to facilitate role-specific cognitive dynamics under continuous information flows. We will release the code and data to enable further research.

(Nov 13): 7:45:45 (Morning) - Gather

Locally Measuring Cross-lingual Lexical Alignment: A Domain and Word Level Perspective

Taelin Karidi, Eitan Grossman, Omri Abend

NLP research on aligning lexical representation spaces to one another has so far focused on aligning language spaces in their entirety. However, cognitive science has long focused on a local perspective, investigating whether translation equivalents truly share the same meaning or the extent that cultural and regional influences result in meaning variations. With recent technological advances and the increasing amounts of available data, the longstanding question of cross-lingual lexical alignment can now be approached in a more data-driven manner. However, developing metrics for the task requires some methodology for comparing metric efficacy. We address this gap and present a methodology for analyzing both synthetic validations and a novel naturalistic validation using lexical gaps in the kinship domain. We further propose new metrics, hitherto unexplored on this task, based on contextualized embeddings. Our analysis spans 16 diverse languages, demonstrating that there is substantial room for improvement with the use of newer language models. Our research paves the way for more accurate and nuanced cross-lingual lexical alignment methodologies and evaluation.

¹⁰<https://github.com/CONE-MT/LLaMAX/>.

¹¹<https://huggingface.co/LLaMAX/>.

(Nov 13): 7:458:45 (Morning) - Gather

Why do objects have many names? A study on word informativeness in language use and lexical systems.

Eleonora Guardoni, Gemma Boleda

Human lexicons contain many different words that speakers can use to refer to the same object, e.g., *purple* or *magenta* for the same shade of color. On the one hand, studies on language use have explored how speakers adapt their referring expressions to successfully communicate in context, without focusing on properties of the lexical system. On the other hand, studies in language evolution have discussed how competing pressures for informativeness and simplicity shape lexical systems, without tackling in-context communication. We aim at bridging the gap between these traditions, and explore why a soft mapping between referents and words is a good solution for communication, by taking into account both in-context communication and the structure of the lexicon. We propose a simple measure of informativeness for words and lexical systems, grounded in a visual space, and analyze color naming data for English and Mandarin Chinese. We conclude that optimal lexical systems are those where multiple words can apply to the same referent, conveying different amounts of information. Such systems allow speakers to maximize communication accuracy and minimize the amount of information they convey when communicating about referents in contexts.

(Nov 13): 7:458:45 (Morning) - Gather

HCEG: Improving the Abstraction Ability of Language Models with Hierarchical Conceptual Entailment Graphs

Juncui Li, Ru Li, Xiaoli Li, Qinghua Chai, Jeff Z. Pan

The abstract inference capability of the Language Model plays a pivotal role in boosting its generalization and reasoning prowess in Natural Language Inference (NLI). Entailment graphs are crafted precisely for this purpose, focusing on learning entailment relations among predicates. Yet, prevailing approaches overlook the *polysemy* and *hierarchical nature of concepts* during entity conceptualization. This oversight disregards how arguments might entail differently across various concept levels, thereby missing potential entailment connections. To tackle this hurdle, we introduce the *concept pyramid* and propose the HiCon-EG (Hierarchical Conceptual Entailment Graph) framework, which organizes arguments hierarchically, delving into entailment relations at diverse concept levels. By learning entailment relationships at different concept levels, the model is guided to better understand concepts so as to improve its abstract inference capabilities. Our method enhances scalability and efficiency in acquiring common-sense knowledge through leveraging statistical language distribution instead of manual labeling. Experimental results show that entailment relations derived from HiCon-EG significantly bolster abstract detection tasks. Our code is available at <https://github.com/SXUCFN/HiCon-EG>

(Nov 13): 7:458:45 (Morning) - Gather

Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood

Yang Xu, Yu Wang, Hao An, Yongyuan Li, Zhichen Liu

Human and model-generated texts can be distinguished by examining the magnitude of likelihood in language. However, it is becoming increasingly difficult as language model's capabilities of generating human-like texts keep evolving. This study provides a new perspective by using the relative likelihood values instead of absolute ones, and extracting useful features from the spectrum-view of likelihood for the human-model text detection task. We propose a detection procedure with two classification methods, supervised and heuristic-based, respectively, which results in competitive performances with previous zero-shot detection methods and a new state-of-the-art on short-text detection. Our method can also reveal subtle differences between human and model languages, which find theoretical roots in psycholinguistics studies.

Low-resource Methods for NLP

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

Clustering and Ranking: Diversity-preserved Instruction Selection through Expert-aligned Quality Estimation

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, shimin tao, Xiaofeng Zhao, Mahongxia, Zhang Li, Boxing Chen, Hao Yang, Bei Li, Tong Xiao, JingBo Zhu

With contributions from the open-source community, a vast amount of instruction tuning (IT) data has emerged. Given the significant resource allocation required by training and evaluating models, it is advantageous to have an efficient method for selecting high-quality IT data. However, existing methods for instruction data selection have limitations such as relying on fragile external APIs, being affected by biases in GPT models, or reducing the diversity of the selected instruction dataset. In this paper, we propose an industrial-friendly, expert-aligned and diversity-preserved instruction data selection method: Clustering and Ranking (CaR). CaR consists of two steps. The first step involves ranking instruction pairs using a scoring model that is well aligned with expert preferences (achieving an accuracy of 84.25%). The second step involves preserving dataset diversity through a clustering process. In our experiment, CaR selected a subset containing only 1.96% of Alpaca's IT data, yet the underlying AlpacaR model trained on this subset outperforms Alpaca by an average of 32.1% in GPT-4 evaluations. Furthermore, our method utilizes small models (550M parameters) and requires only 11.2% of the monetary cost compared to existing methods, making it easily deployable in industrial scenarios.

(Nov 13): 7:458:45 (Morning) - Gather

Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks

Haoyuan Wu, Haisheng Zheng, Zhiwulin He, Bei Yu

Large language models (LLMs) have demonstrated considerable proficiency in general natural language processing (NLP) tasks. Instruction tuning, a successful paradigm, enhances the ability of LLMs to follow natural language instructions and exhibit robust generalization across general tasks. However, these models often encounter performance limitations across multiple tasks due to constrained model capacity. Expanding this capacity during the instruction tuning phase poses significant challenges. To address this issue, we introduce parameter-efficient sparsity crafting (PESC), which crafts dense models into sparse models using the mixture-of-experts (MoE) architecture. PESC integrates adapters into the MoE layers of sparse models, differentiating experts without altering the individual weights within these layers. This method significantly reduces computational costs and GPU memory requirements, facilitating model capacity expansion through a minimal parameter increase when guaranteeing the quality of approximation in function space compared to original sparse upcycling. Our empirical evaluation demonstrates the effectiveness of the PESC method. Using PESC during instruction tuning, our best sparse model outperforms other sparse and dense models and exhibits superior general capabilities compared to GPT-3.5. Our code is available at <https://github.com/wuhy68/Parameter-Efficient-MoE>.

(Nov 13): 7:458:45 (Morning) - Gather

Effective Demonstration Annotation for In-Context Learning via Language Model-Based Determinantal Point Process

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, Yong Jiang

In-context learning (ICL) is a few-shot learning paradigm that involves learning mappings through input-output pairs and appropriately ap-

plying them to new instances. Despite the remarkable ICL capabilities demonstrated by Large Language Models (LLMs), existing works are highly dependent on large-scale labeled support sets, not always feasible in practical scenarios. To refine this approach, we focus primarily on an innovative selective annotation mechanism, which precedes the standard demonstration retrieval. We introduce the Language Model-based Determinant Point Process (LM-DPP) that simultaneously considers the uncertainty and diversity of unlabeled instances for optimal selection. Consequently, this yields a subset for annotation that strikes a trade-off between the two factors. We apply LM-DPP to various language models, including GPT-J, LLaMA, and GPT-3. Experimental results on 9 NLU and 2 Generation datasets demonstrate that LM-DPP can effectively select canonical examples. Further analysis reveals that LLMs benefit most significantly from subsets that are both low uncertainty and high diversity.

(Nov 13): 7:45:45 (Morning) - Gather

QUIK: Towards End-to-end 4-Bit Inference on Generative Large Language Models

Saleh Ashkboos, Ilia Markov, Elias Frantar, Tingxuan Zhong, Xincheng Wang, Jie Ren, Torsten Hoefler, Dan Alistarh

Large Language Models (LLMs) from the GPT family have become extremely popular, leading to a race towards reducing their inference costs to allow for efficient local computation. However, the vast majority of existing work focuses on weight-only quantization, which can reduce runtime costs in the memory-bound one-token-at-a-time generative setting, but does not address costs in compute-bound scenarios, such as batched inference or prompt processing. In this paper, we address the general quantization problem, where *both weights and activations* should be quantized, which leads to computational improvements in general. We show that the majority of inference computations for large generative models can be performed with both weights and activations being cast to 4 bits, while at the same time maintaining good accuracy. We achieve this via a hybrid quantization strategy called QUIK that compresses most of the weights and activations to 4-bit, while keeping a small fraction of “outlier” weights and activations in higher-precision. QUIK is that it is designed with computational efficiency in mind: we provide GPU kernels matching the QUIK format with highly-efficient layer-wise runtimes, which lead to practical end-to-end throughput improvements of up to 3.4x relative to FP16 execution. We provide detailed studies for models from the OPT, LLaMA-2 and Falcon families, as well as a first instance of accurate inference using quantization plus 2:4 sparsity. Anonymized code is available.

(Nov 13): 7:45:45 (Morning) - Gather

Teaching Small Language Models Reasoning through Counterfactual Distillation

FengTao, Yicheng Li, Li Chenglin, Hao Chen, Fei Yu, Yin Zhang

With the rise of large language models (LLMs), many studies are interested in transferring the reasoning capabilities of LLMs to small language models (SLMs). Previous distillation methods usually utilize the capabilities of LLMs to generate chain-of-thought (CoT) samples and teach SLMs via fine-tuning. However, such a standard distillation approach performs poorly when applied to out-of-distribution (OOD) examples, and the diversity of the generated CoT samples is insufficient. In this work, we propose a novel counterfactual distillation framework. Firstly, we leverage LLMs to automatically generate high-quality counterfactual data. Given an input text example, our method generates a counterfactual example that is very similar to the original input, but its task label has been changed to the desired one. Then, we utilize multi-view CoT to enhance the diversity of reasoning samples. Experiments on four NLP benchmarks show that our approach enhances the reasoning capabilities of SLMs and is more robust to OOD data. We also conduct extensive ablations and sample studies to understand the reasoning capabilities of SLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Self-Training for Sample-Efficient Active Learning for Text Classification with Pre-Trained Language Models

Christopher Schröder, Gerhard Heyer

Active learning is an iterative labeling process that is used to obtain a small labeled subset, despite the absence of labeled data, thereby enabling to train a model for supervised tasks such as text classification. While active learning has made considerable progress in recent years due to improvements provided by pre-trained language models, there is untapped potential in the often neglected unlabeled portion of the data, although it is available in considerably larger quantities than the usually small set of labeled data. In this work, we investigate how self-training, a semi-supervised approach that uses a model to obtain pseudo-labels for unlabeled data, can be used to improve the efficiency of active learning for text classification. Building on a comprehensive reproduction of four previous self-training approaches, some of which are evaluated for the first time in the context of active learning or natural language processing, we introduce HAST, a new and effective self-training strategy, which is evaluated on four text classification benchmarks. Our results show that it outperforms the reproduced self-training approaches and reaches classification results comparable to previous experiments for three out of four datasets, using as little as 25% of the data. The code is publicly available at <https://github.com/chschroeder/self-training-for-sample-efficient-active-learning>.

(Nov 13): 7:45:45 (Morning) - Gather

Structured Optimal Brain Pruning for Large Language Models

Jiatieng Wei, Quan Lu, ning Jiang, Sisi Li, Jingyang Xiang, Jun Chen, Yong Liu

The massive parameters and computational demands hinder the widespread application of Large Language Models (LLMs). Network pruning provides a practical solution to this problem. However, existing pruning works for LLMs mainly focus on unstructured pruning or necessitate post-pruning fine-tuning. The former relies on special hardware to accelerate computation, while the latter may need substantial computational resources. In this paper, we introduce a retraining-free structured pruning method called SoBP (Structured Optimal Brain Pruning). It leverages global first-order information to select pruning structures, then refines them with a local greedy approach, and finally adopts module-wise reconstruction to mitigate information loss. We assess the effectiveness of SoBP across 14 models from 3 LLM families on 8 distinct datasets. Experimental results demonstrate that SoBP outperforms current state-of-the-art methods.

(Nov 13): 7:45:45 (Morning) - Gather

Dual-Space Knowledge Distillation for Large Language Models

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, Jinan Xu

Knowledge distillation (KD) is known as a promising solution to compress large language models (LLMs) via transferring their knowledge to smaller models. During this process, white-box KD methods usually minimize the distance between the output distributions of the two models so that more knowledge can be transferred. However, in the current white-box KD framework, the output distributions are from the respective output spaces of the two models, using their own prediction heads. We argue that the space discrepancy will lead to low similarity between the teacher model and the student model on both representation and distribution levels. Furthermore, this discrepancy also hinders the KD process between models with different vocabularies, which is common for current LLMs. To address these issues, we propose a dual-space knowledge distillation (DSKD) framework that unifies the output spaces of the two models for KD. On the basis of DSKD, we further develop a cross-model attention mechanism, which can automatically align the representations of the two models with different vocabularies. Thus, our framework is not only compatible with various distance functions for KD (e.g., KL divergence) like the current framework, but also supports KD between any two LLMs regardless of their vocabularies. Experiments on task-agnostic instruction-following benchmarks show that DSKD significantly outperforms the current white-box KD framework with various distance functions, and also surpasses existing KD methods for LLMs with different vocabularies.

(Nov 13): 7:45:45 (Morning) - Gather

Efficient Active Learning with Adapters

Daria Galimzianova, Leonid Sanovichin

One of the main obstacles for deploying Active Learning (AL) in practical NLP tasks is high computational cost of modern deep learning models. This issue can be partially mitigated by applying lightweight models as an acquisition model, but it can lead to the acquisition-successor mismatch (ASM) problem. Previous works show that the ASM problem can be partially alleviated by using distilled versions of a successor models as acquisition ones. However, distilled versions of pretrained models are not always available. Also, the exact pipeline of model distillation that does not lead to the ASM problem is not clear. To address these issues, we propose to use adapters as an alternative to full fine-tuning for acquisition model training. Since adapters are lightweight, this approach reduces the training cost of the model. We provide empirical evidence that it does not cause the ASM problem and can help to deploy active learning in practical NLP tasks.

(Nov 13): 7:458:45 (Morning) - Gather

From Reading to Compressing: Exploring the Multi-document Reader for Prompt Compression

Eunseong Choi, Sunkyung Lee, Minjin Choi, June Park, Jongwuk Lee

Large language models (LLMs) have achieved significant performance gains using advanced prompting techniques over various tasks. However, the increasing length of prompts leads to high computational costs and often obscures crucial information. Prompt compression has been proposed to alleviate these issues, but it faces challenges in (i) capturing the global context and (ii) training the compressor effectively. To tackle these challenges, we introduce a novel prompt compression method, namely Reading To Compressing (R2C), utilizing the Fusion-in-Decoder (FiD) architecture to identify the important information in the prompt. Specifically, the cross-attention scores of the FiD are used to discern essential chunks and sentences from the prompt. R2C effectively captures the global context without compromising semantic consistency while detouring the necessity of pseudo-labels for training the compressor. Empirical results show that R2C retains key contexts, enhancing the LLM performance by 6% in out-of-domain evaluations while reducing the prompt length by 80%.

(Nov 13): 7:458:45 (Morning) - Gather

VE-KD: Vocabulary-Expansion Knowledge-Distillation for Training Smaller Domain-Specific Language Models

Pengju Gao, Tomohiro Yamasaki, Kazunori Imoto

We propose VE-KD, a novel method that balances knowledge distillation and vocabulary expansion with the aim of training efficient domain-specific language models. Compared with traditional pre-training approaches, VE-KD exhibits competitive performance in downstream tasks while reducing model size and using fewer computational resources. Additionally, VE-KD refrains from overfitting in domain adaptation. Our experiments with different biomedical domain tasks demonstrate that VE-KD performs well compared with models such as BioBERT (+1% at HoC) and PubMedBERT (+1% at PubMedQA), with about 96% less training time. Furthermore, it outperforms DistilBERT and Adapt-and-Distill, showing a significant improvement in document-level tasks. Investigation of vocabulary size and tolerance, which are hyperparameters of our method, provides insights for further model optimization. The fact that VE-KD consistently maintains its advantages, even when the corpus size is small, suggests that it is a practical approach for domain-specific language tasks and is transferable to different domains for broader applications.

(Nov 13): 7:458:45 (Morning) - Gather

Head-wise Shareable Attention for Large Language Models

zouying cao, Yifei Yang, hai zhao

Large Language Models (LLMs) suffer from huge number of parameters, which restricts their deployment on edge devices. Weight sharing is one promising solution that encourages weight reuse, effectively reducing memory usage with less performance drop. However, current weight sharing techniques primarily focus on small-scale models like BERT and employ coarse-grained sharing rules, e.g., layer-wise. This becomes limiting given the prevalence of LLMs and sharing an entire layer or block obviously diminishes the flexibility of weight sharing. In this paper, we present a perspective on head-wise shareable attention for large language models. We further propose two memory-efficient methods that share parameters across attention heads, with a specific focus on LLMs. Both of them use the same dynamic strategy to select the shared weight matrices. The first method directly reuses the pre-trained weights without retraining, denoted as **DirectShare**. The second method first post-trains with constraint on weight matrix similarity and then shares, denoted as **PostShare**. Experimental results reveal our head-wise shared models still maintain satisfactory capabilities, demonstrating the feasibility of fine-grained weight sharing applied to LLMs.

(Nov 13): 7:458:45 (Morning) - Gather

Normalized Narrow Jump To Conclusions: Normalized Narrow Shortcuts for Parameter Efficient Early Exit Transformer Prediction

Amrit Diggavi Seshadri

With the size and cost of large transformer-based language models growing, recently, there has been interest in shortcut casting of early transformer hidden-representations to final-representations for cheaper model inference. In particular, shortcircuiting pre-trained transformers with linear transformations over early layers has been shown to improve precision in early inference. However, for large language models, even this becomes computationally expensive. In this work, we propose Narrow Jump to Conclusions (NJTC) and Normalized Narrow Jump to Conclusions (N-NJTC) - parameter efficient alternatives to standard linear shortcircuiting that reduces shortcut parameter count by over 97%. We show that N-NJTC reliably outperforms Identity shortcuts at early stages and offers stable precision from all transformer block levels for GPT-2-XL, Phi3-Mini and Llama2-7B transformer models, demonstrating the viability of more parameter efficient short-cutting approaches.

(Nov 13): 7:458:45 (Morning) - Gather

LaCo: Large Language Model Pruning via Layer Collapse

Yifei Yang, zouying cao, hai zhao

Large language models (LLMs) based on transformer are witnessing a notable trend of size expansion, which brings considerable costs to both model training and inference. However, existing methods such as model quantization, knowledge distillation, and model pruning are constrained by various issues, including hardware support limitations, the need for extensive training, and alterations to the model internal structure. In this paper, we propose a concise layer-wise structured pruner called *Layer Collapse* (*LaCo*), in which rear model layers collapse into a prior layer, enabling a rapid reduction in model size while preserving the model structure. Comprehensive experiments show that our method maintains an average task performance of over 80% at pruning ratios of 25-30%, significantly outperforming existing state-of-the-art structured pruning methods. We also conduct post-training experiments to confirm that the *LaCo* effectively inherits the parameters of the original model. Additionally, we perform ablation studies on various settings of *LaCo*. Finally, we discuss our motivation from the perspective of layer-wise similarity and evaluate the performance of the pruned LLMs across various pruning ratios.

(Nov 13): 7:458:45 (Morning) - Gather

Efficient Unseen Language Adaptation for Multilingual Pre-Trained Language Models

Po-Heng Chen, Yun-Nung Chen

Multilingual pre-trained language models (mPLMs) have demonstrated notable effectiveness in zero-shot cross-lingual transfer tasks. Specifically, they can be fine-tuned solely on tasks in the source language and subsequently applied to tasks in the target language. However, for low-resource languages unseen during pre-training, relying solely on zero-shot language transfer often yields sub-optimal results. One common strategy is to continue training PLMs using masked language modeling objectives on the target language. Nonetheless, this approach can

be inefficient due to the need to adjust all parameters for language adaptation. In this paper, we propose a more efficient solution: soft-prompt tuning for language adaptation. Our experiments demonstrate that with carefully designed prompts, soft-prompt tuning enables mPLMs to achieve effective zero-shot cross-lingual transfer to downstream tasks in previously unseen languages. Notably, we found that prompt tuning outperforms continuously trained baselines on two text classification benchmarks, encompassing 20 low-resource languages while utilizing a mere 0.28% of the tuned parameters. These results underscore the superior adaptability of mPLMs to previously unseen languages afforded by soft-prompt tuning compared to traditional fine-tuning methods.

Machine Learning for NLP

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Towards Online Continuous Sign Language Recognition and Translation

Ronglai Zuo, Fangyun Wei, Brian Mak

Research on continuous sign language recognition (CSLR) is essential to bridge the communication gap between deaf and hearing individuals. Numerous previous studies have trained their models using the connectionist temporal classification (CTC) loss. During inference, these CTC-based models generally require the entire sign video as input to make predictions, a process known as offline recognition, which suffers from high latency and substantial memory usage. In this work, we take the first step towards online CSLR. Our approach consists of three phases: 1) developing a sign dictionary; 2) training an isolated sign language recognition model on the dictionary; and 3) employing a sliding window approach on the input sign sequence, feeding each sign clip to the optimized model for online recognition. Additionally, our online recognition model can be extended to support online translation by integrating a gloss-to-text network and can enhance the performance of any offline model. With these extensions, our online approach achieves new state-of-the-art performance on three popular benchmarks across various task settings. Code and models are available at <https://github.com/FangyunWei/SLRT>.

(Nov 13): 7:45:45 (Morning) - Gather

FlipGuard: Defending Preference Alignment against Update Regression with Constrained Optimization

Mingye Zhu, Yi Liu, Quan Wang, Junbo Guo, Zhendong Mao

Recent breakthroughs in preference alignment have significantly improved Large Language Models' ability to generate texts that align with human preferences and values. However, current alignment metrics typically emphasize the post-hoc overall improvement, while overlooking a critical aspect: *regression*, which refers to the backsliding on previously correctly-handled data after updates. This potential pitfall may arise from excessive fine-tuning on already well-aligned data, which subsequently leads to over-alignment and degeneration. To address this challenge, we propose *FlipGuard*, a constrained optimization approach to detect and mitigate update regression with focal attention. Specifically, *FlipGuard* identifies performance degradation using a customized reward characterization and strategically enforces a constraint to encourage conditional congruence with the pre-aligned model during training. Comprehensive experiments demonstrate that *FlipGuard* effectively alleviates update regression while demonstrating excellent overall performance, with the added benefit of knowledge preservation while aligning preferences.

(Nov 13): 7:45:45 (Morning) - Gather

Quantum Recurrent Architectures for Text Classification

Wenduan Xu, Stephen Clark, Douglas Brown, Gabriel Matos, Konstantinos Meichanetzidis

We develop quantum RNNs with cells based on Parametrised Quantum Circuits (PQCs). PQCs can provide a form of hybrid quantum-classical computation where the input and the output is in the form of classical data. The previous "hidden" state is the quantum state from the previous time-step, and an angle encoding is used to define a (non-linear) mapping from a classical word embedding into the quantum Hilbert space. Measurements of the quantum state provide classical statistics which are used for classification. We report results which are competitive with various RNN baselines on the Rotten Tomatoes dataset, as well as emulator results which demonstrate the feasibility of running such models on quantum hardware.

(Nov 13): 7:45:45 (Morning) - Gather

Breaking Language Barriers: Cross-Lingual Continual Pre-Training at Scale

Wenchen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, Ming Zhou

In recent years, Large Language Models (LLMs) have made significant strides towards Artificial General Intelligence. However, training these models from scratch requires substantial computational resources and vast amounts of text data. In this paper, we explore an alternative approach to constructing a LLM for a new language by continually pre-training (CPT) from existing pre-trained LLMs, instead of using randomly initialized parameters. Based on parallel experiments on 40 model sizes ranging from 40M to 5B parameters, we find that 1) CPT converges faster and saves significant resources in a scalable manner. 2) CPT adheres to an extended scaling law derived from with a joint data-parameter scaling term. 3) The compute-optimal data-parameter allocation for CPT markedly differs based on our estimated scaling factors. 4) The effectiveness of transfer scale is influenced by training duration and linguistic properties, while robust to data replaying, a method that effectively mitigates catastrophic forgetting in CPT. We hope our findings provide deeper insights into the transferability of LLMs at scale for the research community.

(Nov 13): 7:45:45 (Morning) - Gather

Preference-Guided Reflective Sampling for Aligning Language Models

Hai Ye, Hwee Tou Ng

Iterative data generation and model re-training can effectively align large language models (LLMs) to human preferences. The process of data sampling is crucial, as it significantly influences the success of policy improvement. Repeated random sampling is a widely used method that independently queries the model multiple times to generate outputs. In this work, we propose a more effective sampling method, named Preference-Guided Reflective Sampling (PRS). Unlike random sampling, PRS employs a tree-based generation framework to enable more efficient sampling. It leverages adaptive self-refinement techniques to better explore the sampling space. By specifying user preferences in natural language, PRS can further optimize response generation according to these preferences. As a result, PRS can align models to diverse user preferences. Our experiments demonstrate that PRS generates higher-quality responses with significantly higher rewards. On AlpacaEval and Arena-Hard, PRS substantially outperforms repeated random sampling in best-of- N sampling. Moreover, PRS shows strong performance when applied in iterative offline RL training.

(Nov 13): 7:45:45 (Morning) - Gather

LPZero: Language Model Zero-cost Proxy Search from Zero

Peijie Dong, Lujun Li, Xiang Liu, Zhenheng Tang, Xuebo Liu, Qiang Wang, Xiaowen Chu

Despite the outstanding performance, Neural Architecture Search (NAS) is criticized for massive computation. Recently, Zero-shot NAS has emerged as a promising approach by exploiting Zero-cost (ZC) proxies, which markedly reduce computational demands. Despite this, existing ZC proxies heavily rely on expert knowledge and incur significant trial-and-error costs. Particularly in NLP tasks, most existing ZC proxies fail to surpass the performance of the naive baseline. To address these challenges, we introduce a novel framework, LPZero, which is the first to automatically design zero-cost (ZC) proxies for various tasks, achieving higher ranking consistency than human-designed proxies. Specifically, we model the ZC proxy as a symbolic equation and incorporate a unified proxy search space that encompasses existing ZC proxies, which are composed of a predefined set of mathematical symbols. To heuristically search for the best ZC proxy, LPZero incorporates genetic programming to find the optimal symbolic composition. We propose a Predictive-Pruning Strategy (PPS), which preemptively eliminates unpromising proxies, thereby mitigating the risk of proxy degradation. Extensive experiments on FlexiBERT, GPT-2, and LLaMA-7B demonstrate LPZero's superior ranking ability and performance on downstream tasks compared to current approaches.

Machine Translation

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

LLMs Are Zero-Shot Context-Aware Simultaneous Translators

Roman Koskin, Katsuhiro Sudoh, Satoshi Nakamura

The advent of transformers has fueled progress in machine translation. More recently large language models (LLMs) have come to the spotlight thanks to their generality and strong performance in a wide range of language tasks, including translation. Here we show that open-source LLMs perform on par with or better than some state-of-the-art baselines in simultaneous machine translation (SiMT) tasks, zero-shot. We also demonstrate that injection of minimal background information, which is easy with an LLM, brings further performance gains, especially on challenging technical subject-matter. This highlights LLMs' potential for building next generation of massively multilingual, context-aware and terminologically accurate SiMT systems that require no resource-intensive training or fine-tuning.

(Nov 13): 7:458:45 (Morning) - Gather

Finding the Optimal Byte-Pair Encoding Merge Operations for Neural Machine Translation in a Low-Resource Setting

Kristine Mae M. Adlaon

This paper investigates the impact of different Byte Pair Encoding (BPE) configurations, specifically, merge operations on neural machine translation (NMT) performance for the Filipino-Cebuano language pair across various text domains. Results demonstrate that smaller BPE configurations, notably 2k, 5k, and 8k consistently yield higher BLEU scores, indicating improved translation quality through finer tokenization granularity. Conversely, larger BPE configurations and the absence of BPE result in lower BLEU scores, suggesting a decline in translation quality due to coarser tokenization. Additionally, these findings help us understand how the size of the model and how finely we break down words affect the quality of translations. This knowledge will be useful for improving translation systems, especially for languages that don't have many parallel texts available for training.

Multilinguality and Language Diversity

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

MTLS: Making Texts into Linguistic Symbols

Wenlong Fei, Xiaohua Wang, Min Hu, Qingyu Zhang, Hongbo Li

In linguistics, all languages can be considered as symbolic systems, with each language relying on symbolic processes to associate specific symbols with meanings. In the same language, there is a fixed correspondence between linguistic symbol and meaning. In different languages, universal meanings follow varying rules of symbolization in one-to-one correspondence with symbols. Most work overlooks the properties of languages as symbol systems. In this paper, we shift the focus to the symbolic properties and introduce MTLS: a pre-training method to improve the multilingual capability of models by Making Texts into Linguistic Symbols. Initially, we replace the vocabulary in pre-trained language models by mapping relations between linguistic symbols and semantics. Subsequently, universal semantics within the symbolic system serve as bridges, linking symbols from different languages to the embedding space of the model, thereby enabling the model to process linguistic symbols. To evaluate the effectiveness of MTLS, we conducted experiments on multilingual tasks using BERT and RoBERTa, respectively, as the backbone. The results indicate that despite having just over 12,000 pieces of English data in pre-training, the improvement that MTLS brings to multilingual capabilities is remarkably significant.

(Nov 13): 7:458:45 (Morning) - Gather

What is lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations

Kavya Manohar, Leena G Pillai

This paper explores the pitfalls in evaluating multilingual automatic speech recognition (ASR) models, with a particular focus on Indic language scripts. We investigate the text normalization routine employed by leading ASR models, including OpenAI Whisper, Meta's MMS, Seamless, and Assembly AI's Conformer, and their unintended consequences on performance metrics. Our research reveals that current text normalization practices, while aiming to standardize ASR outputs for fair comparison, by removing inconsistencies such as variations in spelling, punctuation, and special characters, are fundamentally flawed when applied to Indic scripts. Through empirical analysis using text similarity scores and in-depth linguistic examination, we demonstrate that these flaws lead to artificially improved performance metrics for Indic languages. We conclude by proposing a shift towards developing text normalization routines that leverage native linguistic expertise, ensuring more robust and accurate evaluations of multilingual ASR models.

(Nov 13): 7:458:45 (Morning) - Gather

LLM-based Code-Switched Text Generation for Grammatical Error Correction

Tom Potter, Zheng Yuan

With the rise of globalisation, code-switching (CSW) has become a ubiquitous part of multilingual conversation, posing new challenges for natural language processing (NLP), especially in Grammatical Error Correction (GEC). This work explores the complexities of applying GEC systems to CSW texts. Our objectives include evaluating the performance of state-of-the-art GEC systems on an authentic CSW dataset from English as a Second Language (ESL) learners, exploring synthetic data generation as a solution to data scarcity, and developing a model

capable of correcting grammatical errors in monolingual and CSW texts. We generated synthetic CSW GEC data, resulting in one of the first substantial datasets for this task, and showed that a model trained on this data is capable of significant improvements over existing systems. This work targets ESL learners, aiming to provide education technologies that aid in the development of their English grammatical correctness without constraining their natural multilingualism.

(Nov 13): 7:45:45 (Morning) - Gather

Compression Parity: Measuring and Predicting the Multilingual Capabilities of Language Models

Alexander Tsvetkov, Alon Kipnis

Large Language Models (LLMs) are increasingly deployed in user-facing applications worldwide, necessitating handling multiple languages across various tasks. We propose a metric called Information Parity (IP) that can predict an LLM's capabilities across multiple languages in a task-agnostic manner. IP is well-motivated from an information theoretic perspective: it is associated with the LLMs efficiency of compressing the text in a given language compared to a reference language. We evaluate IP and other popular metrics such as Tokenization Parity (TP) and Tokenizer Fertility (TF) on several variants of open-sourced LLMs (Llama2, Gemma, Mistral). Among all metrics known to us, IP is better correlated with existing task-specific benchmark scores from the literature and thus better predicts such scores in a certain language. These findings show that IP may be useful for ranking multilingual LLMs' capabilities regardless of the downstream task.

(Nov 13): 7:45:45 (Morning) - Gather

PreAlign: Boosting Cross-Lingual Transfer by Early Establishment of Multilingual Alignment

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, Jiajun Chen

Large language models demonstrate reasonable multilingual abilities, despite predominantly English-centric pretraining. However, the spontaneous multilingual alignment in these models is shown to be weak, leading to unsatisfactory cross-lingual transfer and knowledge sharing. Previous works attempt to address this issue by explicitly injecting multilingual alignment information during or after pretraining. Thus for the early stage in pretraining, the alignment is weak for sharing information or knowledge across languages. In this paper, we propose PreAlign, a framework that establishes multilingual alignment prior to language model pretraining. PreAlign injects multilingual alignment by initializing the model to generate similar representations of aligned words and preserves this alignment using a code-switching strategy during pretraining. Extensive experiments in a synthetic English to English-Clone setting demonstrate that PreAlign significantly outperforms standard multilingual joint training in language modeling, zero-shot cross-lingual transfer, and cross-lingual knowledge application. Further experiments in real-world scenarios further validate PreAlign's effectiveness across various model sizes.

(Nov 13): 7:45:45 (Morning) - Gather

Towards a Greek Proverb Atlas: Computational Spatial Exploration and Attribution of Greek Proverbs

John Pavlopoulos, Panos Louridas, Panagiotis Filios

Proverbs carry wisdom transferred orally from generation to generation. Based on the place they were recorded, this study introduces a publicly-available and machine-actionable dataset of more than one hundred thousand Greek proverb variants. By quantifying the spatial distribution of proverbs, we show that the most widespread proverbs come from the mainland while the least widespread proverbs come primarily from the islands. By focusing on the least dispersed proverbs, we present the most frequent tokens per location and undertake a benchmark in geographical attribution, using text classification and regression (text geocoding). Our results show that this is a challenging task for which specific locations can be attributed more successfully compared to others. The potential of our resource and benchmark is showcased by two novel applications. First, we extracted terms moving the regression prediction toward the four cardinal directions. Second, we leveraged conformal prediction to attribute 3,676 unregistered proverbs with statistically rigorous predictions of locations each of these proverbs was possibly registered in.

Multimodality and Language Grounding to Vision, Robotics and Beyond

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Mitigating Language Bias of LMMs in Social Intelligence Understanding with Virtual Counterfactual Calibration

Peng Chen, Xiao-Yu Guo, Yuan-Fang Li, Xiaowang Zhang, Zhiyong Feng

Social intelligence is essential for understanding complex human expressions and social interactions. While large multimodal models (LMMs) have demonstrated remarkable performance in social intelligence question answering (SIQA), they are still inclined to generate responses relying on language priors and ignoring the relevant context, due to the dominant prevalence of text-based data in the pre-training stage. To interpret the aforementioned language bias of LMMs, we employ a structure causal model and posit that counterfactual reasoning can mitigate the bias by avoiding spurious correlations between LMMs' internal commonsense knowledge and the given context. However, it is costly and challenging to construct multimodal counterfactual samples. To tackle above challenges, we propose an output Distribution Calibration network with Virtual Counterfactual (DCVC) data augmentation framework. DCVC devises a novel output distribution calibration network to mitigate the impact of negative language biases while preserving beneficial priors. Perturbations are introduced to the output distributions of LMMs to simulate the distribution shifts from counterfactual manipulations of the context, which is employed to construct counterfactual augmented data virtually. Experiments on multiple datasets demonstrate the effectiveness and generalizability of our proposed method.

(Nov 13): 7:45:45 (Morning) - Gather

Predicate Debiasing in Vision-Language Models Integration for Scene Graph Generation Enhancement

Yuxuan Wang, Xiaoyuan Liu

Scene Graph Generation (SGG) provides basic language representation of visual scenes, requiring models to grasp complex and diverse semantics between objects. This complexity and diversity in SGG leads to underrepresentation, where parts of triplet labels are rare or even unseen during training, resulting in imprecise predictions. To tackle this, we propose integrating the pretrained Vision-language Models to enhance representation. However, due to the gap between pretraining and SGG, direct inference of pretrained VLMs on SGG leads to severe bias, which stems from the imbalanced predicates distribution in the pretraining language set. To alleviate the bias, we introduce a novel LM Estimation to approximate the unattainable predicates distribution. Finally, we ensemble the debiased VLMs with SGG models to enhance the representation, where we design a certainty-aware indicator to score each sample and dynamically adjust the ensemble weights. Our training-free method effectively addresses the predicates bias in pretrained VLMs, enhances SGG's representation, and significantly improve the performance.

(Nov 13): 7:45:45 (Morning) - Gather

Enhancing Advanced Visual Reasoning Ability of Large Language Models

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, Weidong Cai

Recent advancements in Vision-Language (VL) research have sparked new benchmarks for complex visual reasoning, challenging models' advanced reasoning ability. Traditional Vision-Language models (VLMs) perform well in visual perception tasks while struggling with complex reasoning scenarios. Conversely, Large Language Models (LLMs) demonstrate robust text reasoning capabilities; however, they lack visual acuity. To bridge this gap, we propose **C**omplex **V**isual **R**easoning **L**arge **L**anguage **M**odels (**CVR-LLM**), capitalizing on VLMs' visual perception proficiency and LLMs' extensive reasoning capability. Unlike recent multimodal large language models (MLLMs) that require a projection layer, our approach transforms images into detailed, context-aware descriptions using an iterative self-refinement loop and leverages LLMs' text knowledge for accurate predictions without extra training. We also introduce a novel multi-modal in-context learning (ICL) methodology to enhance LLMs' contextual understanding and reasoning. Additionally, we introduce Chain-of-Comparison (CoC), a step-by-step comparison technique enabling contrasting various aspects of predictions. Our CVR-LLM presents the first comprehensive study across a wide array of complex visual reasoning tasks and achieves SOTA performance among all.

(Nov 13): 7:45:45 (Morning) - Gather

Empowering Large Language Model for Continual Video Question Answering with Collaborative Prompting

Chen Cai, Zheng Wang, Jianjun Gao, Wenyang Liu, Ye Lu, Runzhong Zhang, Kim-Hui Yap

In recent years, the rapid increase in online video content has underscored the limitations of static Video Question Answering (VideoQA) models trained on fixed datasets, as they struggle to adapt to new questions or tasks posed by newly available content. In this paper, we explore the novel challenge of VideoQA within a continual learning framework, and empirically identify a critical issue: fine-tuning a large language model (LLM) for a sequence of tasks often results in catastrophic forgetting. To address this, we propose Collaborative Prompting (CoPro), which integrates specific question constraint prompting, knowledge acquisition prompting, and visual temporal awareness prompting. These prompts aim to capture textual question context, visual content, and video temporal dynamics in VideoQA, a perspective underexplored in prior research. Experimental results on the NExT-QA and DramaQA datasets show that CoPro achieves superior performance compared to existing approaches, achieving 55.14% accuracy on NExT-QA and 71.24% accuracy on DramaQA, highlighting its practical relevance and effectiveness.

(Nov 13): 7:45:45 (Morning) - Gather

World to Code: Multi-modal Data Generation via Self-Instructed Compositional Captioning and Filtering

Jiacong Wang, Bohong Wu, Haiyong Jiang, Haoyuan Guo, Xin Xiao, Zhou Xun, Jun Xiao

Recent advances in Vision-Language Models (VLMs) and the scarcity of high-quality multi-modal alignment data have inspired numerous researches on synthetic VLM data generation. The conventional norm in VLM data construction uses a mixture of specialists in caption and OCR, or stronger VLM APIs and expensive human annotation. In this paper, we present World to Code (*W2C*), a meticulously curated multi-modal data construction pipeline that organizes the final generation output into a Python code format. The pipeline leverages the VLM itself to extract cross-modal information via different prompts and filter the generated outputs again via a consistency filtering strategy. Experiments have demonstrated the high quality of *W2C* by improving various existing visual question answering and visual grounding benchmarks across different VLMs. Further analysis also demonstrates that the new code parsing ability of VLMs presents better cross-modal equivalence than the commonly used detail caption ability. Our code is available at <https://github.com/foundation-multimodal-models/World2Code>.

(Nov 13): 7:45:45 (Morning) - Gather

RWKV-CLIP: A Robust Vision-Language Representation Learner

Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, Jiankang Deng

Contrastive Language-Image Pre-training (CLIP) has significantly improved performance in various vision-language tasks by expanding the dataset with image-text pairs obtained from the web. This paper further explores CLIP from the perspectives of data and model architecture. To mitigate the impact of the noise data and enhance the quality of large-scale image-text data crawled from the internet, we introduce a diverse description generation framework that can leverage Large Language Models (LLMs) to combine and refine information from web-based image-text pairs, synthetic captions, and detection tags. Additionally, we propose RWKV-CLIP, the first RWKV-driven vision-language representation learning model that combines the effective parallel training of transformers with the efficient inference of RNNs. Extensive experiments across different model scales and pre-training datasets demonstrate that RWKV-CLIP is a robust vision-language representation learner and it achieves state-of-the-art performance across multiple downstream tasks, including linear probing, zero-shot classification, and zero-shot image-text retrieval. To facilitate future research, the code and pre-trained models are released at <https://github.com/deepglint/RWKV-CLIP>.

(Nov 13): 7:45:45 (Morning) - Gather

Beyond Embeddings: The Promise of Visual Table in Visual Reasoning

Yiwei Zhong, Zi-Yuan Hu, Michael Lyu, Liwei Wang

Visual representation learning has been a cornerstone in computer vision, involving typical forms such as visual embeddings, structural symbols, and text-based representations. Despite the success of CLIP-type visual embeddings, they often lack access to world knowledge critical for visual reasoning. In this work, we propose Visual Table, a novel form of visual representation tailored for visual reasoning. Visual tables are constructed as hierarchical descriptions of visual scenes, featuring a scene description and multiple object-centric descriptions covering categories, attributes, and knowledge. Thanks to the structural and textual formats, visual tables offer unique properties over mere visual embeddings, such as explainability and controllable editing. Furthermore, they deliver instance-level world knowledge and detailed attributes that are essential for visual reasoning. To create visual tables, we develop a generator trained on the dataset with collected, small-scale annotations. Extensive results on 11 visual reasoning benchmarks demonstrate that the generated visual tables significantly outperform previous structural and text-based representations. Moreover, they consistently enhance state-of-the-art multi-modal large language models across diverse benchmarks, showcasing their potential for advancing visual reasoning tasks. Our code is available at <https://github.com/LaVi-Lab/Visual-Table>.

(Nov 13): 7:45:45 (Morning) - Gather

DAMRO: Dive into the Attention Mechanism of LVLML to Reduce Object Hallucination

Xuan Gong, Tianshi Ming, Xinpeng Wang, Zhihua Wei

Despite the great success of Large Vision-Language Models (LVLMs), they inevitably suffer from hallucination. As we know, both the visual encoder and the Large Language Model (LLM) decoder in LVLMs are Transformer-based, allowing the model to extract visual information and generate text outputs via attention mechanisms. We find that the attention distribution of LLM decoder on image tokens is highly consistent with the visual encoder and both distributions tend to focus on particular background tokens rather than the referred objects in the image. We attribute to the unexpected attention distribution to an inherent flaw in the visual encoder itself, which misguides LLMs to over emphasize the redundant information and generate object hallucination. To address the issue, we propose DAMRO, a novel training-free strategy that **P**rojects **A**ttention **M**echanism of LVLM to **R**educe **O**bject Hallucination. Specifically, our approach employs classification token (CLS) of ViT to filter out high-attention tokens scattered in the background and then eliminate their influence during decoding stage. We evaluate our method on LVLMs including LLaVA-1.5, LLaVA-NeXT and InstructBLIP, using various benchmarks such as POPE, CHAIR, MME and GPT-4V Aided Evaluation. The results demonstrate that our approach significantly reduces the impact of these outlier tokens, thus effectively alleviating the hallucination of LVLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Efficient Vision-Language pre-training via domain-specific learning for human activities

Adrian Bulat, Yassine Ouali, Ricardo Guerrero, Brais Martinez, Georgios Tzimiropoulos

Current Vision-Language (VL) models owe their success to large-scale pre-training on web-collected data, which in turn requires high-capacity architectures and large compute resources for training. We posit that when the downstream tasks are known in advance, which is in practice common, the pretraining process can be aligned to the downstream domain, leading to more efficient and accurate models, while shortening the pretraining step. To this end, we introduce a domain-aligned pretraining strategy that, without additional data collection, improves the accuracy on a domain of interest, herein, that of human activities, while largely preserving the generalist knowledge. At the core of our approach stands a new LLM-based method that, provided with a simple set of concept seeds, produces a concept hierarchy with high coverage of the target domain. The concept hierarchy is used to filter a large-scale web-crawled dataset and, then, enhances the resulting instances with targeted synthetic labels. We study in depth how to train such approaches and their resulting behavior. We further show generalization to video-based data by introducing a fast adaptation approach for transitioning from a static (image) model to a dynamic one (i.e. with temporal modeling). On the domain of interest, our approach significantly outperforms models trained on up to $60 \times$ more samples and between $10 - 100 \times$ shorter training schedules for image retrieval, video retrieval and action recognition. Code will be released.

(Nov 13): 7:45:45 (Morning) - Gather

Modeling Layout Reading Order as Ordering Relations for Visually-rich Document Understanding

Chong Zhang, Yi Tu, Yixi Zhao, Chenshu Yuan, Huan Chen, Yue Zhang, Mingxu Chai, Ya Guo, Huijia Zhu, Qi Zhang, Tao Gui

Modeling and leveraging layout reading order in visually-rich documents (VrDs) is critical in document intelligence as it captures the rich structure semantics within documents. Previous works typically formulated layout reading order as a permutation of layout elements, i.e. a sequence containing all the layout elements. However, we argue that this formulation does not adequately convey the complete reading order information in the layout, which may potentially lead to performance decline in downstream tasks. To address this issue, we propose to model the layout reading order as ordering relations over the set of layout elements, which have sufficient expressive capability for the complete reading order information. To enable empirical evaluation on methods towards the improved form of reading order prediction (ROP), we establish a comprehensive benchmark dataset including the reading order annotation as relations over layout elements, together with a relation-extraction-based method that outperforms previous models. Moreover, we propose a reading-order-relation-enhancing pipeline to improve model performance on any arbitrary VrD task by introducing additional reading order relation inputs. We conduct comprehensive experiments to demonstrate that the pipeline generally benefits downstream VrD tasks: (1) with utilizing the reading order relation information, the enhanced downstream models achieve SOTA results on both two task settings of the targeted dataset; (2) with utilizing the pseudo reading order information generated by the proposed ROP model, the performance of the enhanced models has improved across all three models and eight cross-domain VrD-IE/QA task settings without targeted optimization.

(Nov 13): 7:45:45 (Morning) - Gather

Large Language Models Know What is Key Visual Entity: An LLM-assisted Multimodal Retrieval for VQA

Pu Jian, Donglei Yu, Jiajun Zhang

Visual question answering (VQA) tasks, often performed by visual language model (VLM), face challenges with long-tail knowledge. Recent retrieval-augmented VQA (RA-VQA) systems address this by retrieving and integrating external knowledge sources. However, these systems still suffer from redundant visual information irrelevant to the question during retrieval. To address these issues, in this paper, we propose LLM-RA, a novel method leveraging the reasoning capability of a large language model (LLM) to identify key visual entities, thus minimizing the impact of irrelevant information in the query of retriever. Furthermore, key visual entities are independently encoded for multimodal joint retrieval, preventing cross-entity interference. Experimental results demonstrate that our method outperforms other strong RA-VQA systems. In two knowledge-intensive VQA benchmarks, our method achieves the new state-of-the-art performance among those with similar scale of parameters and even performs comparably to models with 1-2 orders larger parameters.

(Nov 13): 7:45:45 (Morning) - Gather

VideoCLIP-XL: Advancing Long Description Understanding for Video CLIP Models

Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, Lianwen Jin

Contrastive Language-Image Pre-training (CLIP) has been widely studied and applied in numerous applications. However, the emphasis on brief summary texts during pre-training prevents CLIP from understanding long descriptions. This issue is particularly acute regarding videos given that videos often contain abundant detailed contents. In this paper, we propose the VideoCLIP-XL (eXtra Length) model, which aims to unleash the long-description understanding capability of video CLIP models. Firstly, we establish an automatic data collection system and gather a large-scale VILD pre-training dataset with Video and Long-Description pairs. Then, we propose Text-similarity-guided Primary Component Matching (TPCM) to better learn the distribution of feature space while expanding the long description capability. We also introduce two new tasks namely Detail-aware Description Ranking (DDR) and Hallucination-aware Description Ranking (HDR) for further understanding improvement. Finally, we construct a Long Video Description Ranking (LVDR) benchmark for evaluating the long-description capability more comprehensively. Extensive experimental results on widely-used text-video retrieval benchmarks with both short and long descriptions and our LVDR benchmark can fully demonstrate the effectiveness of our method.

(Nov 13): 7:45:45 (Morning) - Gather

In-Context Compositional Generalization for Large Vision-Language Models

Chuanhao Li, Chenchen Jing, Zhen Li, Mingliang Zhai, Yuwei Wu, Yunde Jia

Recent work has revealed that in-context learning for large language models exhibits compositional generalization capacity, which can be enhanced by selecting in-context demonstrations similar to test cases to provide contextual information. However, how to exhibit in-context compositional generalization (ICCG) of large vision-language models (LVLMs) is non-trivial. Due to the inherent asymmetry between visual and linguistic modalities, ICCG in LVLMs faces an inevitable challenge—redundant information on the visual modality. The redundant information affects in-context learning from two aspects: (1) Similarity calculation may be dominated by redundant information, resulting in sub-optimal demonstration selection. (2) Redundant information in in-context demonstrations brings misleading contextual information to in-context learning. To alleviate these problems, we propose a demonstration selection method to achieve ICCG for LVLMs, by considering two key factors of demonstrations: content and structure, from a multimodal perspective. Specifically, we design a diversity-coverage-based matching score to select demonstrations with maximum coverage, and avoid selecting demonstrations with redundant information via their content redundancy and structural complexity. We build a QQA-ICCG dataset to simulate the ICCG setting, and conduct experiments on GQA-ICCG and the VQA v2 dataset. Experimental results demonstrate the effectiveness of our method.

(Nov 13): 7:45:45 (Morning) - Gather

I-AM-G: Interest Augmented Multimodal Generator for Item Personalization

Xianquan Wang, Likang Wu, Shukang Yin, Zhi Li, Yanjiang Chen, hufeng, Yu Su, Qi Liu

The emergence of personalized generation has made it possible to create texts or images that meet the unique needs of users. Recent advances mainly focus on style or scene transfer based on given keywords. However, in e-commerce and recommender systems, it is almost an untouched area to explore user historical interactions, automatically mine user interests with semantic associations, and create item representations that closely align with user individual interests. In this paper, we propose a brand new framework called **I**-interest-**A**-augmented

Multimodal **G**enerator (**I-AM-G**). The framework first extracts tags from the multimodal information of items that the user has interacted with, and the most frequently occurred ones are extracted to rewrite the text description of the item. Then, the framework uses a decoupled text-to-text and image-to-image retriever to search for the top- K similar item text and image embeddings from the item pool. Finally, the Attention module for user interests fuses the retrieved information in a cross-modal manner and further guides the personalized generation process in collaboration with the rewritten text. We conducted extensive and comprehensive experiments to demonstrate that our framework can effectively generate results aligned with user preferences, which potentially provides a new paradigm of **Rewrite and Retrieve** for personalized generation.

(Nov 13): 7:45:45 (Morning) - Gather

Visual Pivoting Unsupervised Multimodal Machine Translation in Low-Resource Distant Language Pairs

Turghun Tayir, Linlin Tao, Mieradilijiang Maimaiti, Ming Li, Jianquan Liu

Unsupervised Multi-modal machine translation (UMMT) aims to leverage vision information as a pivot between two languages to achieve better performance on low-resource language pairs. However, there is presently a challenge: how to handle alignment between distant language pairs (DLPs) in UMMT. To this end, this paper proposes a visual pivoting UMMT method for DLPs. Specifically, we first construct a dataset containing two DLPs, including English-Uyghur and Chinese-Uyghur. We then apply the visual pivoting method for both to pre-training and fine-tuning, and we observe that the images on the encoder and decoder of UMMT have noticeable effects on DLPs. Finally, we introduce informative multi-granularity image features to facilitate further alignment of the latent space between the two languages. Experimental results show that the proposed method significantly outperforms several baselines on DLPs and close language pairs (CLPs).

(Nov 13): 7:45:45 (Morning) - Gather

Query-based Cross-Modal Projector Bolstering Mamba Multimodal LLM

SooHwan Eom, Jay Shim, Gwanhyeong Koo, Haebin Na, Mark A. Hasegawa-Johnson, Sungwoong Kim, Chang D. Yoo

The Transformer's quadratic complexity with input length imposes an unsustainable computational load on large language models (LLMs). In contrast, the Selective Scan Structured State-Space Model, or Mamba, addresses this computational challenge effectively. This paper explores a query-based cross-modal projector designed to bolster Mamba's efficiency for vision-language modeling by compressing visual tokens based on input through the cross-attention mechanism. This innovative projector also removes the need for manually designing the 2D scan order of original image features when converting them into an input sequence for Mamba LLM. Experimental results across various vision-language understanding benchmarks show that the proposed cross-modal projector enhances Mamba-based multimodal LLMs, boosting both performance and throughput.

(Nov 13): 7:45:45 (Morning) - Gather

MultiSkill: Evaluating Large Multimodal Models for Fine-grained Alignment Skills

Zhenran Xu, Senbao Shi, Baotian Hu, Longyue Wang, Min Zhang

We propose MultiSkill, an evaluation protocol that assesses large multimodal models (LMMs) across multiple fine-grained skills for alignment with human values. Recent LMMs have shown various intriguing abilities, such as solving graph theory problems and explaining various jokes. However, existing multimodal benchmarks have mainly focused on coarse-grained evaluation (e.g., accuracy), without considering the skill composition required by specific instructions. To this end, we present MultiSkill, designed to decompose coarse-level scoring to a fine-grained skill set-level scoring tailored to each instruction. MultiSkill defines five core vision-language capabilities and divides into 12 skills that are necessary to align with user instructions. For evaluation metrics on specific skills, we propose an LMM-based evaluator for open-ended outputs. Based on the diverse instructions collected from 66 datasets spanning 10 domains, we compare multiple representative open-source and proprietary LMMs and find a high correlation between model-based and human-based evaluations. Our experiments underscore the importance of fine-grained evaluation in providing a holistic view of model performance and enhancing the reliability of the evaluation.

(Nov 13): 7:45:45 (Morning) - Gather

MantisScore: A Reliable Fine-grained Metric for Video Generation

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuanzheng Ni, Bohan Lyu, Yaswanth Narsapalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, Wenhui Chen

The recent years have witnessed great advances in video generation. However, the development of automatic video metrics is lagging significantly behind. None of the existing metric is able to provide reliable scores over generated videos. The main barrier is the lack of large-scale human-annotated dataset. In this paper, we release VideoFeedback, the first large-scale dataset containing human-provided multi-aspect score over 37.6K synthesized videos from 11 existing video generative models. We train VideoScore (initialized from Mantis) based on VideoFeedback to enable automatic video quality assessment. Experiments show that the Spearman's correlation between VideoScore and humans can reach 77.1 on VideoFeedback-test, beating the prior best metrics by about 50 points. Further result another held-out EvalCrafter, GenAI-Bench, and VBench show that VideoScore has consistently much higher correlation with humanjudges than other metrics. Due to these results, we believe VideoScore can serve as a great proxy for human raters to (1) rate different video models to track progress (2) simulate fine-grained human feedback in Reinforcement Learning with Human Feedback (RLHF) to improve current video generation models.

(Nov 13): 7:45:45 (Morning) - Gather

FineCops-Ref: A New Dataset and Task for Fine-Grained Compositional Referring Expression Comprehension

Junzhou Liu, Xusheng Yang, WEIWEI LI, Peng Wang

Referring Expression Comprehension (REC) is a crucial cross-modal task that objectively evaluates the capabilities of language understanding, image comprehension, and language-to-image grounding. Consequently, it serves as an ideal testing ground for Multi-modal Large Language Models (MLLMs). In pursuit of this goal, we have established a new REC dataset characterized by two key features: Firstly, it is designed with controllable varying levels of difficulty, necessitating multi-level fine-grained reasoning across object categories, attributes, and multi-hop relationships. Secondly, it includes negative text and images created through fine-grained editing and generation based on existing data, thereby testing the model's ability to correctly reject scenarios where the target object is not visible in the image. An essential aspect often overlooked in existing datasets and approaches. Utilizing this high-quality dataset, we conducted comprehensive evaluations of both state-of-the-art specialist models and MLLMs. Our findings indicate that there remains a significant gap in achieving satisfactory grounding performance. We anticipate that our dataset will inspire new approaches to enhance visual reasoning and develop more advanced cross-modal interaction strategies, ultimately unlocking the full potential of MLLMs.

(Nov 13): 7:45:45 (Morning) - Gather

GRIZAL: Generative Prior-guided Zero-Shot Temporal Action Localization

Onkar Kishor Sudladkar, Gayatri Sudhir Deshmukh, Vandna Gorade, Sparsh Mittal

Zero-shot temporal action localization (TAL) aims to temporally localize actions in videos without prior training examples. To address the challenges of TAL, we offer GRIZAL, a model that uses multimodal embeddings and dynamic motion cues to localize actions effectively. GRIZAL achieves sample diversity by using large-scale generative models such as GPT-4 for generating textual augmentations and DALL-E for generating image augmentations. Our model integrates vision-language embeddings with optical flow insights, optimized through a blend

of supervised and self-supervised loss functions. On ActivityNet, Thumos14 and Charades-STA datasets, GRIZAL greatly outperforms state-of-the-art zero-shot TAL models, demonstrating its robustness and adaptability across a wide range of video content. We will make all the models and code publicly available by open-sourcing them.

(Nov 13): 7:45:45 (Morning) - Gather

Rationale-based Ensemble of Multiple QA Strategies for Zero-shot Knowledge-based VQA

Miaoyu Li, Haixin Li, Zilin Du, Boyang Li

Knowledge-based Visual Question-answering (K-VQA) often requires the use of background knowledge beyond the image. However, we discover that a single knowledge generation strategy is often insufficient for all K-VQA questions. To this end, we propose Diversification, Evidence Truncation, and Combination for Knowledge-based Elucidation (DietCoke), which utilizes a bundle of complementary question-answering tactics and aggregates their answers using textual rationales. DietCoke comprises of three stages: diversification, rationalization, and ensemble. The diversification stage generates three distinctive decision contexts, each leading to its own answer candidate. The rationalization stage generates two rationales, the automatic rationale and the mechanistic rationale, for each answer candidate using decorrelated techniques. Finally, in the ensemble stage, an LLM informed by the rationales selects one answer from the three candidates. Experiments show that DietCoke significantly outperforms state-of-the-art LLM-based baselines by 2.8% on OK-VOA and 4.7% on A-OKVOA and that the strategies in the ensembles are highly complementary.

(Nov 13): 7:45:45 (Morning) - Gather

MaPPER: Multimodal Prior-guided Parameter Efficient Tuning for Referring Expression Comprehension

Ting Liu, Zunman Xu, Zhigang Wang, Yue Hu, Liangtao Shi, Quanjun Yin

Referring Expression Comprehension (REC), which aims to ground a local visual region via natural language, is a task that heavily relies on multimodal alignment. Most existing methods utilize powerful pre-trained models to transfer visual/linguistic knowledge by full fine-tuning. However, full fine-tuning the entire backbone not only breaks the rich prior knowledge embedded in the pre-training, but also incurs significant computational costs. Motivated by the recent emergence of Parameter-Efficient Transfer Learning (PETL) methods, we aim to solve the REC task in an effective and efficient manner. Directly applying these PETL methods to the REC task is inappropriate, as they lack the specific-domain abilities for precise local visual perception and visual-language alignment. Therefore, we propose a novel framework of Multimodal Prior-guided Parameter Efficient Tuning, namely MaPPER. Specifically, MaPPER comprises Dynamic Prior Adapters guided by a aligned prior, and Local Convolution Adapters to extract precise local semantics for better visual perception. Moreover, the Prior-Guided Text module is proposed to further utilize the prior for facilitating the cross-modal alignment. Experimental results on three widely-used benchmarks demonstrate that MaPPER achieves the best accuracy compared to the full fine-tuning and other PETL methods with only 1.41% tunable backbone parameters.

(Nov 13): 7:45:45 (Morning) - Gather

VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazhen Yang, Benyon Wang, Lingpeng Kong, Qi Liu

As large vision-language models (VLVMs) evolve rapidly, the demand for high-quality and diverse data to align these models becomes increasingly crucial. However, the creation of such data with human supervision proves costly and time-intensive. In this paper, we investigate the efficacy of AI feedback to scale supervision for aligning VLVMs. We introduce VLFeedback, the first large-scale vision-language feedback dataset, comprising over 82K multi-modal instructions and comprehensive rationales generated by off-the-shelf models without human annotations. To evaluate the effectiveness of AI feedback for vision-language alignment, we train Silkie, an LVLM fine-tuned via direct preference optimization on VLFeedback. Silkie showcases exceptional performance regarding helpfulness, visual faithfulness, and safety metrics. It outperforms its base model by 6.9% and 9.5% in perception and cognition tasks, reduces hallucination issues on MMHal-Bench, and exhibits enhanced resilience against red-teaming attacks. Furthermore, our analysis underscores the advantage of AI feedback, particularly in fostering preference diversity to deliver more comprehensive improvements. Our dataset, training code and models are available at <https://vlf-silkie.github.io>.

(Nov 13): 7:45:45 (Morning) - Gather

AudioVSR: Enhancing Video Speech Recognition with Audio Data

Xiaodan Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, Tao Jin

Visual Speech Recognition (VSR) aims to predict spoken content by analyzing lip movements in videos. Recently reported state-of-the-art results in VSR often rely on increasingly large amounts of video data, while the publicly available transcribed video datasets are insufficient compared to the audio data. To further enhance the VSR model using the audio data, we employed a generative model for data inflation, integrating the synthetic data with the authentic visual data. Essentially, the generative model incorporates another insight, which enhances the capabilities of the recognition model. For the cross-language issue, previous work has shown poor performance with non-Indo-European languages. We trained a multi-language-family modal fusion model, AudioVSR. Leveraging the concept of modal transfer, we achieved significant results in downstream VSR tasks under conditions of data scarcity. To the best of our knowledge, AudioVSR represents the first work on cross-language-family audio-lip alignment, achieving a new SOTA in the cross-language scenario.

(Nov 13): 7:45:45 (Morning) - Gather

GOME: Grounding-based Metaphor Binding With Conceptual Elaboration For Figurative Language Illustration

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, Guangluan Xu

The illustration or visualization of figurative language, such as linguistic metaphors, is an emerging challenge for existing Large Language Models (LLMs) and multimodal models. Due to their comparison of seemingly unrelated concepts in metaphors, existing LLMs have a tendency of over-literatization, which illustrates figurative language solely based on literal objects, ignoring the underlying groundings and associations across disparate metaphorical domains. Furthermore, prior approaches have ignored the binding process between visual objects and metaphorical attributes, which further intensifies the infidelity of visual metaphors. To address the issues above, we propose GOME (Grounding-based Metaphor Binding), which illustrates linguistic metaphors from the grounding perspective elaborated through LLMs. GOME consists of two steps for metaphor illustration, including grounding-based elaboration and scenario visualization. In the elaboration step, metaphorical knowledge is integrated into systematic instructions for LLMs, which employs a CoT prompting method rooted in rhetoric. This approach specifies metaphorical devices such as vehicles and groundings, to ensure accurate and faithful descriptions consumed by text-to-image models. In the visualization step, an inference-time metaphor binding method is realized based on elaboration outputs, which register attentional control during the diffusion process, and captures the underlying attributes from the abstract metaphorical domain. Comprehensive evaluations using multiple downstream tasks confirm that, GOME is superior to isolated LLMs, diffusion models, or their direct collaboration.

(Nov 13): 7:45:45 (Morning) - Gather

mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding

Anwen Hu, Haiyang Xu, Jiaobo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, Jingren Zhou

Structure information is critical for understanding the semantics of text-rich images, such as documents, tables, and charts. Existing Mul-

timodal Large Language Models (MLLMs) for Visual Document Understanding are equipped with text recognition ability but lack general structure understanding abilities for text-rich document images. In this work, we emphasize the importance of structure information in Visual Document Understanding and propose Unified Structure Learning to boost the performance of MLLMs. Based on publicly available text-rich images, we build a comprehensive training set DocStruct4M to support structure-aware parsing tasks and multi-grained text localization tasks across 5 domains: document, webpage, table, chart, and natural image. To better encode structure information, we design a simple and effective vision-to-text module H-Reducer, which can not only maintain the layout information but also reduce the length of visual features by merging horizontal adjacent patches through convolution, enabling the LLM to understand high-resolution images more efficiently. Our model DocOwl 1.5 achieves state-of-the-art performance on 10 visual document understanding benchmarks. All codes, models, and datasets are publicly available at <https://github.com/X-PLUG/mPLUG-DocOwl/tree/main/DocOwl1.5>.

(Nov 13): 7:458:45 (Morning) - Gather

Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training

Wenbo Li, Guohao Li, Zhibin Lan, Xue Xu, Wanru Zhuang, Jiachen Liu, Xinyan Xiao, Jinsong Su

Diffusion-based text-to-image models have demonstrated impressive achievements in diversity and aesthetics but struggle to generate images with legible visual texts. Existing backbone models have limitations such as misspelling, failing to generate texts, and lack of support for Chinese texts, but their development shows promising potential. In this paper, we propose a series of methods, aiming to empower backbone models to generate visual texts in English and Chinese. We first conduct a preliminary study revealing that BPE tokenization and insufficient learning of cross-attention modules restrict the performance of the backbone models. Based on these observations, we make the following improvements: (1) We design a mixed granularity input strategy to provide more suitable text representations; (2) We propose to augment the conventional training objective with three glyph-aware training losses, which enhance the learning of cross-attention modules and encourage the model to focus on visual texts. Through experiments, we demonstrate that our methods can effectively empower backbone models to generate semantic relevant, aesthetically appealing, and accurate visual text images, while maintaining their fundamental image generation quality.

(Nov 13): 7:458:45 (Morning) - Gather

How Does the Textual Information Affect the Retrieval of Multimodal In-Context Learning?

Yang Luo, Zangwei Zheng, Zirui Zhu, Yang You

The increase in parameter size of multimodal large language models (MLLMs) introduces significant capabilities, particularly multimodal in-context learning, where MLLMs enhance task performance without updating pre-trained parameters. However, this effectiveness hinges on the appropriate selection of in-context examples, a process currently biased towards visual data, overlooking textual information. More importantly, the area of supervised retrievers for retrieval of multimodal in-context learning, crucial for optimal in-context example selection, continues to be investigated. Our study provides an in-depth evaluation of the impact of textual information on the unsupervised selection of in-context examples in multimodal contexts, uncovering a notable sensitivity of retriever performance to the employed modalities. Based on the above finding, we introduce a novel supervised MLLM prompt retriever MSIER that leverages a trained retriever based on MLLM's confidence to select examples, which enhances multimodal in-context learning efficiency. This approach is validated through extensive testing across three different tasks, demonstrating the method's effectiveness. Additionally, we investigate the influence of modalities on our supervised retrieval method's training and explore the transferability of the supervised prompt retriever. This exploration paves the way for future advancements, highlighting the potential for refined in-context learning in MLLMs through the strategic use of multimodal data. The public code is available at <https://github.com/NUS-HPC-AI-Lab/Multimodal-ICL-Retriever>.

(Nov 13): 7:458:45 (Morning) - Gather

Divide and Conquer Radiology Report Generation via Observation Level Fine-grained Pretraining and Prompt Tuning

Yuanpin Zhou, Huogen Wang

The automation of radiology report generation (RRG) holds immense potential to alleviate radiologists' workloads and improve diagnostic accuracy. Despite advancements in image captioning and vision-language pretraining, RRG remains challenging due to the lengthy and complex nature of radiology reports. In this work, we propose the Divide and Conquer Radiology Report Generation (DCRRG) model, which breaks down full-text radiology reports into concise observation descriptions. This approach enables the model to capture fine-grained representations from each observation through a two-stage process: an encoding stage focusing on observation prediction tasks to learn fine-grained representations, and a decoding stage for integrating these descriptions into cohesive and comprehensive radiology reports. Experimental results on two benchmark datasets demonstrate that DCRRG achieves significant improvements across all evaluated metrics, underscoring its capability to generate semantically coherent and clinically accurate radiology reports.

NLP Applications

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

An Inversion Attack Against Obfuscated Embedding Matrix in Language Model Inference

Yu Lin, Qizhi Zhang, Quanwei Cai, Jue Hong, Wu Ye, Huiqi Liu, Bing Duan

With the rapidly-growing deployment of large language model (LLM) inference services, privacy concerns have arisen regarding to the user input data. Recent studies are exploring transforming user inputs to obfuscated embedded vectors, so that the data will not be eavesdropped by service providers. However, in this paper we show that again, without a solid and deliberate security design and analysis, such embedded vector obfuscation failed to protect users' privacy. We demonstrate the conclusion via conducting a novel inversion attack called Element-wise Differential Nearest Neighbor (EDNN) on the glide-reflection proposed in mishra2024sentinelllms, and the result showed that the original user input text can be 100% recovered from the obfuscated embedded vectors. We further analyze security requirements on embedding obfuscation and present several remedies to our proposed attack.

(Nov 13): 7:458:45 (Morning) - Gather

Take Off the Training Wheels! Progressive In-Context Learning for Effective Alignment

zhenyu liu, Dongfang Li, Xinshuo Hu, Xinpeng Zhao, Yibin Chen, Baotian Hu, Min Zhang

Recent studies have explored the working mechanisms of In-Context Learning (ICL). However, they mainly focus on classification and simple generation tasks, limiting their broader application to more complex generation tasks in practice. To address this gap, we investigate the impact of demonstrations on token representations within the practical alignment tasks. We find that the transformer embeds the task function learned from demonstrations into the separator token representation, which plays an important role in the generation of prior response tokens. Once the prior response tokens are determined, the demonstrations become redundant. Motivated by this finding, we propose an efficient Progressive In-Context Alignment (PICA) method consisting of two stages. In the first few-shot stage, the model generates several prior response

tokens via standard ICL while concurrently extracting the ICL vector that stores the task function from the separator token representation. In the following zero-shot stage, this ICL vector guides the model to generate responses without further demonstrations. Extensive experiments demonstrate that our PICA not only surpasses vanilla ICL but also achieves comparable performance to other alignment tuning methods. The proposed training-free method reduces the time cost (e.g., 5.45E) with improved alignment performance (e.g., 6.57+). Consequently, our work highlights the application of ICL for alignment and calls for a deeper understanding of ICL for complex generations. The code will be available at <https://github.com/HITSz-TMG/PICA>.

(Nov 13): 7:45:45 (Morning) - Gather

DockKD: Knowledge Distillation from LLMs for Open-World Document Understanding Models

Sungyun Kim, Haofu Liao, Srikanth Appalaraju, Peng Tang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, Vijay Mahadevan, Stefano Soatto

Visual document understanding (VDU) is a challenging task that involves understanding documents across various modalities (text and image) and layouts (forms, tables, etc.). This study aims to enhance generalizability of small VDU models by distilling knowledge from LLMs. We identify that directly prompting LLMs often fails to generate informative and useful data. In response, we present a new framework (called DocKD) that enriches the data generation process by integrating external document knowledge. Specifically, we provide an LLM with various document elements like key-value pairs, layouts, and descriptions, to elicit open-ended answers. Our experiments show that DocKD produces high-quality document annotations and surpasses the direct knowledge distillation approach that does not leverage external document knowledge. Moreover, student VDU models trained with solely DocKD-generated data is not only comparable to those trained with human-annotated data on in-domain tasks but also significantly excel them on out-of-domain tasks.

(Nov 13): 7:45:45 (Morning) - Gather

Multi-pass Decoding for Grammatical Error Correction

Xiaoying Wang, Lingling Mu, Jingyi Zhang, Hongfei Xu

Sequence-to-sequence (seq2seq) models achieve comparable or better grammatical error correction performance compared to sequence-to-edit (seq2edit) models. Seq2edit models normally iteratively refine the correction result, while seq2seq models decode only once without aware of subsequent tokens. Iteratively refining the correction results of seq2seq models via Multi-Pass Decoding (MPD) may lead to better performance. However, MPD increases the inference costs. Deleting or replacing corrections in previous rounds may lose useful information in the source input. We present an early-stop mechanism to alleviate the efficiency issue. To address the source information loss issue, we propose to merge the source input with the previous round correction result into one sequence. Experiments on the CoNLL-14 test set and BEA-19 test set show that our approach can lead to consistent and significant improvements over strong BART and T5 baselines (+1.80, +1.35, and +2.02 F0.5 for BART 12-2, large and T5 large respectively on CoNLL-14 and +2.99, +1.82, and +2.79 correspondingly on BEA-19), obtaining F0.5 scores of 68.41 and 75.36 on CoNLL-14 and BEA-19 respectively.

(Nov 13): 7:45:45 (Morning) - Gather

CommonIT: Commonality-Aware Instruction Tuning for Large Language Models via Data Partitions

Jun Rao, Xuebo Liu, Lian Lian, shengjun cheng, Yunjie Liao, Min Zhang

With instruction tuning, Large Language Models (LLMs) can enhance their ability to adhere to commands. Diverging from most works focusing on data mixing, our study concentrates on enhancing the model's capabilities from the perspective of data sampling during training. Drawing inspiration from the human learning process, where it is generally easier to master solutions to similar topics through focused practice on a single type of topic, we introduce a novel instruction tuning strategy termed CommonIT: Commonality-aware Instruction Tuning. Specifically, we cluster instruction datasets into distinct groups with three proposed metrics: Task, Embedding and Length. We ensure each training mini-batch, or "partition", consists solely of data from a single group, which brings about both data randomness across mini-batches and intra-batch data similarity. Rigorous testing on LLaMa models demonstrates CommonIT's effectiveness in enhancing the instruction-following capabilities of LLMs through IT datasets (FLAN, COT, and Alpaca) and models (LLaMa2-7B, Qwen2-7B, LLaMa 13B, and BLOOM 7B). CommonIT consistently boosts an average improvement of 2.1% on the general domain (i.e., the average score of Knowledge, Reasoning, Multilinguality and Coding) with the Length metric, and 5.2% on the special domain (i.e., GSM, Openfunctions and Code) with the Task metric, and 3.8% on the specific tasks (i.e., MMLU) with the Embedding metric. Code is available at <https://github.com/rfaoy7/CommonIT>.

(Nov 13): 7:45:45 (Morning) - Gather

Applying Contrastive Learning to Code Vulnerability Type Classification

Chen Ji, Su Yang, Hongyu Sun, Yuqing Zhang

Vulnerability classification is a crucial task in software security analysis, essential for identifying and mitigating potential security risks. Learning-based methods often perform poorly due to the long-tail distribution of vulnerability classification datasets. Recent approaches try to address the problem but treat each CWE class in isolation, ignoring their relationships. This results in non-scalable code vector representations, causing significant performance drops when handling complex real-world vulnerabilities. We propose a hierarchical contrastive learning framework for code vulnerability type classification to bring vector representations of related CWEs closer together. To address the issue of class collapse and enhance model robustness, we mix self-supervised contrastive learning loss into our loss function. Additionally, we employ max-pooling to enable the model to handle longer vulnerability code inputs. Extensive experiments demonstrate that our proposed framework outperforms state-of-the-art methods by 2.97–17.90% on accuracy and 0.98%–22.27% on weighted-F1, with even better performance on higher-quality datasets. We also utilize an ablation study to prove each component's contribution. These findings underscore the potential and advantages of our approach in the multi-class vulnerability classification task.

(Nov 13): 7:45:45 (Morning) - Gather

FIRST: Teach A Reliable Large Language Model Through Efficient Trustworthy Distillation

KaShun SHUM, Minrui Xu, Jianshu Zhang, Zixin CHEN, Shizhe Diao, Hanze Dong, Jipeng Zhang, Muhammad Omer Raza

Large language models (LLMs) have become increasingly prevalent in our daily lives, leading to an expectation for LLMs to be trustworthy — both accurate and well-calibrated (the prediction confidence should align with its ground truth correctness likelihood). Nowadays, fine-tuning has become the most popular method for adapting a model to practical usage by significantly increasing accuracy on downstream tasks. Despite the great accuracy it achieves, we found fine-tuning is still far away from satisfactory trustworthiness due to "tuning-induced mis-calibration". In this paper, we delve deeply into why and how mis-calibration exists in fine-tuned models, and how distillation can alleviate the issue. Then we further propose a brand new method named Efficient Trustworthy Distillation (FIRST), which utilizes a small portion of teacher's knowledge to obtain a reliable language model in a cost-efficient way. Specifically, we identify the "concentrated knowledge" phenomenon during distillation, which can significantly reduce the computational burden. Then we apply a "trustworthy maximization" process to optimize the utilization of this small portion of concentrated knowledge before transferring it to the student. Experimental results demonstrate the effectiveness of our method, where better accuracy (+2.3%) and less mis-calibration (-10%) are achieved on average across both in-domain and out-of-domain scenarios, indicating better trustworthiness.

(Nov 13): 7:45:45 (Morning) - Gather

Leveraging Context-aware Prompting for Commit Message Generation

Zhihua Jiang, Jianwei Chen, Dongning Rao, Guanghui Ye

Writing comprehensive commit messages is tedious yet important, because these messages describe changes of code, such as fixing bugs or adding new features. However, most existing methods focus on either only the changed lines or nearest context lines, without considering the effectiveness of selecting useful contexts. On the other hand, it is possible that introducing excessive contexts can lead to noise. To this end, we propose a code model COMMIT (Context-aware prOMpting based comMIT-message generaTion) in conjunction with a code dataset CODEC (COntext and metaData Enhanced Code dataset). Leveraging program slicing, CODEC consolidates code changes along with related contexts via property graph analysis. Further, utilizing CodeT5+ as the backbone model, we train COMMIT via context-aware prompt on CODEC. Experiments show that COMMIT can surpass all compared models including pre-trained language models for code (code-PLMs) such as CommitBART and large language models for code (code-LLMs) such as Code-LLaMa. Besides, we investigate several research questions (RQs), further verifying the effectiveness of our approach. We release the data and code at: <https://github.com/Jnunplab/COMMIT.git>.

(Nov 13): 7:45:45 (Morning) - Gather

Improving Knowledge Graph Completion with Structure-Aware Supervised Contrastive Learning

Jiaoshi Lin, Lijiang Wang, Xinyu Lu, Zhongtian Hu, Wei Zhang, Wenxuan Lu

Knowledge Graphs (KGs) often suffer from incomplete knowledge, which restricts their utility. Recently, Contrastive Learning (CL) has been introduced to Knowledge Graph Completion (KGC), significantly improving the discriminative capabilities of KGC models and setting new benchmarks in performance. However, existing contrastive methods primarily focus on individual triples, overlooking the broader structural connectivities and topologies of KGs. This narrow focus limits a comprehensive understanding of the graph's structural knowledge. To address this gap, we propose StructKGC, a novel contrastive learning framework designed to flexibly accommodate the diverse topologies inherent in KGs. Additionally, we introduce four contrastive tasks specifically tailored to KG data: Vertex-level CL, Neighbor-level CL, Path-level CL, and Relation composition level CL. These tasks are trained synergistically during the fine-tuning of pre-trained language models (PLMs), allowing for a more nuanced capture of subgraph semantics. To validate the effectiveness of our method, we perform a comprehensive set of experiments on several real-world datasets. The experimental results demonstrate that our approach achieves SOTA performance under standard supervised and low-resource settings. Furthermore, the different levels of structure-aware tasks introduced can mutually reinforce each other, leading to consistent performance improvements.

(Nov 13): 7:45:45 (Morning) - Gather

Human-LLM Hybrid Text Answer Aggregation for Crowd Annotations

Jiyi Li

The quality is a crucial issue for crowd annotations. Answer aggregation is an important type of solution. The aggregated answers estimated from multiple crowd answers to the same instance are the eventually collected annotations, rather than the individual crowd answers themselves. Recently, the capability of Large Language Models (LLMs) on data annotation tasks has attracted interest from researchers. Most of the existing studies mainly focus on the average performance of individual crowd workers; several recent works studied the scenarios of aggregation on categorical labels and LLMs used as label creators. However, the scenario of aggregation on text answers and the role of LLMs as aggregators are not yet well-studied. In this paper, we investigate the capability of LLMs as aggregators in the scenario of close-ended crowd text answer aggregation. We propose a human-LLM hybrid text answer aggregation method with a Creator-Aggregator Multi-Stage (CAMS) crowdsourcing framework. We make the experiments based on public crowdsourcing datasets. The results show the effectiveness of our approach based on the collaboration of crowd workers and LLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Free your mouse! Command Large Language Models to Generate Code to Format Word Documents

Shihao Rao, Liang Li, Jiapeng Liu, Guan Weinix, Xiyani Gao, bing lin

Recently, LLMs have significantly improved code generation, making it increasingly accessible to users. As a result, LLM-powered code generation applications have sprung up, vastly boosting user productivity. This paper mainly explores how to improve the efficiency and experience of users in formatting the document. Specifically, we propose an automatic document formatting method, Text-to-Format, which is driven by various prompting strategies. Text-to-Format takes the user's formatting instructions and then generates code that can be run in Microsoft Word to format the content in a document. Further, to evaluate automatic document formatting approaches and advance the document formatting task, we built an evaluation specification including a high-quality dataset DocFormEval data, a code runtime environment, and evaluation metrics. Extensive experimental results on data reveal that the prompting strategy's effect positively correlates with how much knowledge it introduces related to document formatting task. We believe the constructed DocFormEval data and the exploration about Text-to-Format can help developers build more intelligent tools for automatic document formatting, especially in offline scenarios, where the data privacy is the top priority.

(Nov 13): 7:45:45 (Morning) - Gather

ATAP: Automatic Template-Augmented Commonsense Knowledge Graph Completion via Pre-Trained Language Models

Fu Zhang, Yifan Ding, Jingwei Cheng

The mission of commonsense knowledge graph completion (CKGC) is to infer missing facts from known commonsense knowledge. CKGC methods can be roughly divided into two categories: triple-based methods and text-based methods. Due to the imbalanced distribution of entities and limited structural information, triple-based methods struggle with long-tail entities. Text-based methods alleviate this issue, but require extensive training and fine-tuning of language models, which reduces efficiency. To alleviate these problems, we propose ATAP, the first CKGC framework that utilizes automatically generated continuous prompt templates combined with pre-trained language models (PLMs). Moreover, ATAP uses a carefully designed new prompt template training strategy, guiding PLMs to generate optimal prompt templates for CKGC tasks. Combining the rich knowledge of PLMs with the template automatic augmentation strategy, ATAP effectively mitigates the long-tail problem and enhances CKGC performance. Results on benchmark datasets show that ATAP achieves state-of-the-art performance overall.

(Nov 13): 7:45:45 (Morning) - Gather

ClimRetriever: A Benchmarking Dataset for Information Retrieval from Corporate Climate Disclosures

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martin, Nicola Ranger, Markus Leippold

To handle the vast amounts of qualitative data produced in corporate climate communication, stakeholders increasingly rely on Retrieval Augmented Generation (RAG) systems. However, a significant gap remains in evaluating domain-specific information retrieval – the basis for answer generation. To address this challenge, this work simulates the typical tasks of a sustainability analyst by examining 30 sustainability reports with 16 detailed climate-related questions. As a result, we obtain a dataset with over 8.5K unique question-source-answer pairs labeled by different levels of relevance. Furthermore, we develop a use case with the dataset to investigate the integration of expert knowledge into information retrieval with embeddings. Although we show that incorporating expert knowledge works, we also outline the critical limitations of embeddings in knowledge-intensive downstream domains like climate change communication.

(Nov 13): 7:45:45 (Morning) - Gather

Can Large Language Models Enhance Predictions of Disease Progression? Investigating Through Disease Network Link Prediction

Haohui Lu, Usman Nasreen

Large Language Models (LLMs) have made significant strides in various tasks, yet their effectiveness in predicting disease progression remains relatively unexplored. To fill this gap, we use LLMs and employ advanced graph prompting and Retrieval-Augmented Generation (RAG) to predict disease comorbidity within disease networks. Specifically, we introduce a disease Comorbidity prediction model using LLM, named ComLLM, which leverages domain knowledge to enhance the prediction performance. Based on the comprehensive experimental results, ComLLM consistently outperforms conventional models, such as Graph Neural Networks, achieving average area under the curve (AUC) improvements of 10.70% and 6.07% over the best baseline models in two distinct disease networks. ComLLM is evaluated across multiple settings for disease progression prediction, employing various prompting strategies, including zero-shot, few-shot, Chain-of-Thought, graph prompting and RAG. Our results show that graph prompting and RAG enhance LLM performance in disease progression prediction tasks. ComLLM exhibits superior predictive capabilities and serves as a proof-of-concept for LLM-based systems in disease progression prediction, highlighting its potential for broad applications in healthcare.

(Nov 13): 7:45:45 (Morning) - Gather

SimLLM: Detecting Sentences Generated by Large Language Models Using Similarity between the Generation and its Re-generation

Hoang-Quoc Nguyen-Son, Minh-Son Dao, Koji Zettu

Large language models have emerged as a significant phenomenon due to their ability to produce natural text across various applications. However, the proliferation of generated text raises concerns regarding its potential misuse in fraudulent activities such as academic dishonesty, spam dissemination, and misinformation propagation. Prior studies have detected the generation of non-analogous text, which manifests numerous differences between original and generated text. We have observed that the similarity between the original text and its generation is notably higher than that between the generated text and its subsequent regeneration. To address this, we propose a novel approach named SimLLM, aimed at estimating the similarity between an input sentence and its generated counterpart to detect analogous machine-generated sentences that closely mimic human-written ones. Our empirical analysis demonstrates SimLLM's superior performance compared to existing methods.

(Nov 13): 7:45:45 (Morning) - Gather

Context-aware Watermark with Semantic Balanced Green-red Lists for Large Language Models

Yuxuan Guo, Zhiliang Tian, YIPING SONG, Tianlun Liu, Liang Ding, Dongsheng Li

Watermarking enables people to determine whether the text is generated by a specific model. It injects a unique signature based on the "green-red" list that can be tracked during detection, where the words in green lists are encouraged to be generated. Recent researchers propose to fix the green/red lists or increase the proportion of green tokens to defend against paraphrasing attacks. However, these methods cause degradation of text quality due to semantic disparities between the watermarked text and the unwatermarked text. In this paper, we propose a semantic-aware watermark method that considers contexts to generate a semantic-aware key to split a semantically balanced green/red list for watermark injection. The semantic balanced list reduces the performance drop due to adding bias on green lists. To defend against paraphrasing attacks, we generate the watermark key considering the semantics of contexts via locally sensitive hashing. To improve the text quality, we propose to split green/red lists considering semantics to enable the green list to cover almost all semantics. We also dynamically adapt the bias to balance text quality and robustness. The experiments show our advantages in both robustness and text quality comparable to existing baselines.

(Nov 13): 7:45:45 (Morning) - Gather

Toolken+: Improving LLM Tool Usage with Reranking and a Reject Option

Konstantin Yakovlev, Sergey Nikolenko, Andrey Bout

The recently proposed ToolkenGPT tool learning paradigm demonstrates promising performance but suffers from two major issues: first, it cannot benefit from tool documentation, and second, it often makes mistakes in whether to use a tool at all. We introduce Toolken+ that mitigates the first problem by reranking top-k tools selected by ToolkenGPT and the second problem with a special REJECT option such that the model will generate a vocabulary token if REJECT is ranked first. We demonstrate the effectiveness of Toolken+ on multistep numerical reasoning and tool selection tasks.

(Nov 13): 7:45:45 (Morning) - Gather

See Detail Say Clear: Towards Brain CT Report Generation via Pathological Clue-driven Representation Learning

Chengxin Zheng, Junzhong Ji, Yanzhao Shi, Xiaodan Zhang, Liangqiong Qu

Brain CT report generation is significant to aid physicians in diagnosing cranial diseases. Recent studies concentrate on handling the consistency between visual and textual pathological features to improve the coherence of report. However, there exist some challenges: 1) Redundant visual representing: Massive irrelevant areas in 3D scans distract models from representing salient visual contexts. 2) Shifted semantic representing: Limited medical corpus causes difficulties for models to transfer the learned textual representations to generative layers. This study introduces a Pathological Clue-driven Representation Learning (PCRL) model to build cross-modal representations based on pathological clues and naturally adapt them for accurate report generation. Specifically, we construct pathological clues from perspectives of segmented regions, pathological entities, and report themes, to fully grasp visual pathological patterns and learn cross-modal feature representations. To adapt the representations for the text generation task, we bridge the gap between representation learning and report generation by using a unified large language model (LLM) with task-tailored instructions. These crafted instructions enable the LLM to be flexibly fine-tuned across tasks and smoothly transfer the semantic representation for report generation. Experiments demonstrate that our method outperforms previous methods and achieves SoTA performance. Our code is available at <https://github.com/Chauncey-Jheng/PCRL-MRG>.

(Nov 13): 7:45:45 (Morning) - Gather

Robust AI-Generated Text Detection by Restricted Embeddings

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, Irina Piontovskaya

Growing amount and quality of AI-generated texts makes detecting such content more difficult. In most real-world scenarios, the domain (style and topic) of generated data and the generator model are not known in advance. In this work, we focus on the robustness of classifier-based detectors of AI-generated text, namely their ability to transfer to unseen generators or semantic domains. We investigate the geometry of the embedding space of Transformer-based text encoders and show that clearing out harmful linear subspaces helps to train a robust classifier, ignoring domain-specific spurious features. We investigate several subspace decomposition and feature selection strategies and achieve significant improvements over state of the art methods in cross-domain and cross-generator transfer. Our best approaches for head-wise and coordinate-based subspace removal increase the mean out-of-distribution (OOD) classification score by up to 9% and 14% in particular setups for RoBERTa and BERT embeddings respectively. We release our code and data: <https://github.com/SilverSolver/RobustATD>

(Nov 13): 7:45:45 (Morning) - Gather

Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering

Yubo Wang, Xueguang Ma, Wenhui Chen

Large-scale language models (LLMs) like ChatGPT have demonstrated impressive abilities in generating responses based on human instruc-

tions. However, their use in the medical field can be challenging due to their lack of specific, in-depth knowledge. In this study, we present a system called LLMs Augmented with Medical Textbooks (LLM-AMT) designed to enhance the proficiency of LLMs in specialized domains. LLM-AMT integrates authoritative medical textbooks into the LLMs' framework using plug-and-play modules. These modules include a Query Augmenter, a Hybrid Textbook Retriever, and a Knowledge Self-Refiner. Together, they incorporate authoritative medical knowledge. Additionally, an *LLM Reader* aids in contextual understanding. Our experimental results on three medical QA tasks demonstrate that LLM-AMT significantly improves response quality, with accuracy gains ranging from 11.6% to 16.6%. Notably, with GPT-4-Turbo as the base model, LLM-AMT outperforms the specialized Med-PaLM 2 model pre-trained on a massive amount of medical corpus by 2-3%. We found that despite being 100smaller in size, medical textbooks as a retrieval corpus are proven to be a more effective knowledge database than Wikipedia in the medical domain, boosting performance by 7.8%-13.7%.

(Nov 13): 7:458:45 (Morning) - Gather

Exploring Open Graph Models with Large Language Models

Lianghao Xia, Ben Kao, Chao Huang

Graph learning has become essential in various domains, including recommendation systems and social network analysis. Graph Neural Networks (GNNs) have emerged as promising techniques for encoding structural information and improving performance in tasks like link prediction and node classification. However, a key challenge remains: the difficulty of generalizing to unseen graph data with different properties. In this work, we propose a novel graph foundation model, called OpenGraph, to address this challenge. Our approach tackles several technical obstacles. Firstly, we enhance data augmentation using a large language model (LLM) to overcome data scarcity in real-world scenarios. Secondly, we introduce a unified graph tokenizer that enables the model to generalize effectively to diverse graph data, even when encountering unseen properties during training. Thirdly, our developed scalable graph transformer captures node-wise dependencies within the global topological context. Extensive experiments validate the effectiveness of our framework. By adapting OpenGraph to new graph characteristics and comprehending diverse graphs, our approach achieves remarkable zero-shot graph learning performance across various settings. We release the model implementation at <https://github.com/HKUDS/OpenGraph>.

(Nov 13): 7:458:45 (Morning) - Gather

MM-ChatAlign: A Novel Multimodal Reasoning Framework based on Large Language Models for Entity Alignment

Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Huang Zihé, Jian Guo, Yuanzhuo Wang

Multimodal entity alignment (MMEA) integrates multi-source and cross-modal knowledge graphs, a crucial yet challenging task for data-centric applications. Traditional MMEA methods derive the visual embeddings of entities and combine them with other modal data for alignment by embedding similarity comparison. However, these methods are hampered by the limited comprehension of visual attributes and deficiencies in realizing and bridging the semantics of multimodal data. To address these challenges, we propose MM-ChatAlign, a novel framework that utilizes the visual reasoning abilities of MLLMs for MMEA. The framework features an embedding-based candidate collection module that adapts to various knowledge representation strategies, effectively filtering out irrelevant reasoning candidates. Additionally, a reasoning and rethinking module, powered by MLLMs, enhances alignment by efficiently utilizing multimodal information. Extensive experiments on four MMEA datasets demonstrate MM-ChatAlign's superiority and underscore the significant potential of MLLMs in MMEA tasks. The source code is available at <https://github.com/jxh4945777/MMEA/>.

(Nov 13): 7:458:45 (Morning) - Gather

Divide and Conquer: Legal Concept-guided Criminal Court View Generation

Qi Xu, Xiao Wei, Hang Yu, Qian Liu, Hao Fei

The Criminal Court View Generation task aims to produce explanations that inform judicial decisions. This necessitates a nuanced understanding of diverse legal concepts, such as Recidivism, Confess, and Robbery, which often coexist within cases, complicating holistic analysis. However, existing methods mainly rely on the generation capability of language models, without paying enough attention to the important legal concepts. To enhance the precision and depth of such explanations, we introduce Legal Concept-guided Criminal Court Views Generation (LeGen), a three-stage approach designed for iterative reasoning tailored to individual legal constructs. Specifically, in the first stage, we design a decomposer to divide the court views into focused sub-views, each anchored around a distinct legal concept. Next, a concept reasoning module generates targeted rationales by intertwining the deconstructed facts with their corresponding legal frameworks, ensuring contextually relevant interpretations. Finally, a verifier and a generator are employed to align the rationale with the case fact and obtain synthesized comprehensive and legally sound final court views, respectively. We evaluate LeGen by conducting extensive experiments on a real-world dataset and experimental results validate the effectiveness of our proposed model. Our codes are available at <https://anonymous.4open.science/r/LeGen-5625>.

(Nov 13): 7:458:45 (Morning) - Gather

ProTrix: Building Models for Planning and Reasoning over Tables with Sentence Context

Zirui Wu, Yansong Feng

Tables play a crucial role in conveying information in various domains. We propose a Plan-then-Reason framework to answer different types of user queries over tables with sentence context. The framework first plans the reasoning paths over the context, then assigns each step to program-based or textual reasoning to reach the final answer. This framework enhances the table reasoning abilities for both in-context learning and fine-tuning methods. GPT-3.5-Turbo following Plan-then-Reason framework surpasses other prompting baselines without self-consistency while using less API calls and in-context demonstrations. We also construct an instruction tuning set TrixiInstruct to evaluate the effectiveness of fine-tuning with this framework. We present ProTrix model family by finetuning models on TrixiInstruct. Our experiments show that ProTrix family generalizes to diverse unseen tabular tasks with only 6k training instances. We further demonstrate that ProTrix can generate accurate and faithful explanations to answer complex free-form questions. Our work underscores the importance of the planning and reasoning abilities towards a model over tabular tasks with generalizability and interpretability. We will open-source our dataset and models at <https://anonymous.4open.science/r/LeGen-5625>.

(Nov 13): 7:458:45 (Morning) - Gather

GeoGPT4V: Towards Geometric Multi-modal Large Language Models with Geometric Image Generation

Shihao Cai, Kegin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, Bo Zheng

Large language models have seen widespread adoption in math problem-solving, yet for geometry problems, which often necessitate visual aids even for humans, the most advanced multi-modal models still struggle to effectively utilize image information. High-quality data is crucial for enhancing the geometric capabilities of multi-modal models, yet existing open-source datasets and related efforts are either too challenging for direct model learning or suffer from misalignment between text and images. To overcome this issue, we introduce a novel pipeline that leverages GPT-4 and GPT-4V to generate relatively basic geometry problems with aligned text and images, facilitating model learning. We have produced a dataset of 4.9K geometry problems and combined it with 19K open-source data to form our GeoGPT4V dataset. Experimental results demonstrate that the GeoGPT4V dataset significantly improves the geometry performance of various models on the MathVista and MathVision benchmarks. The code is available at <https://anonymous.4open.science/r/GeoGPT4V-08B2>.

(Nov 13): 7:458:45 (Morning) - Gather

Gold Panning in Vocabulary: An Adaptive Method for Vocabulary Expansion of Domain-Specific LLMs

Chengyuan Liu, Shihang Wang, Lizi Qing, Kun Kuang, Yangyang Kang, Changlong Sun, Fei Wu

While Large Language Models (LLMs) demonstrate impressive generation abilities, they frequently struggle when it comes to specialized domains due to their limited domain-specific knowledge. Studies on domain-specific LLMs resort to expanding the vocabulary before fine-tuning on domain-specific corpus, aiming to decrease the sequence length and enhance efficiency during decoding, without thoroughly investigating the results of vocabulary expansion to LLMs over different domains. Our pilot study reveals that expansion with only a subset of the entire vocabulary may lead to superior performance. Guided by the discovery, this paper explores how to identify a vocabulary subset to achieve the optimal results. We introduce VEGAD, an adaptive method that automatically identifies valuable words from a given domain vocabulary. Our method has been validated through experiments on three Chinese datasets, demonstrating its effectiveness. Additionally, we have undertaken comprehensive analyses of the method. The selection of a optimal subset for expansion has shown to enhance performance on both domain-specific tasks and general tasks, showcasing the potential of VEGAD.

(Nov 13): 7:45:45 (Morning) - Gather

More Than Catastrophic Forgetting: Integrating General Capabilities For Domain-Specific LLMs

Chengyuan Liu, Shihang Wang, Yangyang Kang, Lizi Qing, Fabang Zhao, Chao Wu, Changlong Sun, Kun Kuang, Fei Wu

The performance on general tasks decreases after Large Language Models (LLMs) are fine-tuned on domain-specific tasks, the phenomenon is known as Catastrophic Forgetting (CF). However, this paper presents a further challenge for real application of domain-specific LLMs beyond CF, called General Capabilities Integration (GCI), which necessitates the integration of both the general capabilities and domain knowledge within a single instance. The objective of GCI is not merely to retain previously acquired general capabilities alongside new domain knowledge, but to harmonize and utilize both sets of skills in a cohesive manner to enhance performance on domain-specific tasks. Taking legal domain as an example, we carefully design three groups of training and testing tasks without lacking practicability, and construct the corresponding datasets. To better incorporate general capabilities across domain-specific scenarios, we introduce ALoRA, which utilizes a multi-head attention module upon LoRA, facilitating direct information transfer from preceding tokens to the current one. This enhancement permits the representation to dynamically switch between domain-specific knowledge and general competencies according to the attention. Extensive experiments are conducted on the proposed tasks. The results exhibit the significance of our setting, and the effectiveness of our method.

(Nov 13): 7:45:45 (Morning) - Gather

MoCoKG: Momentum Contrast Entity Encoding for Knowledge Graph Completion

Qingyang Li, Yanru Zhong, Yuchu Qin

In recent years, numerous studies have sought to enhance the capabilities of pretrained language models (PLMs) for Knowledge Graph Completion (KGC) tasks by integrating structural information from knowledge graphs. However, existing approaches have not effectively combined the structural attributes of knowledge graphs with the textual descriptions of entities to generate robust entity encodings. To address this issue, this paper proposes MoCoKG (Momentum Contrast Entity Encoding for Knowledge Graph Completion), which incorporates three primary encoders: the entity-relation encoder, the entity encoder, and the momentum entity encoder. Momentum contrastive learning not only provides more negative samples but also allows for the gradual updating of entity encodings. Consequently, we reintroduce the generated entity encodings into the encoder to incorporate the graph's structural information. Additionally, MoCoKG enhances the inferential capabilities of the entity-relation encoder through deep prompts of relations. On the standard evaluation metric, Mean Reciprocal Rank (MRR), the MoCoKG model demonstrates superior performance, achieving a 7.1% improvement on the WN18RR dataset and an 11% improvement on the Wikidata5M dataset, while also surpassing the current best model on the FB15k-237 dataset. Through a series of experiments, this paper thoroughly examines the role and contribution of each component and parameter of the model.

(Nov 13): 7:45:45 (Morning) - Gather

BAPO: Base-Anchored Preference Optimization for Personalized Alignment in LLMs

Gihun Lee, Minchan Jeong, Yujin Kim, Hojung Jung, Jaehoon Oh, SangMook Kim, Se-Young Yun

While learning to align Large Language Models (LLMs) with human preferences has shown remarkable success, aligning these models to meet the diverse user preferences presents further challenges in preserving previous knowledge. This paper examines the impact of personalized preference optimization on LLMs, revealing that the extent of knowledge loss varies significantly with preference heterogeneity. Although previous approaches have utilized the KL constraint between the reference model and the policy model, we observe that they fail to maintain general knowledge and alignment when facing personalized preferences. To this end, we introduce Base-Anchored Preference Optimization (BAPO), a simple yet effective approach that utilizes the initial responses of reference model to mitigate forgetting while accommodating personalized alignment. BAPO effectively adapts to diverse user preferences while minimally affecting global knowledge or general alignment. Our experiments demonstrate the efficacy of BAPO in various setups.

(Nov 13): 7:45:45 (Morning) - Gather

ExpertEase: A Multi-Agent Framework for Grade-Specific Document Simplification with Large Language Models

Kaijie Mo, Renfen Hu

Text simplification is crucial for making texts more accessible, yet current research primarily focuses on sentence-level simplification, neglecting document-level simplification and the different reading levels of target audiences. To bridge these gaps, we introduce ExpertEase, a multi-agent framework for grade-specific document simplification using Large Language Models (LLMs). ExpertEase simulates real-world text simplification by introducing expert, teacher, and student agents that cooperate on the task and rely on external tools for calibration. Experiments demonstrate that this multi-agent approach significantly enhances LLMs' ability to simplify reading materials for diverse audiences. Furthermore, we evaluate the performance of LLMs varying in size and type, and compare LLM-generated texts with human-authored ones, highlighting their potential in educational resource development and guiding future research.

(Nov 13): 7:45:45 (Morning) - Gather

Learning to Plan by Updating Natural Language

Yiduo Guo, Yaobo Liang, Chenfei Wu, Wenshan Wu, Dongyan Zhao, Nan Duan

Large Language Models (LLMs) have shown remarkable performance in various basic natural language tasks. For completing the complex task, we still need a plan for the task to guide LLMs to generate the specific solutions step by step. LLMs can directly generate task plans, but these plans may still contain factual errors or are incomplete. A high-quality task plan contains correct step-by-step solutions for solving all situations and behavioral instructions for avoiding mistakes. To obtain it, we propose the Learning to Plan method, which involves two phases: (1) In the first learning task plan phase, it iteratively updates the task plan with new step-by-step solutions and behavioral instructions, which are obtained by prompting LLMs to derive from training error feedback. (2) In the subsequent test phase, the LLM uses the learned task plan to guide the inference of LLM on the test set. We demonstrate the effectiveness of our method on the five different reasoning type tasks (8 datasets). Further, our analysis experiment shows that the task plan learned by one LLM can directly guide another LLM to improve its performance, which reveals a new transfer learning paradigm.

(Nov 13): 7:45:45 (Morning) - Gather

Denoising Rationalization for Multi-hop Fact Verification via Multi-granular Explainer

Jiasheng Si, Yingjie Zhu, Wenpeng Lu, Deyu Zhou

The success of deep learning models on multi-hop fact verification has prompted researchers to understand the behavior behind their veracity. One feasible way is erasure search: obtaining the rationale by entirely removing a subset of input without compromising verification accuracy. Despite extensive exploration, current rationalization methods struggle to discern nuanced composition within the correlated evidence, which inevitably leads to noise rationalization in multi-hop scenarios. To address this issue, this paper explores the multi-granular rationale extraction method, aiming to realize the denoising rationalization for multi-hop fact verification. Specifically, given a pretrained veracity prediction model, two independent external explainers are introduced and trained collaboratively to enhance the discriminating ability by imposing varied constraints. Meanwhile, three key properties (Fidelity, Consistency, Salience) are introduced to regularize the denoising and faithful rationalization process. Additionally, a new Noiselessness metric is proposed to measure the purity of the rationales. Experimental results on three multi-hop fact verification datasets show that the proposed approach outperforms 12 baselines.

(Nov 13): 7:458:45 (Morning) - Gather

Reasoning Paths Optimization: A Framework For Exploring And Learning From Diverse Reasoning Paths

Yew Ken Chia, Guizhen Chen, Weiwen Xu, Anh Tuan Luu, Soujanya Poria, Lidong Bing

Advanced models such as OpenAI exhibit impressive problem-solving capabilities through step-by-step reasoning. However, they may still falter on more complex problems, making errors that disrupt their reasoning paths. We attribute this to the expansive solution space, where each step has the risk of diverging into mistakes. To enhance language model reasoning, we introduce a specialized training framework called Reasoning Paths Optimization (RPO), which enables learning to reason and explore from diverse paths. Our approach encourages favorable branches at each reasoning step while penalizing unfavorable ones, enhancing the model's overall problem-solving performance. Reasoning Paths Optimization does not rely on large-scale human-annotated rationales or outputs from closed-source models, making it scalable and data-efficient. We focus on multi-step reasoning tasks, such as math word problems and science-based exam questions. The experiments demonstrate that our framework significantly enhances the reasoning performance of large language models, with up to 3.1% and 4.3% improvement on GSM8K and MMLU (STEM) respectively. Our data and code can be found at <https://reasoning-paths.github.io>.

(Nov 13): 7:458:45 (Morning) - Gather

A Reflective LLM-based Agent to Guide Zero-shot Cryptocurrency Trading

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, Bingsheng He

The utilization of Large Language Models (LLMs) in financial trading has primarily been concentrated within the stock market, aiding in economic and financial decisions. Yet, the unique opportunities presented by the cryptocurrency market, noted for its on-chain data's transparency and the critical influence of off-chain signals like news, remain largely untapped by LLMs. This work aims to bridge the gap by developing an LLM-based trading agent, CryptoTrade, which uniquely combines the analysis of on-chain and off-chain data. This approach leverages the transparency and immutability of on-chain data, as well as the timeliness and influence of off-chain signals, providing a comprehensive overview of the cryptocurrency market. CryptoTrade incorporates a reflective mechanism specifically engineered to refine its daily trading decisions by analyzing the outcomes of prior trading decisions. This research makes two significant contributions. Firstly, it broadens the applicability of LLMs to the domain of cryptocurrency trading. Secondly, it establishes a benchmark for cryptocurrency trading strategies. Through extensive experiments, CryptoTrade has demonstrated superior performance in maximizing returns compared to time-series baselines, but not compared to traditional trading signals, across various cryptocurrencies and market conditions. Our code and data are available at <https://github.com/Xtra-Computing/CryptoTrade>

(Nov 13): 7:458:45 (Morning) - Gather

A Simple yet Effective Training-free Prompt-free Approach to Chinese Spelling Correction Based on Large Language Models

Houqian Zhou, Zhenghua Li, Bo Zhang, Chen Li, Shaopeng Lai, Ji Zhang, Fei Huang, Min Zhang

This work proposes a simple training-free prompt-free approach to leverage large language models (LLMs) for the Chinese spelling correction (CSC) task, which is totally different from all previous CSC approaches. The key idea is to use an LLM as a pure language model in a conventional manner. The LLM goes through the input sentence from the beginning, and at each inference step, produces a distribution over its vocabulary for deciding the next token, given a partial sentence. To ensure that the output sentence remains faithful to the input sentence, we design a minimal distortion model that utilizes pronunciation or shape similarities between the original and replaced characters. Furthermore, we propose two useful reward strategies to address practical challenges specific to the CSC task. Experiments on five public datasets demonstrate that our approach significantly improves LLM performance, enabling them to compete with state-of-the-art domain-general CSC models.

Question Answering

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

Making Large Language Models Better Reasoners with Orchestrated Streaming Experiences

Xiangyang Liu, Junliang He, Xipeng Qiu

Large language models (LLMs) can perform complex reasoning by generating intermediate reasoning steps using chain-of-thought prompting under zero-shot or few-shot settings. However, zero-shot prompting always encounters low performance, and the superior performance of few-shot prompting hinges on the manual-crafting of task-specific demonstrations one by one. In this paper, we present **RoSE** (**R**easoning with **O**rchestrated **S**treaming **E**xperiences), a general framework for solving reasoning tasks that can self-improve as it answers various reasoning questions. To enable RoSE, we describe an architecture that extends an LLM to store all answered reasoning questions and their reasoning steps in a streaming experience pool and orchestrates helpful questions from the pool to assist itself in answering new questions. To set up a question-aware orchestration mechanism, RoSE first calculates the similarity of each question in the pool with the question to be answered. Since the solution to each question in the experience pool is not always correct, RoSE will sort the questions according to their similarity with the question to be answered, and then uniformly divide them into multiple buckets. It finally extracts one question from each bucket to make the extracted questions more diverse. To make the extracted questions help RoSE answer new questions as much as possible, we introduce two other attributes of uncertainty and complexity for each question. RoSE will preferentially select the questions with low uncertainty and high complexity from each bucket. We evaluate the versatility of RoSE in various complex reasoning tasks and LLMs, such as arithmetic and commonsense reasoning, and find that it can achieve excellent performance without any labeled data and pre-set unlabeled data.

(Nov 13): 7:458:45 (Morning) - Gather

MAR: Matching-Augmented Reasoning for Enhancing Visual-based Entity Question Answering

Zhengxuan Zhang, Yin Wu, Yuyu Luo, Nan Tang

A multimodal large language model MLLMs may struggle with answering visual-based (personal) entity questions (VEQA), such as "who is A?" or "who is that B is talking to?" for various reasons, e.g., the absence of the name of A in the caption or the inability of MLLMs to recognize A, particularly for less common entities. Furthermore, even if the MLLMs can identify A, it may refrain from answering due to privacy concerns. In this paper, we introduce a novel method called Matching-Augmented Reasoning (MAR) to enhance VEQA. Given a collection of visual objects with captions, MAR preprocesses each object individually, identifying faces, names, and their alignments within the object. It encodes this information and stores their vector representations in vector databases. When handling VEQA, MAR retrieves matching faces and names and organizes these entities into a matching graph. MAR then derives the answer to the query by reasoning over this matching graph. Extensive experiments show that MAR significantly improves VEQA compared with the state-of-the-art methods using MLLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering

Dongze Hao, Qunbo Wang, Longteng Guo, Jie Jiang, Jing Liu

While large pre-trained visual-language models have shown promising results on traditional visual question answering benchmarks, it is still challenging for them to answer complex VQA problems which require diverse world knowledge. Motivated by the research of retrieval-augmented generation in the field of natural language processing, we use Dense Passage Retrieval (DPR) to retrieve related knowledge to help the model answer questions. However, DPR conduct retrieving in natural language space, which may not ensure comprehensive acquisition of image information. Thus, the retrieved knowledge is not truly conducive to helping answer the question, affecting the performance of the overall system. To address this issue, we propose a novel framework that leverages the visual-language model to select the key knowledge retrieved by DPR and answer questions. The framework consists of two modules: Selector and Answerer, where both are initialized by the MLLM and parameter-efficiently finetuned by self-bootstrapping: find key knowledge in the retrieved knowledge documents using the Selector, and then use them to finetune the Answerer to predict answers; obtain the pseudo-labels of key knowledge documents based on the predictions of the Answerer and weak supervision labels, and then finetune the Selector to select key knowledge; repeat. Our framework significantly enhances the performance of the baseline on the challenging open-domain Knowledge-based VQA benchmark, OK-VQA, achieving a state-of-the-art accuracy of 62.83%.

(Nov 13): 7:45:45 (Morning) - Gather

TimeR⁴ : Time-aware Rewrite-Augmented Large Language Models for Temporal Knowledge Graph Question Answering

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, Kehui Song

Temporal Knowledge Graph Question Answering (TKGQA) aims to answer temporal questions using knowledge in Temporal Knowledge Graphs (TKGs). Previous works employ pre-trained TKG embeddings or graph neural networks to incorporate the knowledge of TKGs. However, these methods fail to fully understand the complex semantic information of time constraints in questions. In contrast, Large Language Models (LLMs) have shown exceptional performance in knowledge graph reasoning, unifying both semantic understanding and structural reasoning. To further enhance LLMs temporal reasoning ability, this paper aims to integrate relevant temporal knowledge from TKGs into LLMs through a Time-aware Retrieve-Rewrite-Retrieve-Rerank framework, which we named TimeR⁴. Specifically, to reduce temporal hallucination in LLMs, we propose a retrieve-rewrite module to rewrite questions using background knowledge stored in the TKGs, thereby acquiring explicit time constraints. Then, we implement a retrieve-rerank module aimed at retrieving semantically and temporally relevant facts from the TKGs and reranking them according to the temporal constraints. To achieve this, we fine-tune a retriever using the contrastive time-aware learning framework. Our approach achieves great improvements, with relative gains of 47.8% and 22.5% on two datasets, underscoring its effectiveness in boosting the temporal reasoning abilities of LLMs. Our code is available at <https://github.com/qianxinying/TimeR4>.

(Nov 13): 7:45:45 (Morning) - Gather

Python Is Not Always the Best Choice: Embracing Multilingual Program of Thoughts

Xianzhen Luo, Qinshi Zhu, Zhiming Zhang, Libo Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, Wanxiang Che

Program of Thoughts (PoT) is an approach characterized by its executable intermediate steps, which ensure the accuracy of the logical calculations in the reasoning process. Currently, PoT primarily uses Python. However, relying solely on a single language may result in suboptimal solutions and overlook the potential benefits of other programming languages. In this paper, we conduct comprehensive experiments on the programming languages used in PoT and find that no single language consistently delivers optimal performance across all tasks and models. The effectiveness of each language varies depending on the specific scenarios. Inspired by this, we propose a task and model agnostic approach called MultiPoT, which harnesses strength and diversity from various languages. Experimental results reveal that it significantly outperforms Python Self-Consistency. Furthermore, it achieves comparable or superior performance compared to the best monolingual PoT in almost all tasks across all models. In particular, MultiPoT achieves more than 4.6% improvement on average on ChatGPT (gpt-3.5-turbo-0701).

(Nov 13): 7:45:45 (Morning) - Gather

Position Engineering: Boosting Large Language Models through Positional Information Manipulation

Zhiyuan He, Huiqiang Jiang, Zilong Wang, Yiqing Yang, Luna K. Qiu, Lili Qiu

The performance of large language models (LLMs) is significantly influenced by the quality of the prompts provided. In response, researchers have developed enormous prompt engineering strategies aimed at modifying the prompt text to enhance task performance. In this paper, we introduce a novel technique termed position engineering, which offers a more efficient way to guide large language models. Unlike prompt engineering, which requires substantial effort to modify the text provided to LLMs, position engineering merely involves altering the positional information in the prompt without modifying the text itself. We have evaluated position engineering in two widely-used LLM scenarios: retrieval-augmented generation (RAG) and in-context learning (ICL). Our findings show that position engineering substantially improves upon the baseline in both cases. Position engineering thus represents a promising new strategy for exploiting the capabilities of large language models.

(Nov 13): 7:45:45 (Morning) - Gather

Fine-tuning Smaller Language Models for Question Answering over Financial Documents

Karmvir Singh Phogat, Sai Akhil Puranam, Srihar Dasaratha, Chetan Harsha, Shashishekhar Ramakrishna

Recent research has shown that smaller language models can acquire substantial reasoning abilities when fine-tuned with reasoning exemplars crafted by a significantly larger teacher model. We explore this paradigm for the financial domain, focusing on the challenge of answering questions that require multi-hop numerical reasoning over financial texts. We assess the performance of several smaller models that have been fine-tuned to generate programs that encode the required financial reasoning and calculations. Our findings demonstrate that these fine-tuned smaller models approach the performance of the teacher model. To provide a granular analysis of model performance, we propose an approach to investigate the specific student model capabilities that are enhanced by fine-tuning. Our empirical analysis indicates that fine-tuning refines the student models ability to express and apply the required financial concepts along with adapting the entity extraction for the specific data format. In addition, we hypothesize and demonstrate that comparable financial reasoning capability can be induced using relatively smaller datasets.

(Nov 13): 7:45:45 (Morning) - Gather

Correct after Answer: Enhancing Multi-Span Question Answering with Post-Processing Method**Jiayi Lin, Chenyang Zhang, Haibo Tong, Dongyu Zhang, Qingqing Hong, Bingxuan Hou, Junli Wang**

Multi-Span Question Answering (MSQA) requires models to extract one or multiple answer spans from a given context to answer a question. Prior work mainly focuses on designing specific methods or applying heuristic strategies to encourage models to predict more correct predictions. However, these models are trained on gold answers and fail to consider the incorrect predictions. Through a statistical analysis, we observe that models with stronger abilities do not predict less incorrect predictions compared with other models. In this work, we propose Answering-Classifying-Correcting (ACC) framework, which employs a post-processing strategy to handle incorrect predictions. Specifically, the ACC framework first introduces a **classifier** to classify the predictions into three types and exclude "wrong predictions", then introduces a **corrector** to modify "partially correct predictions". Experiments on several MSQA datasets show that ACC framework significantly improves the Exact Match (EM) scores, and further analysis demonstrates that ACC framework efficiently reduces the number of incorrect predictions, improving the quality of predictions. Our code and data are available at <https://github.com/TongjiNLP/ACC>.

(Nov 13): 7:45:45 (Morning) - Gather

Context-Driven Index Trimming: A Data Quality Perspective to Enhancing Precision of RALMs**Kexin Ma, Ruochun Jin, Wang Haotian, Wang Xi, Huan Chen, Yuhua Tang, Qian Wang**

Retrieval-Augmented Large Language Models (RALMs) have made significant strides in enhancing the accuracy of generated responses. However, existing research often overlooks the data quality issues within retrieval results, often caused by inaccurate existing vector-distance-based retrieval methods. We propose to boost the precision of RALMs answers from a data quality perspective through the Context-Driven Index Trimming (CDIT) framework, where Context Matching Dependencies (CMDs) are employed as logical data quality rules to capture and regulate the consistency between retrieved contexts. Based on the semantic comprehension capabilities of Large Language Models (LLMs), CDIT can effectively identify and discard retrieval results that are inconsistent with the query context and further modify indexes in the database, thereby improving answer quality. Experiments demonstrate average improvement of 3.75% in accuracy on challenging open-domain question-answering tasks. Also, the flexibility of CDIT is verified through its compatibility with various language models and indexing methods, which offers a promising approach to bolster RALMs data quality and retrieval precision jointly.

(Nov 13): 7:45:45 (Morning) - Gather

Exploring Union and Intersection of Visual Regions for Generating Questions, Answers, and Distractors**Wenjian Ding, YAO ZHANG, Jun Wang, Adam Jatowt, Zhenglu Yang**

Multiple-choice visual question answering (VQA) is to automatically choose a correct answer from a set of choices after reading an image. Existing efforts have been devoted to a separate generation of an image-related question, a correct answer, or challenge distractors. By contrast, we turn to a holistic generation and optimization of questions, answers, and distractors (QADs) in this study. This integrated generation strategy eliminates the need for human curation and guarantees information consistency. Furthermore, we first propose to put the spotlight on different image regions to diversify QADs. Accordingly, a novel framework ReBo is formulated in this paper. ReBo cyclically generates each QAD based on a recurrent multimodal encoder, and each generation is focusing on a different area of the image compared to those already concerned by the previously generated QADs. In addition to traditional VQA comparisons with state-of-the-art approaches, we also validate the capability of ReBo in generating augmented data to benefit VQA models.

(Nov 13): 7:45:45 (Morning) - Gather

Advancing Large Language Model Attribution through Self-Improving**Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin**

Teaching large language models (LLMs) to generate text with citations to evidence sources can mitigate hallucinations and enhance verifiability in information-seeking systems. However, improving this capability requires high-quality attribution data, which is costly and labor-intensive. Inspired by recent advances in self-improvement that enhance LLMs without manual annotation, we present START, a Self-Taught AttrIBUTion framework for iteratively improving the attribution capability of LLMs. First, to prevent models from stagnating due to initially insufficient supervision signals, START leverages the model to self-construct synthetic training data for warming up. To further self-improve the model's attribution ability, START iteratively utilizes fine-grained preference supervision signals constructed from its sampled responses to encourage robust, comprehensive, and attributable generation. Experiments on three open-domain question-answering datasets, covering long-form QA and multi-step reasoning, demonstrate significant performance gains of 25.13% on average without relying on human annotations and more advanced models. Further analysis reveals that START excels in aggregating information across multiple sources.

(Nov 13): 7:45:45 (Morning) - Gather

DVD: Dynamic Contrastive Decoding for Knowledge Amplification in Multi-Document Question Answering**Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, Zhijiang Guo**

Large language models (LLMs) are widely used in question-answering (QA) systems but often generate information with hallucinations. Retrieval-augmented generation (RAG) offers a potential remedy, yet the uneven retrieved quality and irrelevant contents may distract LLMs. In this work, we address these issues at the generation phase by treating RAG as a multi-document QA task. We propose a novel decoding strategy, Dynamic Contrastive Decoding, which dynamically amplifies knowledge from selected documents during the generation phase. method involves constructing inputs batchwise, designing new selection criteria to identify documents worth amplifying, and applying contrastive decoding with a specialized weight calculation to adjust the final logits used for sampling answer tokens. Zero-shot experimental results on ALCE-ASQA, NQ, TQA and PopQA benchmarks show that our method outperforms other decoding strategies. Additionally, we conduct experiments to validate the effectiveness of our selection criteria, weight calculation, and general multi-document scenarios. Our method requires no training and can be integrated with other methods to improve the RAG performance. Our codes will be publicly available at https://github.com/JulieJin-km/Dynamic_Contrastive_Decoding.

(Nov 13): 7:45:45 (Morning) - Gather

CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity**Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, Peter Iszak**

State-of-the-art performance in QA tasks is currently achieved by systems employing Large Language Models (LLMs), however these models tend to hallucinate information in their responses. One approach focuses on enhancing the generation process by incorporating attribution from the given input to the output. However, the challenge of identifying appropriate attributions and verifying their accuracy against a source is a complex task that requires significant improvements in assessing such systems. We introduce an attribution-oriented Chain-of-Thought reasoning method to enhance the accuracy of attributions. This approach focuses the reasoning process on generating an attribution-centric output. Evaluations on two context enhanced question-answering datasets using GPT-4 demonstrate improved accuracy and correctness of attributions. In addition, the combination of our method with finetuning enhances the response and attribution accuracy of two smaller LLMs, showing their potential to outperform GPT-4 in some cases.

(Nov 13): 7:45:45 (Morning) - Gather

Improving Zero-shot LLM Re-Ranker with Risk Minimization

Xiaowei Yuan, Zhao Yang, Yequan Wang, Jun Zhao, Kang Liu

In the Retrieval-Augmented Generation (RAG) system, advanced Large Language Models (LLMs) have emerged as effective Query Likelihood Models (QLMs) in an unsupervised way, which re-rank documents based on the probability of generating the query given the content of a document. However, directly prompting LLMs to approximate QLMs inherently is biased, where the estimated distribution might diverge from the actual document-specific distribution. In this study, we introduce a novel framework, UR³, which leverages Bayesian decision theory to both quantify and mitigate this estimation bias. Specifically, UR³ reformulates the problem as maximizing the probability of document generation, thereby harmonizing the optimization of query and document generation probabilities under a unified risk minimization objective. Our empirical results indicate that UR³ significantly enhances re-ranking, particularly in improving the Top-1 accuracy. It benefits the QA tasks by achieving higher accuracy with fewer input documents.

(Nov 13): 7:45:45 (Morning) - Gather

PCQPR: Proactive Conversational Question Planning with Reflection

Shasha Guo

Conversational Question Generation (CQG) enhances the interactivity of conversational question-answering systems in fields such as education, customer service, and entertainment. However, traditional CQG, focusing primarily on the immediate context, lacks the conversational foresight necessary to guide conversations toward specified conclusions. This limitation significantly restricts their ability to achieve conclusion-oriented conversational outcomes. In this work, we redefine the CQG task as Conclusion-driven Conversational Question Generation (CCQG) by focusing on proactivity, not merely reacting to the unfolding conversation but actively steering it towards a conclusion-oriented question-answer pair. To address this, we propose a novel approach, called Proactive Conversational Question Planning with self-Refining (PCQPR). Concretely, by integrating a planning algorithm inspired by Monte Carlo Tree Search (MCTS) with the analytical capabilities of large language models (LLMs), PCQPR predicts future conversation turns and continuously refines its questioning strategies. This iterative self-refining mechanism ensures the generation of contextually relevant questions strategically devised to reach a specified outcome. Our extensive evaluations demonstrate that PCQPR significantly surpasses existing CQG methods, marking a paradigm shift towards conclusion-oriented conversational question-answering systems.

(Nov 13): 7:45:45 (Morning) - Gather

LongRAG: A Dual-perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, Jie Tang

Long-Context Question Answering (LCQA), a challenging task, aims to reason over long-context documents to yield accurate answers to questions. Existing long-context Large Language Models (LLMs) for LCQA often struggle with the "lost in the middle" issue. Retrieval-Augmented Generation (RAG) mitigates this issue by providing external factual evidence. However, its chunking strategy disrupts the global long-context information, and its low-quality retrieval in long contexts hinders LLMs from identifying effective factual details due to substantial noise. To this end, we propose LongRAG, a general, dual-perspective, and robust LLM-based RAG system paradigm for LCQA to enhance RAGs' understanding of complex long-context knowledge (i.e., global information and factual details). We design LongRAG as a plug-and-play paradigm, facilitating adaptation to various domains and LLMs. Extensive experiments on three multi-hop datasets demonstrate that LongRAG significantly outperforms long-context LLMs (up by 6.94%), advanced RAG (up by 6.16%), and Vanilla RAG (up by 17.25%). Furthermore, we conduct quantitative ablation studies and multi-dimensional analyses, highlighting the effectiveness of the system's components and fine-tuning strategies. Data and code are available at <https://github.com/QingFei1/LongRAG>.

Resources and Evaluation

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

To Forget or Not? Towards Practical Knowledge Unlearning for Large Language Models

Bozhong Tian, Xiaozhan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, Ningyu Zhang

Large Language Models (LLMs) trained on extensive corpora inevitably retain sensitive data, such as personal privacy information and copyrighted material. Recent advancements in knowledge unlearning involve updating LLM parameters to erase specific knowledge. However, current unlearning paradigms are mired in vague forgetting boundaries, often erasing knowledge indiscriminately. In this work, we introduce KnowUndo, a benchmark containing copyrighted content and user privacy domains to evaluate if the unlearning process inadvertently erases essential knowledge. Our findings indicate that existing unlearning methods often suffer from excessive unlearning. To address this, we propose a simple yet effective method, MemFlex, which utilizes gradient information to precisely target and unlearn sensitive parameters. Experimental results show that MemFlex is superior to existing methods in both precise knowledge unlearning and general knowledge retaining of LLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Cross-Domain Audio Deepfake Detection: Dataset and Analysis

Yuang Li, Min Zhang, Mengxin Ren, Xiaosong Qiao, Miaomiao Ma, Daimeng Wei, Hao Yang

Audio deepfake detection (ADD) is essential for preventing the misuse of synthetic voices that may infringe on personal rights and privacy. Recent zero-shot text-to-speech (TTS) models pose higher risks as they can clone voices with a single utterance. However, the existing ADD datasets are outdated, leading to suboptimal generalization of detection models. In this paper, we construct a new cross-domain ADD dataset comprising over 300 hours of speech data that is generated by five advanced zero-shot TTS models. To simulate real-world scenarios, we employ diverse attack methods and audio prompts from different datasets. Experiments show that, through novel attack-augmented training, the Wav2Vec2-large and Whisper-medium models achieve equal error rates of 4.1% and 6.5% respectively. Additionally, we demonstrate our models' outstanding few-shot ADD ability by fine-tuning with just one minute of target-domain data. Nonetheless, neural codec compressors greatly affect the detection accuracy, necessitating further research. Our dataset is publicly available (<https://github.com/leolya/CD-ADD>).

(Nov 13): 7:45:45 (Morning) - Gather

UNO Arena for Evaluating Sequential Decision-Making Capability of Large Language Models

Zhanyue Qin, Haochuan Wang, Deyuan Liu, Ziyang Song, Cunhang Fan, Zhao Lv, Jinlin Wu, Zhen Lei, Zhiving Tu, Dianhui Chu, Xiaoyan Yu, Dianbo Sui

Sequential decision-making refers to algorithms that take into account the dynamics of the environment, where early decisions affect subsequent decisions. With large language models (LLMs) demonstrating powerful capabilities between tasks, we can't help but ask: Can Current LLMs Effectively Make Sequential Decisions? In order to answer this question, we propose the UNO Arena based on the card game UNO

to evaluate the sequential decision-making capability of LLMs and explain in detail why we choose UNO. In UNO Arena, We evaluate the sequential decision-making capability of LLMs dynamically with novel metrics based Monte Carlo methods. We set up random players, DQN-based reinforcement learning players, and LLM players (e.g., GPT-4, Gemini-pro) for comparison testing. Furthermore, in order to improve the sequential decision-making capability of LLMs, we propose the TUTRI player, which can involves having LLMs reflect their own actions with the summary of game history and the game strategy. Numerous experiments demonstrate that the TUTRI player achieves a notable breakthrough in the performance of sequential decision-making compared to the vanilla LLM player.

(Nov 13): 7:458:45 (Morning) - Gather

CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios

Zerian Ouyang, Yishuai Qiu, Linlin Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, Liang He

With the proliferation of Large Language Models (LLMs) in diverse domains, there is a particular need for unified evaluation standards in clinical medical scenarios, where models need to be examined very thoroughly. We present CliMedBench, a comprehensive benchmark with 14 expert-guided core clinical scenarios specifically designed to assess the medical ability of LLMs across 7 pivot dimensions. It comprises 33,735 questions derived from real-world medical reports of top-tier tertiary hospitals and authentic examination exercises. The reliability of this benchmark has been confirmed in several ways. Subsequent experiments with existing LLMs have led to the following findings: (i) Chinese medical LLMs underperform on this benchmark, especially where medical reasoning and factual consistency are vital, underscoring the need for advances in clinical knowledge and diagnostic accuracy. (ii) Several general-domain LLMs demonstrate substantial potential in medical clinics, while the limited input capacity of many medical LLMs hinders their practical use. These findings reveal both the strengths and limitations of LLMs in clinical scenarios and offer critical insights for medical research.

(Nov 13): 7:458:45 (Morning) - Gather

RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs

Ekaterina Taktashova, Maxim Bachukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, Vladislav Mikhailov

Minimal pairs are a well-established approach to evaluating the grammatical knowledge of language models. However, existing resources for minimal pairs address a limited number of languages and lack diversity of language-specific grammatical phenomena. This paper introduces the Russian Benchmark of Linguistic Minimal Pairs (RuBLiMP), which includes 45k pairs of sentences that differ in grammaticality and isolate a morphological, syntactic, or semantic phenomenon. In contrast to existing benchmarks of linguistic minimal pairs, RuBLiMP is created by applying linguistic perturbations to automatically annotated sentences from open text corpora and decontaminating test data. We describe the data collection protocol and present the results of evaluating 25 language models in various scenarios. We find that the widely used LMs for Russian are sensitive to morphological and agreement-oriented contrasts, but fall behind humans on phenomena requiring the understanding of structural relations, negation, transitivity, and tense. RuBLiMP, the codebase, and other materials are publicly available.

(Nov 13): 7:458:45 (Morning) - Gather

ToolBeHonest: A Multi-level Hallucination Diagnostic Benchmark for Tool-Augmented Large Language Models

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen WAN, Yujiu Yang, Tetsuya Sakai, Tian Feng, Hayato Yamana

Tool-augmented large language models (LLMs) are rapidly being integrated into real-world applications. Due to the lack of benchmarks, the community has yet to fully understand the hallucination issues within these models. To address this challenge, we introduce a comprehensive diagnostic benchmark, ToolBH. Specifically, we assess the LLM's hallucinations through two perspectives: depth and breadth. In terms of depth, we propose a multi-level diagnostic process, including (1) solvability detection, (2) solution planning, and (3) missing-tool analysis. For breadth, we consider three scenarios based on the characteristics of the toolset: missing necessary tools, potential tools, and limited functionality tools. Furthermore, we developed seven tasks and collected 700 evaluation samples through multiple rounds of manual annotation. The results show the significant challenges presented by the ToolBH benchmark. The current advanced models, Gemini-1.5-Pro and GPT-4o only achieve total scores of 45.3 and 37.0, respectively, on a scale of 100. In this benchmark, larger model parameters do not guarantee better performance; the training data and response strategies also play crucial roles in tool-enhanced LLM scenarios. Our diagnostic analysis indicates that the primary reason for model errors lies in assessing task solvability. Additionally, open-weight models suffer from performance drops with verbose replies, whereas proprietary models excel with longer reasoning.

(Nov 13): 7:458:45 (Morning) - Gather

PrExMe! Large Scale Prompt Exploration of Open Source LLMs for Machine Translation and Summarization Evaluation

Christoph Leiter, Steffen Eger

Large language models (LLMs) have revolutionized NLP research. Notably, in-context learning enables their use as evaluation metrics for natural language generation, making them particularly advantageous in low-resource scenarios and time-restricted applications. In this work, we introduce `**PrExMe**`, a large-scale `**Pr**ompt **Ex**ploration for **Me**trics, where we evaluate more than 720 prompt templates for open-source LLM-based metrics on machine translation (MT) and summarization datasets, totalling over 6.6M evaluations. This extensive comparison (1) benchmarks recent open-source LLMs as metrics and (2) explores the stability and variability of different prompting strategies. We discover that, on the one hand, there are scenarios for which prompts are stable. For instance, some LLMs show idiosyncratic preferences and favor to grade generated texts with textual labels while others prefer to return numeric scores. On the other hand, the stability of prompts and model rankings can be susceptible to seemingly innocuous changes. For example, changing the requested output format from "0 to 100" "-1 to +1" can strongly affect the rankings in our evaluation. Our study contributes to understanding the impact of different prompting approaches on LLM-based metrics for MT and summarization evaluation, highlighting the most stable prompting patterns and potential limitations.`

(Nov 13): 7:458:45 (Morning) - Gather

Are Large Language Models Good Classifiers? A Study on Edit Intent Classification in Scientific Document Revisions

Qian Ruan, Illia Kuznetsov, Iryna Gurevych

Classification is a core NLP task architecture with many potential applications. While large language models (LLMs) have brought substantial advancements in text generation, their potential for enhancing classification tasks remains underexplored. To address this gap, we propose a framework for thoroughly investigating fine-tuning LLMs for classification, including both generation- and encoding-based approaches. We instantiate this framework in edit intent classification (EIC), a challenging and underexplored classification task. Our extensive experiments and systematic comparisons with various training approaches and a representative selection of LLMs yield new insights into their application for EIC. We investigate the generalizability of these findings on five further classification tasks. To demonstrate the proposed methods and address the data shortage for empirical edit analysis, we use our best-performing EIC model to create Re3-Sci2.0, a new large-scale dataset of 1,780 scientific document revisions with over 94k labeled edits. The quality of the dataset is assessed through human evaluation. The new dataset enables an in-depth empirical study of human editing behavior in academic writing. We make our experimental framework, models and data publicly available.

(Nov 13): 7:458:45 (Morning) - Gather

MetaBench: Planning of Multiple APIs from Various APPs for Complex User Instruction

Hongru WANG, Rui Wang, Boyang XUE, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z. Pan, Kam-Fai Wong

Large Language Models (LLMs) can interact with the real world by connecting with versatile external APIs, resulting in better problem-solving and task automation capabilities. Previous research primarily either focuses on APIs with limited arguments from a single source or overlooks the complex dependency relationship between different APIs. However, it is essential to utilize multiple APIs collaboratively from various sources, especially for complex user instructions. In this paper, we introduce **MetaBench**, the first benchmark to evaluate LLMs' ability to plan and execute multiple APIs from various sources in order to complete the user's task. Specifically, we consider two significant challenges in multiple APIs: 1) graph structures: some APIs can be executed independently while others need to be executed one by one, resulting in graph-like execution order; and 2) permission constraints: which source is authorized to execute the API call. We have experimental results on 9 distinct LLMs; e.g., GPT-4o achieves only a 2.0% success rate at the most complex instruction, revealing that the existing state-of-the-art LLMs still cannot perform well in this situation even with the help of in-context learning and finetuning. Our code and data are publicly available at <https://github.com/ruleGreen/AppBench>.

(Nov 13): 7:45:45 (Morning) - Gather

On Creating an English-Thai Code-switched Machine Translation in Medical Domain

Parinthipan Pengpun, Krittamat Tiansanon, Amresi Chinkamol, Jiramek Kinchagawat, Pitchaya Chairuengitjaras, Pasit Supholkhan, Pubodee Aussavavirojekul, Chiraphat Bonnag, Kanyakorn Veerakanjan, Hirunrak Phimsiri, Boonithica Sae-jia, Nattawach Sataudom, Piyalit Ittichaiwong, Peerat Limkonchotiwat

Machine translation (MT) in the medical domain plays a pivotal role in enhancing healthcare quality and disseminating medical knowledge. Despite advancements in English-Thai MT technology, common MT approaches often underperform in the medical field due to their inability to precisely translate medical terminologies. Our research prioritizes not merely improving translation accuracy but also maintaining medical terminology in English within the translated text through code-switched (CS) translation. We developed a method to produce CS medical translation data, fine-tuned a CS translation model with this data, and evaluated its performance against strong baselines, such as Google Neural Machine Translation (NMT) and GPT-3.5/GPT-4. Our model demonstrated competitive performance in automatic metrics and was highly favored in human preference evaluations. Our evaluation result also shows that medical professionals significantly prefer CS translations that maintain critical English terms accurately, even if it slightly compromises fluency. Our code and test set are publicly available https://github.com/preceptorai-org/NLLB_CS_EM_NLP2024.

(Nov 13): 7:45:45 (Morning) - Gather

LongGenBench: Long-context Generation Benchmark

Xiang Liu, Peijie Dong, Xuming Hu, Xiaowen Chu

Current long-context benchmarks primarily focus on retrieval-based tests, requiring Large Language Models (LLMs) to locate specific information within extensive input contexts, such as the needle-in-a-haystack (NIAH) benchmark. Long-context generation refers to the ability of a language model to generate coherent and contextually accurate text that spans across lengthy passages or documents. While recent studies show strong performance on NIAH and other retrieval-based long-context benchmarks, there is a significant lack of benchmarks for evaluating long-context generation capabilities. To bridge this gap and offer a comprehensive assessment, we introduce a synthetic benchmark, LongGenBench, which allows for flexible configurations of customized generation context lengths. LongGenBench advances beyond traditional benchmarks by redesigning the format of questions and necessitating that LLMs respond with a single, cohesive long-context answer. Upon extensive evaluation using LongGenBench, we observe that: (1) both API accessed and open source models exhibit performance degradation in long-context generation scenarios, ranging from 1.2% to 47.1%; (2) different series of LLMs exhibit varying trends of performance degradation, with the Gemini-1.5-Flash model showing the least degradation among API accessed models, and the Qwen2 series exhibiting the least degradation in LongGenBench among open source models.

(Nov 13): 7:45:45 (Morning) - Gather

R-Judge: Benchmarking Safety Risk Awareness for LLM Agents

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, Gongshen Liu

Large language models (LLMs) have exhibited great potential in autonomously completing tasks across real-world applications. Despite this, these LLM agents introduce unexpected safety risks when operating in interactive environments. Instead of centering on the harmlessness of LLM-generated content in most prior studies, this work addresses the imperative need for benchmarking the behavioral safety of LLM agents within diverse environments. We introduce R-Judge, a benchmark crafted to evaluate the proficiency of LLMs in judging and identifying safety risks given agent interaction records. R-Judge comprises 569 records of multi-turn agent interaction, encompassing 27 key risk scenarios among 5 application categories and 10 risk types. It is of high-quality curation with annotated safety labels and risk descriptions. Evaluation of 11 LLMs on R-Judge shows considerable room for enhancing the risk awareness of LLMs: The best-performing model, GPT-4o, achieves 74.42% while no other models significantly exceed the random. Moreover, we reveal that risk awareness in open agent scenarios is a multi-dimensional capability involving knowledge and reasoning, thus challenging for LLMs. With further experiments, we find that fine-tuning on safety judgment significantly improve model performance while straightforward prompting mechanisms fail. R-Judge is publicly available at [Anonymouse](https://github.com/gongshenliu/R-Judge).

(Nov 13): 7:45:45 (Morning) - Gather

MoleculeQA: A Dataset to Evaluate Factual Accuracy in Molecular Comprehension

Xingyu Lu, He CAO, Zijing Liu, Shengyuan Bai, leqingenchen, Yuan Yao, Hai-Tao Zheng, Yu Li

Large language models are playing an increasingly significant role in molecular research, yet existing models often generate erroneous information. Traditional evaluations fail to assess a model's factual correctness. To rectify this absence, we present MoleculeQA, a novel question answering (QA) dataset which possesses 62K QA pairs over 23K molecules. Each QA pair, composed of a manual question, a positive option and three negative options, has consistent semantics with a molecular description from authoritative corpus. MoleculeQA is not only the first benchmark to evaluate molecular factual correctness but also the largest molecular QA dataset. A comprehensive evaluation on MoleculeQA for existing molecular LLMs exposes their deficiencies in specific aspects and pinpoints crucial factors for molecular modeling. Furthermore, we employ MoleculeQA in reinforcement learning to mitigate model hallucinations, thereby enhancing the factual correctness of generated information.

(Nov 13): 7:45:45 (Morning) - Gather

SLANG: New Concept Comprehension of Large Language Models

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Xueqi Cheng

The dynamic nature of language, particularly evident in the realm of slang and memes on the Internet, poses serious challenges to the adaptability of Large Language Models (LLMs). Traditionally anchored to static datasets, these models often struggle to keep up with the rapid linguistic evolution characteristic of online communities. This research aims to bridge this gap by enhancing LLMs' comprehension of the evolving new concepts on the Internet, without the high cost of continual retraining. In pursuit of this goal, we introduce **SLANG**, a benchmark designed to autonomously integrate novel data and assess LLMs' ability to comprehend emerging concepts, alongside **FOCUS**, an approach uses causal inference to enhance LLMs to understand new phrases and their colloquial context. Our benchmark and approach

involves understanding real-world instances of linguistic shifts, serving as contextual beacons, to form more precise and contextually relevant connections between newly emerging expressions and their meanings. The empirical analysis shows that our causal inference-based approach outperforms the baseline methods in terms of precision and relevance in the comprehension of Internet slang and memes.

(Nov 13): 7:458:45 (Morning) - Gather

Multilingual Synopses of Movie Narratives: A Dataset for Story Understanding

Yidan Sun, Jianfei Yu, Boyang Li

Story video-text alignment, a core task in computational story understanding, aims to align video clips with corresponding sentences in their descriptions. However, progress on the task has been held back by the scarcity of manually annotated video-text correspondence and the heavy concentration on English narrations of Hollywood movies. To address these issues, in this paper, we construct a large-scale multilingual video story dataset named Multilingual Synopses of Movie Narratives (M-SyMoN), containing 13,166 movie summary videos from 7 languages, as well as manual annotation of fine-grained video-text correspondences for 101.5 hours of video. Training on the human annotated data from SyMoN outperforms the SOTA methods by 15.7 and 16.2 percentage points on Clip Accuracy and Sentence IoU scores, respectively, demonstrating the effectiveness of the annotations. As benchmarks for future research, we create 6 baseline approaches with different multilingual training strategies, compare their performance in both intra-lingual and cross-lingual setups, exemplifying the challenges of multilingual video-text alignment. The dataset is released at: <https://github.com/insundaycathy/M-SyMoN>

(Nov 13): 7:458:45 (Morning) - Gather

A Novel Metric for Measuring the Robustness of Large Language Models in Non-adversarial Scenarios

Samuel Ackerman, Ella Rabinovitch, Eitan Farchi, Ateret Anaby Tavor

We evaluate the robustness of several large language models on multiple datasets. Robustness here refers to the relative insensitivity of the model's answers to meaning-preserving variants of their input. Benchmark datasets are constructed by introducing naturally-occurring, non-malicious perturbations, or by generating semantically equivalent paraphrases of input questions or statements. We further propose a novel metric for assessing a model robustness, and demonstrate its benefits in the non-adversarial scenario by empirical evaluation of several models on the created datasets.

(Nov 13): 7:458:45 (Morning) - Gather

DetectBench: Can Large Language Model Detect and Piece Together Implicit Evidence?

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, Yanghua Xiao

Detecting evidence within the context is a key step in the process of reasoning task. Evaluating and enhancing the capabilities of LLMs in evidence detection will strengthen context-based reasoning performance. This paper proposes a benchmark called DetectBench for verifying the ability to detect and piece together implicit evidence within a long context. DetectBench contains 3,928 multiple-choice questions, with an average of 994 tokens per question. Each question contains an average of 4.55 pieces of implicit evidence, and solving the problem typically requires 7.62 logical jumps to find the correct answer. To enhance the performance of LLMs in evidence detection, this paper proposes Detective Reasoning Prompt and Finetune. Experiments demonstrate that the existing LLMs' abilities to detect evidence in long contexts are far inferior to humans. However, the Detective Reasoning Prompt effectively enhances the capability of powerful LLMs in evidence detection, while the Finetuning method shows significant effects in enhancing the performance of weaker LLMs. Moreover, when the abilities of LLMs in evidence detection are improved, their final reasoning performance is also enhanced accordingly.

(Nov 13): 7:458:45 (Morning) - Gather

RoTBench: A Multi-Level Benchmark for Evaluating the Robustness of Large Language Models in Tool Learning

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, Xuanjing Huang

Tool learning has generated widespread interest as a vital means of interaction between Large Language Models (LLMs) and the physical world. Current research predominantly emphasizes LLMs' capacity to utilize tools in well-structured environments while overlooking their stability when confronted with the inevitable noise of the real world. To bridge this gap, we introduce *RoTBench*, a multi-level benchmark for evaluating the robustness of LLMs in tool learning. Specifically, we establish five external environments, each featuring varying levels of noise (i.e., Clean, Slight, Medium, Heavy, and Union), providing an in-depth analysis of the model's resilience across three critical phases: tool selection, parameter identification, and content filling. Experiments involving six widely-used models underscore the urgent necessity for enhancing the robustness of LLMs in tool learning. For instance, the performance of GPT-4 even drops significantly from 80.00 to 58.10 when there is no substantial change in manual accuracy. More surprisingly, the noise correction capability inherent in the GPT family paradoxically impedes its adaptability in the face of mild noise. In light of these findings, we propose RoTTuning, a strategy that enriches the diversity of training environments to bolster the robustness of LLMs in tool learning. The code and data are available at <https://github.com/JunjieYe/RoTBench>.

Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

SEAVER: Attention Reallocation for Mitigating Distractions in Language Models for Conditional Semantic Textual Similarity Measurement

Baixuan Li, Yunlong Fan, Zhiqiang Gao

Conditional Semantic Textual Similarity (C-STS) introduces specific limiting conditions to the traditional Semantic Textual Similarity (STS) task, posing challenges for STS models. Language models employing cross-encoding demonstrate satisfactory performance in STS, yet their effectiveness significantly diminishes in C-STS. In this work, we argue that the failure is due to the fact that the redundant information in the text distracts language models from the required condition-relevant information. To alleviate this, we propose Self-Augmentation via Self-Reweighting (SEAVER), which, based solely on models' internal attention and without the need for external auxiliary information, adaptively reallocates the model's attention weights by emphasizing the importance of condition-relevant tokens. On the C-STS-2023 test set, SEAVER consistently improves performance of all million-scale fine-tuning baseline models (up to around 3 points), and even surpasses performance of billion-scale few-shot prompted large language models (such as GPT-4). Our code is available at <https://github.com/BaixuanLi/SEAVER>.

(Nov 13): 7:458:45 (Morning) - Gather

To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models

Bastien Liétard, Pascal Denis, Mikaela Keller

Polysemy and synonymy are two crucial interrelated facets of lexical ambiguity. While both phenomena are widely documented in lexical

resources and have been studied extensively in NLP, leading to dedicated systems, they are often being considered independently in practical problems. While many tasks dealing with polysemy (e.g. Word Sense Disambiguation or Induction) highlight the role of word's senses, the study of synonymy is rooted in the study of concepts, i.e. meaning shared across the lexicon. In this paper, we introduce ConceptInduction, the unsupervised task of learning a soft clustering among words that defines a set of concepts directly from data. This task generalizes Word Sense Induction. We propose a bi-level approach to Concept Induction that leverages both a local lemma-centric view and a global cross-lexicon view to induce concepts. We evaluate the obtained clustering on SemCor's annotated data and obtain good performance (BCubed F1 above 0.60). We find that the local and the global levels are mutually beneficial to induce concepts and also senses in our setting. Finally, we create static embeddings representing our induced concepts and use them on the Word-in-Context task, obtaining competitive performance with the State-of-the-Art.

(Nov 13): 7:45:45 (Morning) - Gather

Scaling Synthetic Logical Reasoning Datasets with Context-Sensitive Declarative Grammars

Damien Silo

Logical reasoning remains a challenge for natural language processing, but it can be improved by training language models to mimic theorem provers on procedurally generated problems. Previous work used domain-specific proof generation algorithms, which biases reasoning toward specific proof traces and limits auditability and extensibility. We present a simpler and more general declarative framework with flexible context-sensitive rules binding multiple languages (specifically, simplified English and the TPTP theorem-proving language). We construct first-order logic problems by selecting up to 32 premises and one hypothesis. We demonstrate that using semantic constraints during generation and careful English verbalization of predicates enhances logical reasoning without hurting natural English tasks. Using relatively small DeBERTa-v3 models, we achieve state-of-the-art accuracy on the FOLIO human-authored logic dataset, surpassing GPT-4 in accuracy with or without an external solver by 12%.

(Nov 13): 7:45:45 (Morning) - Gather

Advancing Semantic Textual Similarity Modeling: A Regression Framework with Translated ReLU and Smooth K2 Loss

Bowen Zhang, Chuning Li

Since the introduction of BERT and RoBERTa, research on Semantic Textual Similarity (STS) has made groundbreaking progress. Particularly, the adoption of contrastive learning has substantially elevated state-of-the-art performance across various STS benchmarks. However, contrastive learning categorizes text pairs as either semantically similar or dissimilar, failing to leverage fine-grained annotated information and necessitating large batch sizes to prevent model collapse. These constraints pose challenges for researchers engaged in STS tasks that involve nuanced similarity levels or those with limited computational resources, compelling them to explore alternatives like Sentence-BERT. Despite its efficiency, Sentence-BERT tackles STS tasks from a classification perspective, overlooking the progressive nature of semantic relationships, which results in suboptimal performance. To bridge this gap, this paper presents an innovative regression framework and proposes two simple yet effective loss functions: Translated ReLU and Smooth K2 Loss. Experimental results demonstrate that our method achieves convincing performance across seven established STS benchmarks and offers the potential for further optimization of contrastive learning pre-trained models.

(Nov 13): 7:45:45 (Morning) - Gather

Pcc-tuning: Breaking the Contrastive Learning Ceiling in Semantic Textual Similarity

Bowen Zhang, Chuning Li

Semantic Textual Similarity (STS) constitutes a critical research direction in computational linguistics and serves as a key indicator of the encoding capabilities of embedding models. Driven by advances in pre-trained language models and contrastive learning, leading sentence representation methods have reached an average Spearman's correlation score of approximately 86 across seven STS benchmarks in SentEval. However, further progress has become increasingly marginal, with no existing method attaining an average score higher than 86.5 on these tasks. This paper conducts an in-depth analysis of this phenomenon and concludes that the upper limit for Spearman's correlation scores under contrastive learning is 87.5. To transcend this ceiling, we propose an innovative approach termed Pcc-tuning, which employs Pearson's correlation coefficient as a loss function to refine model performance beyond contrastive learning. Experimental results demonstrate that Pcc-tuning can markedly surpass previous state-of-the-art strategies with only a minimal amount of fine-grained annotated samples.

(Nov 13): 7:45:45 (Morning) - Gather

Optimizing Chinese Lexical Simplification Across Word Types: A Hybrid Approach

ZHao Xiao, JieJi Gong, Shijun Wang, Wei Song

This paper addresses the task of Chinese Lexical Simplification (CLS). A key challenge in CLS is the scarcity of data resources. We begin by evaluating the performance of various language models at different scales in unsupervised and few-shot settings, finding that their effectiveness is sensitive to word types. Expensive large language models (LLMs), such as GPT-4, outperform small models in simplifying complex content words and Chinese idioms from the dictionary. To take advantage of this, we propose an automatic knowledge distillation framework called PivotKD for generating training data to fine-tune small models. In addition, all models face difficulties with out-of-dictionary (OOD) words such as internet slang. To address this, we implement a retrieval-based interpretation augmentation (RIA) strategy, injecting word interpretations from external resources into the context. Experimental results demonstrate that fine-tuned small models outperform GPT-4 in simplifying complex content words and Chinese idioms. Additionally, the RIA strategy enhances the performance of most models, particularly in handling OOD words. Our findings suggest that a hybrid approach could optimize CLS performance while managing inference costs. This would involve configuring choices such as model scale, linguistic resources, and the use of RIA based on specific word types to strike an ideal balance.

(Nov 13): 7:45:45 (Morning) - Gather

The Emergence of Compositional Languages in Multi-entity Referential Games: from Image to Graph Representations

Daniel Akkerman, Phong Le, Raquel G. Alhama

To study the requirements needed for a human-like language to develop, Language Emergence research uses jointly trained artificial agents which communicate to solve a task, the most popular of which is a referential game. The targets that agents refer to typically involve a single entity, which limits their ecological validity and the complexity of the emergent languages. Here, we present a simple multi-entity game in which targets include multiple entities that are spatially related. We ask whether agents dealing with multi-entity targets benefit from the use of graph representations, and explore four different graph schemes. Our game requires more sophisticated analyses to capture the extent to which the emergent languages are compositional, and crucially, what the decomposed features are. We find that emergent languages from our setup exhibit a considerable degree of compositionality, but not over all features.

(Nov 13): 7:45:45 (Morning) - Gather

Self-supervised Topic Taxonomy Discovery in the Box Embedding Space

Yiyan Lu, Henggan Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, Qing Li

Topic taxonomy discovery aims at uncovering topics of different abstraction levels and constructing hierarchical relations between them. Unfortunately, most of prior work can hardly model semantic scopes of words and topics by holding the Euclidean embedding space assumption.

What's worse, they infer asymmetric hierarchical relations by symmetric distances between topic embeddings. As a result, existing methods suffer from problems of low-quality topics at high abstraction levels and inaccurate hierarchical relations. To alleviate these problems, this paper develops a Box embedding-based Topic Model (BoxTM) that maps words and topics into the box embedding space, where the asymmetric metric is defined to properly infer hierarchical relations among topics. Additionally, our BoxTM explicitly infers upper-level topics based on correlation between specific topics through recursive clustering on topic boxes. Finally, extensive experiments validate high-quality of the topic taxonomy learned by BoxTM.

Sentiment Analysis, Stylistic Analysis, and Argument Mining

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

Dynamic Multi-granularity Attribution Network for Aspect-based Sentiment Analysis

Yanjiang Chen, Kai Zhang, hufeng, Xianquan Wang, Ruikang li, Qi Liu

Aspect-based sentiment analysis (ABSA) aims to predict the sentiment polarity of a specific aspect within a given sentence. Most existing methods predominantly leverage semantic or syntactic information based on attention scores, which are susceptible to interference caused by irrelevant contexts and often lack sentiment knowledge at a data-specific level. In this paper, we propose a novel Dynamic Multi-granularity Attribution Network (DMAN) from the perspective of attribution. Initially, we leverage Integrated Gradients to dynamically extract attribution scores for each token, which contain underlying reasoning knowledge for sentiment analysis. Subsequently, we aggregate attribution representations from multiple semantic granularities in natural language, enhancing a profound understanding of the semantics. Finally, we integrate attribution scores with syntactic information to capture the relationships between aspects and their relevant contexts more accurately during the sentence understanding process. Extensive experiments on five benchmark datasets demonstrate the effectiveness of our proposed method.

(Nov 13): 7:458:45 (Morning) - Gather

TARA: Token-level Attribute Relation Adaptation for Multi-Attribute Controllable Text Generation

Yilin Cao, Jiahao Zhao, Ruike Zhang, Hanyi Zou, Wenji Mao

Multi-attribute controllable text generation (CTG) aims to generate fluent text satisfying multiple attributes, which is an important and challenging task. The majority of previous research on multi-attribute CTG has ignored the interrelations of attributes that affect the performance of text generation. Recently, several work considers the attribute relations by explicitly defining them as *inhibitory*. We argue that for multi-attribute CTG, the attribute relations are not fixed, which can be not only *inhibitory* but *promotive* as well. In this paper, we tackle the multi-attribute CTG problem by explicitly identifying the above attribute relations for the first time and propose TARA, which employs token-level attribute relation adaptation and representation to generate text with the balanced multi-attribute control. Experimental results on the benchmark dataset demonstrate the effectiveness of our proposed method.

(Nov 13): 7:458:45 (Morning) - Gather

Reusing Transferable Weight Increments for Low-resource Style Generation

Chunchen Jin, Eliot Huang, Heng Chang, Yaqi Wang, Peng Cao, Osmar Zaiane

Text style transfer (TST) is crucial in natural language processing, aiming to endow text with a new style without altering its meaning. In real-world scenarios, not all styles have abundant resources. This work introduces TWIST (reusing Transferable Weight Increments for Style Text generation), a novel framework to mitigate data scarcity by utilizing style features in weight increments to transfer low-resource styles effectively. During target style learning, we derive knowledge via a specially designed weight pool and initialize the parameters for the unseen style. To enhance the effectiveness of merging, the target style weight increments are often merged from multiple source style weight increments through singular vectors. Considering the diversity of styles, we also designed a multi-key memory network that simultaneously focuses on task- and instance-level information to derive the most relevant weight increments. Results from multiple style transfer datasets show that TWIST demonstrates remarkable performance across different backbones, achieving particularly effective results in low-resource scenarios.

Special Theme: Efficiency in Model Algorithms, Training, and Inference

(Nov 13): 7:458:45 (Morning) - Room: Gather

(Nov 13): 7:458:45 (Morning) - Gather

EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees

Yuhui Li, Fangyun Wei, Chao Zhang, Hongyang Zhang

Inference with modern Large Language Models (LLMs) is expensive and time-consuming, and speculative sampling has proven to be an effective solution. Most speculative sampling methods such as EAGLE use a static draft tree, implicitly assuming that the acceptance rate of draft tokens depends only on their position. Interestingly, we found that the acceptance rate of draft tokens is also context-dependent. In this paper, building upon EAGLE, we propose EAGLE-2, which introduces a new technique of context-aware dynamic draft tree into drafting modeling. This improvement leverages the fact that the draft model of EAGLE is well-calibrated: the confidence scores from the draft model approximate acceptance rates with small errors. We conducted extensive evaluations on three series of LLMs and six tasks, with EAGLE-2 achieving speedup ratios of up to $**5x**$, which is 1.3x that of EAGLE. EAGLE-2 also ensures that the distribution of the generated text remains unchanged, making it a $**lossless**$ acceleration algorithm.

(Nov 13): 7:458:45 (Morning) - Gather

A Learning Rate Path Switching Training Paradigm for Version Updates of Large Language Models

Zhihao Wang, Shiyu Liu, Jianheng Huang, Wang Zheng, YiXuan Liao, Xiaoxin Chen, Junfeng Yao, Jinsong Su

Due to the continuous emergence of new data, version updates have become an indispensable requirement for Large Language Models (LLMs). The training paradigms for version updates of LLMs include pre-training from scratch (PTFS) and continual pre-training (CPT). Preliminary experiments demonstrate that PTFS achieves better pre-training performance, while CPT has lower training cost. Moreover, their performance and training cost gaps widen progressively with version updates. To investigate the underlying reasons for this phenomenon, we analyze the effect of learning rate adjustments during the two stages of CPT: preparing an initialization checkpoint and continual pre-training based on this checkpoint. We find that a large learning rate in the first stage and a complete learning rate decay process in the second stage are crucial for version updates of LLMs. Hence, we propose a learning rate path switching training paradigm. Our paradigm comprises one main

path, where we pre-train a LLM with the maximal learning rate, and multiple branching paths, each of which corresponds to an update of the LLM with newly-added training data. Extensive experiments demonstrate the effectiveness and generalization of our paradigm. Particularly, when training four versions of LLMs, our paradigm reduces the total training cost to 58% compared to PTFS, while maintaining comparable pre-training performance.

(Nov 13): 7:45:45 (Morning) - Gather

CHESS: Optimizing LLM Inference via Channel-Wise Thresholding and Selective Sparsification

Junhua He, Shanyu Wu, Weidong Wen, Chun Jason Xue, Qingan Li

Deploying large language models (LLMs) on edge devices presents significant challenges due to the substantial computational overhead and memory requirements. Activation sparsification can mitigate these resource challenges by reducing the number of activated neurons during inference. Existing methods typically employ thresholding-based sparsification based on the statistics of activation tensors. However, they do not model the impact of activation sparsification on performance, resulting in suboptimal performance degradation. To address the limitations, this paper reformulates the activation sparsification problem to explicitly capture the relationship between activation sparsity and model performance. Then, this paper proposes CHESS, a general activation sparsification approach via C H annel-wise thrEsholding and Selective Sparsification. First, channel-wise thresholding assigns a unique threshold to each activation channel in the feed-forward network (FFN) layers. Then, selective sparsification involves applying thresholding-based activation sparsification to specific layers within the attention modules. Finally, we detail the implementation of sparse kernels to accelerate LLM inference. Experimental results demonstrate that the proposed CHESS achieves lower performance degradation over eight downstream tasks while activating fewer parameters than existing methods, thus speeding up the LLM inference by up to 1.27x.

(Nov 13): 7:45:45 (Morning) - Gather

Distilling Instruction-following Abilities of Large Language Models with Task-aware Curriculum Planning

Yuanhao Yue, Chengyu Wang, Jun Huang, Peng Wang

Instruction tuning aims to align large language models (LLMs) with open-domain instructions and human-preferred responses. While several studies have explored autonomous approaches to distilling and annotating instructions from powerful proprietary LLMs, such as ChatGPT, they often neglect the impact of the distributions and characteristics of tasks, together with the varying difficulty of instructions in training sets. This oversight can lead to imbalanced knowledge capabilities and poor generalization powers of student LLMs. To address these challenges, we introduce Task-Aware Curriculum Planning for Instruction Refinement (TAPIR), a multi-round distillation framework that utilizes an oracle LLM to select instructions that are difficult for a student LLM to follow. To balance the student's capabilities, task distributions in training sets are adjusted with responses automatically refined according to their corresponding tasks. In addition, by incorporating curriculum planning, our approach systematically escalates the difficulty levels of tasks, progressively enhancing the student LLM's capabilities. We rigorously evaluate TAPIR using several widely recognized benchmarks (such as AlpacaEval 2.0, MT-Bench, etc.) and multiple student LLMs. Empirical results demonstrate that student LLMs, trained with our method and less training data, outperform larger instruction-tuned models and strong distillation baselines.

(Nov 13): 7:45:45 (Morning) - Gather

Make Some Noise: Unlocking Language Model Parallel Inference Capability through Noisy Training

Yixuan Wang, Xianzhen Luo, Fuxuan Wei, Yijun Liu, Qingfu Zhu, Xuanyu Zhang, Qing Yang, Dongliang Xu, Wanxiang Che

Existing speculative decoding methods typically require additional model structure and training processes to assist the model for draft token generation. This makes the migration of acceleration methods to the new model more costly and more demanding on device memory. To address this problem, we propose the Make Some Noise (MSN) training framework as a replacement for the supervised fine-tuning stage of the large language model. The training method simply introduces some noise at the input for the model to learn the denoising task. It significantly enhances the parallel decoding capability of the model without affecting the original task capability. In addition, we propose a tree-based retrieval-augmented Jacobi (TR-Jacobi) decoding strategy to further improve the inference speed of MSN models. Experiments in both the general and code domains have shown that MSN can improve inference speed by 2.3-2.7x times without compromising model performance. The MSN model also achieves comparable acceleration ratios to the SOTA model with additional model structure on Spec-Bench.

(Nov 13): 7:45:45 (Morning) - Gather

Searching for Best Practices in Retrieval-Augmented Generation

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaqing Zheng, Xuanyang Huang

Retrieval-augmented generation (RAG) techniques have proven to be effective in integrating up-to-date information, mitigating hallucinations, and enhancing response quality, particularly in specialized domains. While many RAG approaches have been proposed to enhance large language models through query-dependent retrievals, these approaches still suffer from their complex implementation and prolonged response times. Typically, a RAG workflow involves multiple processing steps, each of which can be executed in various ways. Here, we investigate existing RAG approaches and their potential combinations to identify optimal RAG practices. Through extensive experiments, we suggest several strategies for deploying RAG that balance both performance and efficiency. Moreover, we demonstrate that multimodal retrieval techniques can significantly enhance question-answering capabilities about visual inputs and accelerate the generation of multimodal content using a "retrieval-as generation" strategy.

(Nov 13): 7:45:45 (Morning) - Gather

Expressive and Generalizable Low-rank Adaptation for Large Models via Slow Cascaded Learning

Swee Li, Yifan Yang, Yifei Shen, Fangyun Wei, Zongqiang Lu, Lili Qiu, Yuqing Yang

Efficient fine-tuning plays a fundamental role in modern large models, with low-rank adaptation emerging as a particularly promising approach. However, the existing variants of LoRA are hampered by limited expressiveness, a tendency to overfit, and sensitivity to hyper-parameter settings. This paper presents LoRA Slow Cascade Learning (LoRASC), an innovative technique designed to enhance LoRA's expressiveness and generalization capabilities while preserving its training efficiency. Our approach augments expressiveness through a cascaded learning strategy that enables a mixture-of-low-rank adaptation, thereby increasing the model's ability to capture complex patterns. Additionally, we introduce a slow-fast update mechanism and cascading noisy tuning to bolster generalization. The extensive experiments on various language and vision datasets, as well as robustness benchmarks, demonstrate that the proposed method not only significantly outperforms existing baselines, but also mitigates overfitting, enhances model stability, and improves OOD robustness.

(Nov 13): 7:45:45 (Morning) - Gather

Pruning via Merging: Compressing LLMs via Manifold Alignment Based Layer Merging

Deyuan Liu, Zhanqie Qin, Hairui Wang, Zhao Yang, Zecheng Wang, Fangying Rong, Qingbin Liu, Yanchao Hao, Bo Li, Xi Chen, Cunhang Fan, Zhao Lv, Dianhui Chu, Zhiying Tu, Dianbo Sui

While large language models (LLMs) excel in many domains, their complexity and scale challenge deployment in resource-limited environments. Current compression techniques, such as parameter pruning, often fail to effectively utilize the knowledge from pruned parameters. To address these challenges, we propose Manifold-Based Knowledge Alignment and Layer Merging Compression (MKA), a novel approach that

uses manifold learning and the Information Bottleneck (IB) measure to merge similar layers, reducing model size while preserving essential performance. We evaluate MKA on multiple benchmark datasets and various LLMs. Our findings show that MKA not only preserves model performance but also achieves substantial compression ratios, outperforming traditional pruning methods. Moreover, when coupled with quantization, MKA delivers even greater compression. Specifically, on the MMLU dataset using the Llama3-8B model, MKA achieves a compression ratio of 43.75% with a minimal performance decrease of only 2.82%. The proposed MKA method offers a resource-efficient and performance-preserving model compression technique for LLMs. We make our code available at <https://github.com/SempraETY/Pruning-via-Merging>

Speech Processing and Spoken Language Understanding

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

TCSinger: Zero-Shot Singing Voice Synthesis with Style Transfer and Multi-Level Style Control

Yu Zhang, Ziyue Jiang, Ruiqi Li, Changhao Pan, Jinzheng He, Rongjie Huang, Chuxin Wang, Zhou Zhao

Zero-shot singing voice synthesis (SVS) with style transfer and style control aims to generate high-quality singing voices with unseen timbres and styles (including singing method, emotion, rhythm, technique, and pronunciation) from audio and text prompts. However, the multi-faceted nature of singing styles poses a significant challenge for effective modeling, transfer, and control. Furthermore, current SVS models often fail to generate singing voices rich in stylistic nuances for unseen singers. To address these challenges, we introduce TCSinger, the first zero-shot SVS model for style transfer across cross-lingual speech and singing styles, along with multi-level style control. Specifically, TCSinger proposes three primary modules: 1) the clustering style encoder employs a clustering vector quantization model to stably condense style information into a compact latent space; 2) the Style and Duration Language Model (S&D-LM) concurrently predicts style information and phoneme duration, which benefits both; 3) the style adaptive decoder uses a novel mel-style adaptive normalization method to generate singing voices with enhanced details. Experimental results show that TCSinger outperforms all baseline models in synthesis quality, singer similarity, and style controllability across various tasks, including zero-shot style transfer, multi-level style control, cross-lingual style transfer, and speech-to-singing style transfer.

(Nov 13): 7:45:45 (Morning) - Gather

AlignCap: Aligning Speech Emotion Captioning to Human Preferences

Ziqi Liang, Haixiang Shi, Hanhui Chen

Speech Emotion Captioning (SEC) has gradually become an active research task. The emotional content conveyed through human speech are often complex, and classifying them into fixed categories may not be enough to fully capture speech emotions. Describing speech emotions through natural language may be a more effective approach. However, existing SEC methods often produce hallucinations and lose generalization on unseen speech. To overcome these problems, we propose AlignCap, which Aligning Speech Emotion Captioning to Human Preferences based on large language model (LLM) with two properties: 1) Speech-Text Alignment, which minimizing the divergence between the LLM's response prediction distributions for speech and text inputs using knowledge distillation (KD) Regularization. 2) Human Preference Alignment, where we design Preference Optimization (PO) Regularization to eliminate factuality and faithfulness hallucinations. We also extract emotional clues as a prompt for enriching fine-grained information under KD-Regularization. Experiments demonstrate that AlignCap presents stronger performance to other state-of-the-art methods on Zero-shot SEC task.

(Nov 13): 7:45:45 (Morning) - Gather

IDEAW: Robust Neural Audio Watermarking with Invertible Dual-Embedding

Pengcheng Li, Xulong Zhang, Jing Xiao, Jianzong Wang

The audio watermarking technique embeds messages into audio and accurately extracts messages from the watermarked audio. Traditional methods develop algorithms based on expert experience to embed watermarks into the time-domain or transform-domain of signals. With the development of deep neural networks, deep learning-based neural audio watermarking has emerged. Compared to traditional algorithms, neural audio watermarking achieves better robustness by considering various attacks during training. However, current neural watermarking methods suffer from low capacity and unsatisfactory imperceptibility. Additionally, the issue of watermark locating, which is extremely important and even more pronounced in neural audio watermarking, has not been adequately studied. In this paper, we design a dual-embedding watermarking model for efficient locating. We also consider the impact of the attack layer on the invertible neural network in robustness training, improving the model to enhance both its reasonableness and stability. Experiments show that the proposed model, IDEAW, can withstand various attacks with higher capacity and more efficient locating ability compared to existing methods.

(Nov 13): 7:45:45 (Morning) - Gather

Self-Powered LLM Modality Expansion for Large Speech-Text Models

Tengfei Yu, Xuebo Liu, Zhiyi Hou, Liang Ding, Dacheng Tao, Min Zhang

Large language models (LLMs) exhibit remarkable performance across diverse tasks, indicating their potential for expansion into large speech-text models (LSMs) by integrating speech capabilities. Although unified speech-text pre-training and multimodal data instruction-tuning offer considerable benefits, these methods generally entail significant resource demands and tend to overfit specific tasks. This study aims to refine the use of speech datasets for LSM training by addressing the limitations of vanilla instruction tuning. We explore the instruction-following dynamics within LSMs, identifying a critical issue termed speech anchor biasa tendency for LSMs to over-rely on speech inputs, mistakenly interpreting the entire speech modality as directives, thereby neglecting textual instructions. To counteract this bias, we introduce a self-powered LSM that leverages augmented automatic speech recognition data generated by the model itself for more effective instruction tuning. Our experiments across a range of speech-based tasks demonstrate that self-powered LSM mitigates speech anchor bias and improves the fusion of speech and text modalities in LSMs. Data, code and scripts are freely available at <https://github.com/ytf-philip/Self-powered-LSM>.

(Nov 13): 7:45:45 (Morning) - Gather

Dual-oriented Disentangled Network with Counterfactual Intervention for Multimodal Intent Detection

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, Yuxian Zou

Multimodal intent detection is designed to leverage diverse modalities for a comprehensive understanding of user intentions in real-world scenarios, thus playing a critical role in modern task-oriented dialogue systems. Existing methods have made great progress in modal alignment and fusion, however, two vital limitations are neglected: (I) close entanglement of multimodal semantics with modal structures; (II) insufficient learning of the causal effects of semantic and modality-specific information on the final predictions under the end-to-end training fashion. To alleviate the above limitations, we introduce the Dual-oriented Disentangled Network with Counterfactual Intervention (DuoDN). DuoDN addresses key limitations in current systems by effectively disentangling and utilizing modality-specific and multimodal semantic information. The model consists of a Dual-oriented Disentangled Encoder that decouples semantics-oriented and modality-oriented representations,

alongside a Counterfactual Intervention Module that applies causal inference to understand causal effects by injecting confounders. Experiments on three benchmark datasets demonstrate DuoDN's superiority over existing methods, with extensive analysis validating its advantages.

(Nov 13): 7:45:45 (Morning) - Gather

Beyond Common Words: Enhancing ASR Cross-Lingual Proper Noun Recognition Using Large Language Models

Rishabh Kumar, Sabyasachi Ghosh, Ganesh Ramakrishnan

In this work, we address the challenge of cross-lingual proper noun recognition in automatic speech recognition (ASR), where proper nouns in an utterance may originate from a language different from the language in which the ASR system is trained. We enhance the performance of end-to-end ASR systems by instructing a large language model (LLM) to correct the ASR model's predictions. The LLM's context is augmented with a dictionary of cross-lingual words that are phonetically and graphemically similar to the potentially incorrect proper nouns in the ASR predictions. Our dictionary-based method DiP-ASR (Dictionary-based Prompting for Automatic Speech Recognition) significantly reduces word error rates compared to both the end-to-end ASR baseline and instruction-based prompting of the LLM without the dictionary across cross-lingual proper noun recognition tasks involving three secondary languages.

(Nov 13): 7:45:45 (Morning) - Gather

LaRA: Large Rank Adaptation for Speech and Text Cross-Modal Learning in Large Language Models

Zahair hasan shaik, Pradyot Hegde, Prashant Bamunmath, Deepak K T

Integrating speech and text capabilities into large language models (LLMs) is a challenging task and we present Large Rank Adaptation (LaRA) for effective cross-modal integration of speech and text in the LLM framework. Unlike conventional LoRA, our method requires significantly larger ranks comparable to the pretrained weights to accommodate the complexities of speech-text cross-modality learning. The approach utilizes HuBERT to convert speech into discrete tokens and fine-tunes the pretrained LLM to adapt to cross-modal inputs and outputs. The work employs a Hi-Fi GAN vocoder to synthesize speech waveforms from the generated speech units. The initial studies use the LibriSpeech corpus to teach the model the relationships between speech and text, and Daily Talk, which involves dialog conversations, to adapt for interaction. The proposed work demonstrates adaptation for spoken and text conversations. However, the proposed framework can be easily extended to other cross-modal applications.

Summarization

(Nov 13): 7:45:45 (Morning) - Room: Gather

(Nov 13): 7:45:45 (Morning) - Gather

Improving Factual Consistency of News Summarization by Contrastive Preference Optimization

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, Qianli Ma

Despite the recent progress in news summarization made by large language models (LLMs), they often generate summaries that are factually inconsistent with original articles, known as "hallucinations" in text generation. Unlike previous small models (e.g., BART, T5), current LLMs make fewer silly mistakes but more sophisticated ones, such as imposing cause and effect, adding false details, overgeneralizing, etc. These hallucinations are challenging to detect through traditional methods, which poses great challenges for improving the factual consistency of text summarization. In this paper, we propose Contrastive Preference Optimization (CPO) to disentangle the LLMs' propensities to generate faithful and fake content. Furthermore, we adopt a probing-based specific training method to improve their capacity of distinguishing two types of propensities. In this way, LLMs can execute the instructions more accurately and have enhanced perception of hallucinations. Experimental results show that CPO significantly improves the reliability of summarization based on LLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation

While summarization has been extensively researched in natural language processing (NLP), cross-lingual cross-temporal summarization (CLCTS) is a largely unexplored area that has the potential to improve cross-cultural accessibility and understanding. This article comprehensively addresses the CLCTS task, including dataset creation, modeling, and evaluation. We (1) build the first CLCTS corpus with 328 instances for hDe-En (extended version with 455 instances) and 289 for hEn-De (extended version with 501 instances), leveraging historical fiction texts and Wikipedia summaries in English and German; (2) examine the effectiveness of popular transformer end-to-end models with different intermediate fine-tuning tasks; (3) explore the potential of GPT-3.5 as a summarizer; and (4) report evaluations from humans, GPT-4, and several recent automatic evaluation metrics. Our results indicate that intermediate task finetuned end-to-end models generate bad to moderate quality summaries while GPT-3.5, as a zero-shot summarizer, provides moderate to good quality outputs. GPT-3.5 also seems very adept at normalizing historical text. To assess data contamination in GPT-3.5, we design an adversarial attack scheme in which we find that GPT-3.5 performs slightly worse for unseen source documents compared to seen documents. Moreover, it sometimes hallucinates when the source sentences are inverted against its prior knowledge with a summarization accuracy of 0.67 for plot omission, 0.71 for entity swap, and 0.53 for plot negation. Overall, our regression results of model performances suggest that longer, older, and more complex source texts (all of which are more characteristic for historical language variants) are harder to summarize for all models, indicating the difficulty of the CLCTS task. Regarding evaluation, we observe that both the GPT-4 and BERTScore correlate moderately with human evaluations, implicating great potential for future improvement.

(Nov 13): 7:45:45 (Morning) - Gather

GlobeSumm: A Challenging Benchmark Towards Unifying Multi-lingual, Cross-lingual and Multi-document News Summarization

Yanfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin

News summarization in todays global scene can be daunting with its flood of multilingual content and varied viewpoints from different sources. However, current studies often neglect such real-world scenarios as they tend to focus solely on either single-language or single-document tasks. To bridge this gap, we aim to unify Multi-lingual, Cross-lingual and Multi-document Summarization into a novel task, i.e., MCMS, which encapsulates the real-world requirements all-in-one. Nevertheless, the lack of a benchmark inhibits researchers from adequately studying this invaluable problem. To tackle this, we have meticulously constructed the GLOBESUMM dataset by first collecting a wealth of multilingual news reports and restructuring them into event-centric format. Additionally, we introduce the method of protocol-guided prompting for high-quality and cost-effective reference annotation. In MCMS, we also highlight the challenge of conflicts between news reports, in addition to the issues of redundancies and omissions, further enhancing the complexity of GLOBESUMM. Through extensive experimental analysis, we validate the quality of our dataset and elucidate the inherent challenges of the task. We firmly believe that GLOBESUMM, given its challenging nature, will greatly contribute to the multilingual communities and the evaluation of LLMs.

(Nov 13): 7:45:45 (Morning) - Gather

Identifying Factual Inconsistencies in Summaries: Grounding Model Inference via Task Taxonomy*Liyuan Xu, Zhenlin Su, Mo Yu, Jin Xu, Jinho D. Choi, Jie Zhou, Fei Liu*

Factual inconsistencies pose a significant hurdle for the faithful summarization by generative models. While a major direction to enhance inconsistency detection is to derive stronger Natural Language Inference (NLI) models, we propose an orthogonal aspect that underscores the importance of incorporating task-specific taxonomy into the inference. To this end, we consolidate key error types of inconsistent facts in summaries, and incorporate them to facilitate both the zero-shot and supervised paradigms of LLMs. Extensive experiments on ten datasets of five distinct domains suggest that, zero-shot LLM inference could benefit from the explicit solution space depicted by the error type taxonomy, and achieves state-of-the-art performance overall, surpassing specialized non-LLM baselines, as well as recent LLM baselines. We further distill models that fuse the taxonomy into parameters through our designed prompt completions and supervised training strategies, efficiently substituting state-of-the-art zero-shot inference with much larger LLMs.

Syntax: Tagging, Chunking and Parsing*(Nov 13): 7:45A:45 (Morning) - Room: Gather**(Nov 13): 7:45A:45 (Morning) - Gather***When Generative Adversarial Networks Meet Sequence Labeling Challenges***Yu Tong, Ge Chen, Guokai Zheng, Rui Li, Jiang Dashi*

The current framework for sequence labeling encompasses a feature extractor and a sequence tagger. This study introduces a unified framework named SLGAN, which harnesses the capabilities of Generative Adversarial Networks to address the challenges associated with Sequence Labeling tasks. SLGAN not only mitigates the limitation of GANs in backpropagating loss to discrete data but also exhibits strong adaptability to various sequence labeling tasks. Unlike traditional GANs, the discriminator within SLGAN does not discriminate whether data originates from the discriminator or the generator; instead, it focuses on predicting the correctness of each tag within the tag sequence. We conducted evaluations on six different tasks spanning four languages, including Chinese, Japanese, and Korean Word Segmentation, Chinese and English Named Entity Recognition, and Chinese Part-of-Speech Tagging. Our experimental results illustrate that SLGAN represents a versatile and highly effective solution, consistently achieving state-of-the-art or competitive performance results, irrespective of the specific task or language under consideration.

*(Nov 13): 7:45A:45 (Morning) - Gather***Contribution of Linguistic Typology to Universal Dependency Parsing: An Empirical Investigation***Ali Basirat, Navid Baradarani Hemmati*

Universal Dependencies (UD) is a global initiative to create a standard annotation for the dependency syntax of human languages. Addressing its deviation from typological principles, this study presents an empirical investigation of a typologically motivated transformation of UD proposed by William Croft. Our findings underscore the significance of the transformations across diverse languages and highlight their advantages and limitations.

NLP Applications*(Nov 13): 7:45A:45 (Morning) - Room: Gather**(Nov 13): 7:45A:45 (Morning) - Gather***A Usage-centric Take on Intent Understanding in E-Commerce***Wendi Zhou, Tianyi Li, Pavlos Vougiouklis, Mark Steedman, Jeff Z. Pan*

Identifying and understanding user intents is a pivotal task for E-Commerce. Despite its essential role in product recommendation and business user profiling analysis, intent understanding has not been consistently defined or accurately benchmarked. In this paper, we focus on predicative user intents as how a customer uses a product, and pose intent understanding as a natural language reasoning task, independent of product ontologies. We identify two weaknesses of FolkScope, the SOTA E-Commerce Intent Knowledge Graph: category-rigidity and property-ambiguity. They limit its ability to strongly align user intents with products having the most desirable property, and to recommend useful products across diverse categories. Following these observations, we introduce a Product Recovery Benchmark featuring a novel evaluation framework and an example dataset. We further validate the above FolkScope weaknesses on this benchmark. Our code and dataset are available at <https://github.com/stayones/Usgae-Centric-Intent-Understanding>.

Virtual Poster Session 3 - (Nov 14): 13:0014:00 (Afternoon)**Computational Social Science and Cultural Analytics***(Nov 14): 13:0014:00 (Afternoon) - Room: Gather**(Nov 14): 13:0014:00 (Afternoon) - Gather***Oddballs and Misfits: Detecting Implicit Abuse in Which Identity Groups are Depicted as Deviating from the Norm***Michael Wiegand, Josef Ruppenhofer*

We address the task of detecting abusive sentences in which identity groups are depicted as deviating from the norm (e.g. Gays sprinkle flour over their gardens for good luck). These abusive utterances need not be stereotypes or negative in sentiment. We introduce the first dataset for this task. It is created via crowdsourcing and includes 7 identity groups. We also report on classification experiments.

*(Nov 14): 13:0014:00 (Afternoon) - Gather***Deciphering Rumors: A Multi-Task Learning Approach with Intent-aware Hierarchical Contrastive Learning**

Chang Yang, Peng Zhang, Hui Gao, Jing Zhang

Social networks are rife with noise and misleading information, presenting multifaceted challenges for rumor detection. In this paper, from the perspective of human cognitive subjectivity, we introduce the mining of individual latent intentions and propose a novel multi-task learning framework, the Intent-Aware Rumor Detection Network (IRDNet). IRDNet is designed to discern multi-level rumor semantic features and latent user intentions, addressing the challenges of robustness and key feature mining and alignment that plague existing models. In IRDNet, the multi-level semantic extraction module captures sequential and hierarchical features to generate robust semantic representations. The hierarchical contrastive learning module incorporates two complementary strategies, event-level and intent-level, to establish cognitive anchors that uncover the latent intentions of information disseminators. Event-level contrastive learning employs high-quality data augmentation and adversarial perturbations to enhance model robustness. Intent-level contrastive learning leverages the intent encoder to capture latent intent features and optimize consistency within the same intent while ensuring heterogeneity between different intents to clearly distinguish key features from irrelevant elements. Experimental results demonstrate that IRDNet significantly improves the effectiveness of rumor detection and effectively addresses the challenges present in the field of rumor detection.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Investigating LLMs as Voting Assistants via Contextual Augmentation: A Case Study on the European Parliament Elections 2024

Ilias Chalkidis

In light of the recent 2024 European Parliament elections, we are investigating if LLMs can be used as Voting Advice Applications (VAAs). We audit MISTRAL and MIXTRAL models and evaluate their accuracy in predicting the stance of political parties based on the latest "EU and I" voting assistance questionnaire. Furthermore, we explore alternatives to improve models' performance by augmenting the input context via Retrieval-Augmented Generation (RAG) relying on web search, and Self-Reflection using staged conversations that aim to re-collect relevant content from the model's internal memory. We find that MIXTRAL is highly accurate with an 82% accuracy on average with a significant performance disparity across different political groups (50-95%). Augmenting the input context with expert-curated information can lead to a significant boost of approx. 9%, which remains an open challenge for automated RAG approaches, even considering curated content.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Enhancing Data Quality through Simple De-duplication: Navigating Responsible Computational Social Science Research

Yida Mu, Mali Jin, Xingyi Song, Nikolaos Aletras

Research in natural language processing (NLP) for Computational Social Science (CSS) heavily relies on data from social media platforms. This data plays a crucial role in the development of models for analysing socio-linguistic phenomena within online communities. In this work, we conduct an in-depth examination of 20 datasets extensively used in NLP for CSS to comprehensively examine data quality. Our analysis reveals that social media datasets exhibit varying levels of data duplication. Consequently, this gives rise to challenges like label inconsistencies and data leakage, compromising the reliability of models. Our findings also suggest that data duplication has an impact on the current claims of state-of-the-art performance, potentially leading to an overestimation of model effectiveness in real-world scenarios. Finally, we propose new protocols and best practices for improving dataset development from social media data and its usage.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Quantifying Generative Media Bias with a Corpus of Real-world and Generated News Articles

Filip Trhlík, Pontus Stenetorp

Large language models (LLMs) are increasingly being utilised across a range of tasks and domains, with a burgeoning interest in their application within the field of journalism. This trend raises concerns due to our limited understanding of LLM behaviour in this domain, especially with respect to political bias. Existing studies predominantly focus on LLMs undertaking political questionnaires, which offers only limited insights into their biases and operational nuances. To address this gap, our study establishes a new curated dataset that contains 2,100 human-written articles and utilises their descriptions to generate 56,700 synthetic articles using nine LLMs. This enables us to analyse shifts in properties between human-authored and machine-generated articles, with this study focusing on political bias, detecting it using both supervised models and LLMs. Our findings reveal significant disparities between base and instruction-tuned LLMs, with instruction-tuned models exhibiting consistent political bias. Furthermore, we are able to study how LLMs behave as classifiers, observing their display of political bias even in this role. Overall, for the first time within the journalistic domain, this study outlines a framework and provides a structured dataset for quantifiable experiments, serving as a foundation for further research into LLM political bias and its implications.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Still Not Quite There! Assessing Large Language Models for Comorbid Mental Health Diagnosis

Amey Hengle, Atharva Kulkarni, Shantanu Deepak Patankar, Rashmi Gupta

In this study, we introduce ANGST, a novel, first of its kind benchmark for depression-anxiety comorbidity classification from social media posts. Unlike contemporary datasets that often oversimplify the intricate interplay between different mental health disorders by treating them as isolated conditions, ANGST enables multi-label classification, allowing each post to be simultaneously identified as indicating depression and/or anxiety. Comprising 2876 meticulously annotated posts by expert psychologists and an additional 7667 silver-labeled posts, ANGST posits a more representative sample of online mental health discourse. Moreover, we benchmark ANGST using various state-of-the-art language models, ranging from Mental-BERT to GPT-4. Our results provide significant insights into the capabilities and limitations of these models in complex diagnostic scenarios. While GPT-4 generally outperforms other models, none achieve an F1 score exceeding 72% in multi-class comorbid classification, underscoring the ongoing challenges in applying language models to mental health diagnostics.

(Nov 14): 13:0014:00 (Afternoon) - Gather

F²RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation

Haiyang Wang, Yuchen Pan, Xin Song, Xuechen Zhao, Minghao Hu, Bin Zhou

Hate speech (HS) on social media exacerbates misinformation and baseless prejudices. Evidence-supported counterspeech (CS) is crucial for correcting misinformation and reducing prejudices through facts. Existing methods for generating evidence-supported CS often lack clear guidance with a core claim for organizing evidence and do not adequately address factuality and faithfulness hallucinations in CS within anti-hate contexts. In this paper, to mitigate the aforementioned, we propose F²RL, a Factuality and Faithfulness Reinforcement Learning framework for generating claim-guided and evidence-supported CS. Firstly, we generate counter-claims based on hate speech and design a self-evaluation mechanism to select the most appropriate one. Secondly, we propose a coarse-to-fine evidence retrieval method. This method initially generates broad queries to ensure the diversity of evidence, followed by carefully reranking the retrieved evidence to ensure its relevance to the claim. Finally, we design a reinforcement learning method with a triplet-based factuality reward model and a multi-aspect faithfulness reward model. The method rewards the generator to encourage greater factuality, more accurate refutation of hate speech, consistency with the claim, and better utilization of evidence. Extensive experiments on three benchmark datasets demonstrate that the proposed framework achieves excellent performance in CS generation, with strong factuality and faithfulness.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Moral Foundations of Large Language Models*Marwa Abdulhai, Gregory Serapio-Garcia, Clement CREPY, Daria Valter, John Canny, Natasha Jaques*

Moral foundations theory (MFT) is a psychological assessment tool that decomposes human moral reasoning into five factors, including care/harm, liberty/oppression, and sanctity/degradation (Graham et al., 2009). People vary in the weight they place on these dimensions when making moral decisions, in part due to their cultural upbringing and political ideology. As large language models (LLMs) are trained on datasets collected from the internet, they may reflect the biases that are present in such corpora. This paper uses MFT as a lens to analyze whether popular LLMs have acquired a bias towards a particular set of moral values. We analyze known LLMs and find they exhibit particular moral foundations, and show how these relate to human moral foundations and political affiliations. We also measure the consistency of these biases, or whether they vary strongly depending on the context of how the model is prompted. Finally, we show that we can adversarially select prompts that encourage the model to exhibit a particular set of moral foundations, and that this can affect the model's behavior on downstream tasks. These findings help illustrate the potential risks and unintended consequences of LLMs assuming a particular moral stance.

Demo

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Debug Smarter, Not Harder: AI Agents for Error Resolution in Computational Notebooks*Artem Borzilov, Konstantin Grotov, Maksim Krivobok, Timofey Bryksin, Yaroslav Zharov*

Computational notebooks became indispensable tools for research-related development, offering unprecedented interactivity and flexibility in the development process. However, these benefits come at the cost of reproducibility and an increased potential for bugs. With the rise of code-fluent Large Language Models empowered with agentic techniques, smart bug-fixing tools with a high level of autonomy have emerged. However, those tools are tuned for classical script programming and still struggle with non-linear computational notebooks. In this paper, we present an AI agent designed specifically for error resolution in a computational notebook. We have developed an agentic system capable of exploring a notebook environment by interacting with it—similar to how a user would—and integrated the system into the JetBrains service for collaborative data science called DataLore. We evaluate our approach against the pre-existing single-action solution by comparing costs and conducting a user study. Users rate the error resolution capabilities of the agentic system higher but experience difficulties with UI. We share the results of the study and consider them valuable for further improving user-agent collaboration.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Kandinsky 3: Text-to-Image Synthesis for Multifunctional Generative Framework*Andrey Kuznetsov, Anna Averchenkova, Bakushkin Anton, Evelina Mironova, Igor Pavlov, Julia Agafonova, Konstantin Kulikov, Nikolai Gerasimenko, Viacheslav Vasilev, Andrey Filatov, Arkhipkin Sergeevich Vladimir, Denis Valerievich Dimitrov*

Text-to-image (T2I) diffusion models are popular for introducing image manipulation methods, such as editing, image fusion, inpainting, etc. At the same time, image-to-video (I2V) and text-to-video (T2V) models are also built on top of T2I models. We present Kandinsky 3, a novel T2I model based on latent diffusion, achieving a high level of quality and photorealism. The key feature of the new architecture is the simplicity and efficiency of its adaptation for many types of generation tasks. We extend the base T2I model for various applications and create a multifunctional generation system that includes text-guided inpainting/outpainting, image fusion, text-image fusion, image variations generation, I2V and T2V generation. We also present a distilled version of the T2I model, evaluating inference in 4 steps of the reverse process without reducing image quality and 3 times faster than the base model. We deployed a user-friendly demo system in which all the features can be tested in the public domain. Additionally, we released the source code and checkpoints for the Kandinsky 3 and extended models. Human evaluations show that Kandinsky 3 demonstrates one of the highest quality scores among open source generation systems.

(Nov 14): 13:0014:00 (Afternoon) - Gather

i-Code Studio: A Configurable and Composable Framework for Integrative AI*Chengwu Zhu, Lu Yuan, MAHMOUD KHADEMI, Michael Zeng, Reid Pryzant, Takuya Yoshioka, Xuedong Huang, Yao Qian, Yichong Xu, Yuwei Fang, Ziyi Yang*

Artificial General Intelligence (AGI) requires comprehensive understanding and generation capabilities for a variety of tasks spanning different modalities and functionalities. Integrative AI is one important direction to approach AGI, through combining multiple models to tackle complex multimodal tasks. However, there is a lack of a flexible and composable platform to facilitate efficient and effective model composition and coordination. In this paper, we propose the i-Code Studio, a configurable and composable framework for Integrative AI. The i-Code Studio orchestrates multiple pre-trained models in a finetuning-free fashion to conduct complex multimodal tasks. Instead of simple model composition, the i-Code Studio provides an integrative, flexible, and composable setting for developers to quickly and easily compose cutting-edge services and technologies tailored to their specific requirements. The i-Code Studio achieves impressive results on a variety of zero-shot multimodal tasks, such as video-to-text retrieval, speech-to-speech translation, and visual question answering. We also demonstrate how to quickly build a multimodal agent based on the i-Code Studio that can communicate and personalize for users. The project page with demonstrations and code is at <https://i-code-studio.github.io/>.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Integrating INCEpTION into larger annotation processes*Iryna Gurevych, Jan-Christoph Kie, Richard Eckart de Castilho*

Annotation tools are increasingly only steps in a larger process into which they need to be integrated, for instance by calling out to web services for labeling support or importing documents from external sources. This requires certain capabilities that annotation tools need to support in order to keep up. Here, we define the respective requirements and how popular annotation tools support them. As a demonstration for how these can be implemented, we adapted INCEpTION, a semantic annotation platform offering intelligent assistance and knowledge management. For instance, support for a range of APIs has been added to INCEpTION through which it can be controlled and which allow it to interact with external services such as authorization services, crowdsourcing platforms, terminology services or machine learning services. Additionally, we introduce new capabilities that allow custom rendering of XML documents and even the ability to add new JavaScript-based editor plugins, thereby making INCEpTION usable in an even wider range of annotation tasks.

(Nov 14): 13:0014:00 (Afternoon) - Gather

TruthReader: Towards Trustworthy Document Assistant Chatbot with Reliable Attribution*Baotian Hu, Min Zhang, Qian Chen, Shaolin Ye, Xinshuo Hu, Zetian Sun, Zifei Shan, dongfang li*

Document assistant chatbots are empowered with extensive capabilities by Large Language Models (LLMs) and have exhibited significant advancements. However, these systems may suffer from hallucinations that are difficult to verify in the context of given documents. Moreover, despite the emergence of products for document assistants, they either heavily rely on commercial LLM APIs or lack transparency in

their technical implementations, leading to expensive usage costs and data privacy concerns. In this work, we introduce a fully open-source document assistant chatbot with reliable attribution, named TruthReader, utilizing adapted conversational retriever and LLMs. Our system enables the LLMs to generate answers with detailed inline citations, which can be attributed to the original document paragraphs, facilitating the verification of the factual consistency of the generated text. To further adapt the generative model, we develop a comprehensive pipeline consisting of data construction and model optimization processes. This pipeline equips the LLMs with the necessary capabilities to generate accurate answers, produce reliable citations, and refuse unanswerable questions. Our codebase, data and models are released, and the video demonstration of our system is available at <https://youtu.be/RYVt3itzUQM>.

Ethics, Bias, and Fairness

(Nov 14): 13:00 14:00 (Afternoon) - Room: Gather

(Nov 14): 13:00 14:00 (Afternoon) - Gather

MITTenS: A Dataset for Evaluating Gender Mistranslation

Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, Jasmin Bastings

Translation systems, including foundation models capable of translation, can produce errors that result in gender mistranslation, and such errors can be especially harmful. To measure the extent of such potential harms when translating into and out of English, we introduce a dataset, MITTenS, covering 26 languages from a variety of language families and scripts, including several traditionally under-represented in digital resources. The dataset is constructed with handcrafted passages that target known failure patterns, longer synthetically generated passages, and natural passages sourced from multiple domains. We demonstrate the usefulness of the dataset by evaluating both neural machine translation systems and foundation models, and show that all systems exhibit gender mistranslation and potential harm, even in high resource languages.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Walking in Others' Shoes: How Perspective-Taking Guides Large Language Models in Reducing Toxicity and Bias

Rongwu Xu, Zian Zhou, Tianwei Zhang, Zehan Qi, SÜ YAO, Ke Xu, Wei Xu, Han Qiu

The common toxicity and societal bias in contents generated by large language models (LLMs) necessitate strategies to reduce harm. Present solutions often demand white-box access to the model or substantial training, which is impractical for cutting-edge commercial LLMs. Moreover, prevailing prompting methods depend on external tool feedback and fail to simultaneously lessen toxicity and bias. Motivated by social psychology principles, we propose a novel strategy named perspective-taking prompting (PeT) that inspires LLMs to integrate diverse human perspectives and self-regulate their responses. This self-correction mechanism can significantly diminish toxicity (up to 89%) and bias (up to 73%) in LLMs' responses. Rigorous evaluations and ablation studies are conducted on two commercial LLMs (ChatGPT and GLM) and three open-source LLMs, revealing PeT's superiority in producing less harmful responses, outperforming five strong baselines.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Flex Tape Cant Fix That: Bias and Misinformation in Edited Language Models

Karina H Halevy, Anna Sotnikova, Badr AlKhamissi, Syrielle Montariol, Antoine Bosselut

Weight-based model editing methods update the parametric knowledge of language models post-training. However, these methods can unintentionally alter unlearned parametric knowledge representations, potentially increasing the risk of harm. In this work, we investigate how weight editing methods unexpectedly amplify model biases after edits. We introduce a novel benchmark dataset, Seesaw-CF, for measuring bias amplification of model editing methods for demographic traits such as race, geographic origin, and gender. We use Seesaw-CF to examine the impact of model editing on bias in five large language models. Our results demonstrate that edited models exhibit, to various degrees, more biased behavior for certain demographic groups than before they were edited, specifically becoming less confident in properties for Asian and African subjects. Additionally, editing facts about place of birth, country of citizenship, or gender has particularly negative effects on the model's knowledge about unrelated properties, such as field of work, a pattern observed across multiple models.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Fine-grained Pluggable Gradient Ascent for Knowledge Unlearning in Language Models

XiaoHua Feng, Chaochao Chen, Yuyuan Li, Zibin Lin

Pre-trained language models acquire knowledge from vast amounts of text data, which can inadvertently contain sensitive information. To mitigate the presence of undesirable knowledge, the task of knowledge unlearning becomes crucial for language models. Previous research relies on gradient ascent methods to achieve knowledge unlearning, which is simple and effective. However, this approach calculates all the gradients of tokens in the sequence, potentially compromising the general ability of language models. To overcome this limitation, we propose an adaptive objective that calculates gradients with fine-grained control specifically targeting sensitive tokens. Our adaptive objective is pluggable, ensuring simplicity and enabling extension to the regularization-based framework that utilizes non-target data or other models to preserve general ability. Through extensive experiments targeting the removal of typical sensitive data, we demonstrate that our proposed method enhances the general ability of language models while achieving knowledge unlearning. Additionally, it demonstrates the capability to adapt to behavior alignment, eliminating all the undesirable knowledge within a specific domain.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Split and Merge: Aligning Position Biases in LLM-based Evaluators

Zongjie Li, Chaocheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, Yang Liu

Large language models (LLMs) have shown promise as automated evaluators for assessing the quality of answers generated by AI systems. However, LLM-based evaluators exhibit position bias, or inconsistency, when used to evaluate candidate answers in pairwise comparisons, favoring either the first or second answer regardless of content. To address this limitation, we propose PORTIA, an alignment-based system designed to mimic human comparison strategies to calibrate position bias in a lightweight yet effective manner. Specifically, PORTIA splits the answers into multiple segments, taking into account both length and semantics, and merges them back into a single prompt for evaluation by LLMs. Extensive experiments with six LLMs on 11,520 answer pairs demonstrate that PORTIA markedly enhances the consistency rates for all models and forms of comparison tested, achieving an average relative improvement of 47.46%. It also enables PORTIA-enhanced GPT-3.5 to achieve agreement rates with humans comparable to GPT-4 and elevates GPT-4's consistency rate up to 98%. Subsequent human evaluations indicate that the PORTIA-enhanced GPT-3.5 model can even surpass standalone GPT-4 in terms of alignment with human evaluations, highlighting PORTIA's ability to correct position bias, improve LLM consistency, and boost performance while keeping cost efficiency.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Text Fluoroscopy: Detecting LLM-Generated Text through Intrinsic Features

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, Nenghai Yu

Large language models (LLMs) have revolutionized the domain of natural language processing because of their excellent performance on various tasks. Despite their impressive capabilities, LLMs also have the potential to generate texts that pose risks of misuse. Consequently, detecting LLM-generated text has become increasingly important. Previous LLM-generated text detection methods use semantic features, which are stored in the last layer. This leads to methods that overfit the training set domain and exhibit shortcomings in generalization. Therefore, We argue that utilizing intrinsic features rather than semantic features for detection results in better performance. In this work, we design Text Fluoroscopy, a black-box method with better generalizability for detecting LLM-generated text by mining the intrinsic features of the text to be detected. Our method captures the text's intrinsic features by identifying the layer with the largest distribution difference from the last and first layers when projected to the vocabulary space. Our method achieves 7.36% and 2.84% average improvement in detection performance compared to the baselines in detecting texts from different domains generated by GPT-4 and Claude3, respectively.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Can AI Relate: Testing Large Language Model Response for Mental Health Support

Saadie Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, Marzyeh Ghassemi

Large language models (LLMs) are already being piloted for clinical use in hospital systems like NYU Langone, Dana-Farber and the NHS. A proposed deployment use case is psychotherapy, where a LLM-powered chatbot can treat a patient undergoing a mental health crisis. Deployment of LLMs for mental health response could hypothetically broaden access to psychotherapy and provide new possibilities for personalizing care. However, recent high-profile failures, like damaging dieting advice offered by the Tessa chatbot to patients with eating disorders, have led to doubt about their reliability in high-stakes and safety-critical settings. In this work, we develop an evaluation framework for determining whether LLM response is a viable and ethical path forward for the automation of mental health treatment. Our framework measures equity in empathy and adherence of LLM responses to motivational interviewing theory. Using human evaluation with trained clinicians and automatic quality-of-care metrics grounded in psychology research, we compare the responses provided by peer-to-peer responders to those provided by a state-of-the-art LLM. We show that LLMs like GPT-4 use implicit and explicit cues to infer patient demographics like race. We then show that there are statistically significant discrepancies between patient subgroups: Responses to Black posters consistently have lower empathy than for any other demographic group (2%-13% lower than the control group). Promisingly, we do find that the manner in which responses are generated significantly impacts the quality of the response. We conclude by proposing safety guidelines for the potential deployment of LLMs for mental health response.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling

Matúš Píkuliak, Štefan Oresko, Andrea Hrková, Marian Simko

We present GEST – a new manually created dataset designed to measure gender-stereotypical reasoning in language models and machine translation systems. GEST contains samples for 16 gender stereotypes about men and women (e.g., Women are beautiful, Men are leaders) that are compatible with the English language and 9 Slavic languages. The definition of said stereotypes was informed by gender experts. We used GEST to evaluate English and Slavic masked LMs, English generative LMs, and machine translation systems. We discovered significant and consistent amounts of gender-stereotypical reasoning in almost all the evaluated models and languages. Our experiments confirm the previously postulated hypothesis that the larger the model, the more stereotypical it usually is.

Generation

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Retrieve-Plan-Generation: An Iterative Planning and Answering Framework for Knowledge-Intensive LLM Generation

Yuanjie Lyu, Zihan Niu, Zheyong Xie, Chao Zhang, Tong Xu, Yang Wang, Enhong Chen

Despite the significant progress of large language models (LLMs) in various tasks, they often produce factual errors due to their limited internal knowledge. Retrieval-Augmented Generation (RAG), which enhances LLMs with external knowledge sources, offers a promising solution. However, these methods can be misled by irrelevant paragraphs in retrieved documents. Due to the inherent uncertainty in LLM generation, inputting the entire document may introduce off-topic information, causing the model to deviate from the central topic and affecting the relevance of the generated content. To address these issues, we propose the Retrieve-Plan-Generation (RPG) framework. RPG generates plan tokens to guide subsequent generation in the plan stage. In the answer stage, the model selects relevant fine-grained paragraphs based on the plan and uses them for further answer generation. This plan-answer process is repeated iteratively until completion, enhancing generation relevance by focusing on specific topics. To implement this framework efficiently, we utilize a simple but effective multi-task prompt-tuning method, enabling the existing LLMs to handle both planning and answering. We comprehensively compare RPG with baselines across 5 knowledge-intensive generation tasks, demonstrating the effectiveness of our approach.

Industry

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Generating Vehicular Icon Descriptions and Indications Using Large Vision-Language Models

James Fletcher, Nicholas Dethnen, Seyed Nima Tayarani Bathaei, Ajun An, Heidar Davoudi, Ron Di Carlantonio, Gary Farmaner

To enhance a question-answering system for automotive drivers, we tackle the problem of automatic generation of icon image descriptions. The descriptions can match the drivers query about the icon appearing on the dashboard and tell the driver what is happening so that they may take an appropriate action. We use three state-of-the-art large vision-language models to generate both visual and functional descriptions based on the icon image and its context information in the car manual. Both zero-shot and few-shot prompts are used. We create a dataset containing over 400 icons with their ground-truth descriptions and use it to evaluate model-generated descriptions across several performance metrics. Our evaluation shows that two of these models (GPT-4o and Claude 3.5) performed well on this task, while the third model (LLaVA-NEXT) performs poorly.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Arreec's MergeKit: A Toolkit for Merging Large Language Models

Charles Goddard, Shamane Siriwardhana, Malikeh Elghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, Jacob

Solawetz

The rapid growth of open-source language models provides the opportunity to merge model checkpoints, combining their parameters to improve performance and versatility. Advances in transfer learning have led to numerous task-specific models, which model merging can integrate into powerful multitask models without additional training. MergeKit is an open-source library designed to support this process with an efficient and extensible framework suitable for any hardware. It has facilitated the merging of thousands of models, contributing to some of the world's most powerful open-source model checkpoints.

(Nov 14): 13:00:14:00 (Afternoon) - Gather

QDyLoRA: Quantized Dynamic Low-Rank Adaptation for Efficient Large Language Model Tuning

Hossein Rajabzadeh, Mojtaba Valipour, Tianshu Zhu, Marzieh S. Tahaei, Hyock Ju Kwon, Ali Ghodsi, Boxing Chen, Mehdi Rezagholizadeh
Finetuning large language models requires huge GPU memory, restricting the choice to acquire larger models. While the quantized version of the Low-Rank Adaptation technique, named QLoRA, significantly alleviates this issue, finding the efficient LoRA rank is still challenging. Moreover, QLoRA is trained on a pre-defined rank and, therefore, cannot be reconfigured for its lower ranks without requiring further finetuning steps. This paper proposes QDyLoRA -Quantized Dynamic Low-Rank Adaptation-, as an efficient quantization approach for dynamic low-rank adaptation. Motivated by Dynamic LoRA, QDyLoRA is able to efficiently finetune LLMs on a set of pre-defined LoRA ranks. QDyLoRA enables fine-tuning Falcon-40b for ranks 1 to 64 on a single 32 GB V100-GPU through one round of fine-tuning. Experimental results show that QDyLoRA is competitive to QLoRA and outperforms when employing its optimal rank.

(Nov 14): 13:00:14:00 (Afternoon) - Gather

Retrieval Augmented Spelling Correction for E-Commerce Applications

Xuan Guo, Rohit Patki, Dante Everaert, Christopher Potts

The rapid introduction of new brand names into everyday language poses a unique challenge for spelling correction services, which must distinguish between genuine misspellings and correct-but-unconventional brand names. This paper integrates the Retrieval-Augmented Generation framework with Large Language Models, to enable a spelling correction system to recognize new brands. Through quantitative evaluation (F1 score) and qualitative error analyses, we find improvements in spelling correction utilizing the RAG framework beyond a standalone LLM. We also demonstrate the value of additional finetuning that incorporates retrieved context, to allow LLMs better use context from retriever during text generation.

(Nov 14): 13:00:14:00 (Afternoon) - Gather

PDFTriage: Question Answering over Long, Structured Documents

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Seunghyun Yoon, Ryan A. Rossi, Franck Dernoncourt

Large Language Models (LLMs) have issues with document question answering (QA) in situations where the document is unable to fit in the small context length of an LLM. To overcome this issue, most existing works focus on retrieving the relevant context from the document, representing them as plain text. However, documents such as PDFs, web pages, and presentations are naturally structured with different pages, tables, sections, and so on. Representing such structured documents as plain text is incongruous with the users' mental model of these documents with rich structure. When a system has to query the document for context, this incongruity is brought to the fore, and seemingly trivial questions can trip up the QA system. To bridge this fundamental gap in handling structured documents, we propose an approach called PDFTriage that enables models to retrieve the context based on either structure or content. Our experiments demonstrate the effectiveness of the proposed PDFTriage-augmented models across several classes of questions where existing retrieval-augmented LLMs fail. To facilitate further research on this fundamental problem, we release our benchmark dataset consisting of 900+ human-generated questions over 80 structured documents from 10 different categories of question types for document QA. Our code and datasets will be released soon on Github.

(Nov 14): 13:00:14:00 (Afternoon) - Gather

Fairness-Aware Online Positive-Unlabeled Learning

Hoin Jung, Xiaogiao Wang

Machine learning applications for text classification are increasingly used in domains such as toxicity and misinformation detection in online settings. However, obtaining precisely labeled data for training remains challenging, particularly because not all problematic instances are reported. Positive-Unlabeled (PU) learning, which uses only labeled positive and unlabeled samples, offers a solution for these scenarios. A significant concern in PU learning, especially in online settings, is fairness: specific groups may be disproportionately classified as problematic. Despite its importance, this issue has not been explicitly addressed in research. This paper aims to bridge this gap by investigating the fairness of PU learning in both offline and online settings. We propose a novel approach to achieve more equitable results by extending PU learning methods to online learning for both linear and non-linear classifiers and analyzing the impact of the online setting on fairness. Our approach incorporates a convex fairness constraint during training, applicable to both offline and online PU learning. Our solution is theoretically robust, and experimental results demonstrate its efficacy in improving fairness in PU learning in text classification.

Information Extraction

(Nov 14): 13:00:14:00 (Afternoon) - Room: Gather

(Nov 14): 13:00:14:00 (Afternoon) - Gather

One2Set + Large Language Model: Best Partners for Keyphrase Generation

Liangying Shao, Liang Zhang, Minlong Peng, Guoqi Ma, Hao Yue, Mingming Sun, Jinsong Su

Keyphrase generation (KPG) aims to automatically generate a collection of phrases representing the core concepts of a given document. The dominant paradigm in KPG includes one2seq and one2set. Recently, there has been increasing interest in applying large language models (LLMs) to KPG. Our preliminary experiments reveal that it is challenging for a single model to excel in both recall and precision. Further analysis shows that: 1) the one2set paradigm owns the advantage of high recall, but suffers from improper assignments of supervision signals during training; 2) LLMs are powerful in keyphrase selection, but existing selection methods often make redundant selections. Given these observations, we introduce a generate-then-select framework decomposing KPG into two steps, where we adopt a one2set-based model as generator to produce candidates and then use an LLM as selector to select keyphrases from these candidates. Particularly, we make two important improvements on our generator and selector: 1) we design an Optimal Transport-based assignment strategy to address the above improper assignments; 2) we model the keyphrase selection as a sequence labeling task to alleviate redundant selections. Experimental results on multiple benchmark datasets show that our framework significantly surpasses state-of-the-art models, especially in absent keyphrase prediction.

(Nov 14): 13:00:14:00 (Afternoon) - Gather

Lifelong Event Detection via Optimal Transport

Viet Dao, Van-Cuong Pham, Quyen Tran, Thanh-Thien Le, Linh Van Ngo, Thien Huu Nguyen

Continual Event Detection (CED) poses a formidable challenge due to the catastrophic forgetting phenomenon, where learning new tasks (with new coming event types) hampers performance on previous ones. In this paper, we introduce a novel approach, Lifelong Event Detection via Optimal Transport (**LEDOT**), that leverages optimal transport principles to align the optimization of our classification module with the intrinsic nature of each class, as defined by their pre-trained language modeling. Our method integrates replay sets, prototype latent representations, and an innovative Optimal Transport component. Extensive experiments on MAVEN and ACE datasets demonstrate LEDOT's superior performance, consistently outperforming state-of-the-art baselines. The results underscore LEDOT as a pioneering solution in continual event detection, offering a more effective and nuanced approach to addressing catastrophic forgetting in evolving environments.

Information Retrieval and Text Mining

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

MTA4DPR: Multi-Teaching-Assistants Based Iterative Knowledge Distillation for Dense Passage Retrieval

Qixi Lu, Gongbo Tang

Although Dense Passage Retrieval (DPR) models have achieved significantly enhanced performance, their widespread application is still hindered by the demanding inference efficiency and high deployment costs. Knowledge distillation is an efficient method to compress models, which transfers knowledge from strong teacher models to weak student models. Previous studies have proved the effectiveness of knowledge distillation in DPR. However, there often remains a significant performance gap between the teacher and the distilled student. To narrow this performance gap, we propose MTA4DPR, a Multi-Teaching-Assistants based iterative knowledge distillation method for Dense Passage Retrieval, which transfers knowledge from the teacher to the student with the help of multiple assistants in an iterative manner; with each iteration, the student learns from more performant assistants and more difficult data. The experimental results show that our 66M student model achieves the state-of-the-art performance among models with same parameters on multiple datasets, and is very competitive when compared with larger, even LLM-based, DPR models.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Unsupervised Domain Adaptation for Keyphrase Generation using Citation Contexts

Florian Boudin, Akiko Aizawa

Adapting keyphrase generation models to new domains typically involves few-shot fine-tuning with in-domain labeled data. However, annotating documents with keyphrases is often prohibitively expensive and impractical, requiring expert annotators. This paper presents silk, an unsupervised method designed to address this issue by extracting silver-standard keyphrases from citation contexts to create synthetic labeled data for domain adaptation. Extensive experiments across three distinct domains demonstrate that our method yields high-quality synthetic samples, resulting in significant and consistent improvements in in-domain performance over strong baselines.

Interpretability and Analysis of Models for NLP

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Knowledge Conflicts for LLMs: A Survey

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru WANG, Yue Zhang, Wei Xu

This survey provides an in-depth analysis of knowledge conflicts for large language models (LLMs), highlighting the complex challenges they encounter when blending contextual and parametric knowledge. Our focus is on three categories of knowledge conflicts: context-memory, inter-context, and intra-memory conflict. These conflicts can significantly impact the trustworthiness and performance of LLMs, especially in real-world applications where noise and misinformation are common. By categorizing these conflicts, exploring the causes, examining the behavior of LLMs under such conflicts, and reviewing available solutions, this survey aims to shed light on strategies for improving the robustness of LLMs, thereby serving as a valuable resource for advancing research in this evolving area.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Can We Trust the Performance Evaluation of Uncertainty Estimation Methods in Text Summarization?

Jianfeng He, Runing Yang, Linlin Yu, Changbin Li, Ruoxi Jia, Feng Chen, Ming Jin, Chang-Tien Lu

Text summarization, a key natural language generation (NLG) task, is vital in various domains. However, the high cost of inaccurate summaries in risk-critical applications, particularly those involving human-in-the-loop decision-making, raises concerns about the reliability of uncertainty estimation on text summarization (UE-TS) evaluation methods. This concern stems from the dependency of uncertainty model metrics on diverse and potentially conflicting NLG metrics. To address this issue, we introduce a comprehensive UE-TS benchmark incorporating 31 NLG metrics across four dimensions. The benchmark evaluates the uncertainty estimation capabilities of two large language models and one pre-trained language model on three datasets, with human-annotation analysis incorporated where applicable. We also assess the performance of 14 common uncertainty estimation methods within this benchmark. Our findings emphasize the importance of considering multiple uncorrelated NLG metrics and diverse uncertainty estimation methods to ensure reliable and efficient evaluation of UE-TS techniques. Our code and data are available: <https://github.com/he159ok/Benchmark-of-Uncertainty-Estimation-Methods-in-Text-Summarization>.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Can We Instruct LLMs to Compensate for Position Bias?

Meiru Zhang, Zaiqiao Meng, Nigel Collier

Position bias in large language models (LLMs) leads to difficulty in accessing information retrieved from the retriever, thus downgrading the effectiveness of Retrieval-Augmented Generation (RAG) approaches in open-question answering. Recent studies reveal that this bias is related to disproportional attention across the context. In this work, we examine how to direct LLMs to allocate more attention towards a selected segment of the context through prompting, aiming to compensate for the shortage of attention. We find that language models do not have relative position awareness of the context but can be directed by promoting instruction with an exact document index. Our analysis contributes to a deeper understanding of position bias in LLMs and provides a pathway to mitigate this bias by instruction, thus benefiting LLMs in locating and utilizing relevant information from retrieved documents in RAG applications. The code and data in our study have been made publicly available.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Zero-Resource Hallucination Prevention for Large Language Models

Junyu Luo, Cao Xiao, Fenglong Ma

The prevalent use of large language models (LLMs) in various domains has drawn attention to the issue of “hallucination”, which refers to instances where LLMs generate factually inaccurate or ungrounded information. Existing techniques usually identify hallucinations post-generation that cannot prevent their occurrence and suffer from inconsistent performance due to the influence of the instruction format and model style. In this paper, we introduce a novel pre-detection self-evaluation technique, referred to as SELF-FAMILIARITY, which focuses on evaluating the model’s familiarity with the concepts present in the input instruction and withholding the generation of responses in case of unfamiliar concepts under the zero-resource setting, where external ground-truth or background information is not available. We also propose a new dataset Concept-7 focusing on the hallucinations caused by limited inner knowledge. We validate SELF-FAMILIARITY across four different large language models, demonstrating consistently superior performance compared to existing techniques. Our findings propose a significant shift towards preemptive strategies for hallucination mitigation in LLM assistants, promising improvements in reliability, applicability, and interpretability.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Commonsense Knowledge Editing Based on Free-Text in LLMs

Xiusheng Huang, Yequan Wang, Jun Zhao, Kang Liu

Knowledge editing technology is crucial for maintaining the accuracy and timeliness of large language models (LLMs). However, the setting of this task overlooks a significant portion of commonsense knowledge based on free-text in the real world, characterized by broad knowledge scope, long content and non instantiation. The editing objects of previous methods (e.g., MEMIT) were single token or entity, which were not suitable for commonsense knowledge in free-text form. To address the aforementioned challenges, we conducted experiments from two perspectives: knowledge localization and knowledge editing. Firstly, we introduced Knowledge Localization for Free-Text(KLFT) method, revealing the challenges associated with the distribution of commonsense knowledge in MLP and Attention layers, as well as in decentralized distribution. Next, we propose a Dynamics-aware Editing Method(DEM), which utilizes a Dynamics-aware Module to locate the parameter positions corresponding to commonsense knowledge, and uses Knowledge Editing Module to update knowledge. The DEM method fully explores the potential of the MLP and Attention layers, and successfully edits commonsense knowledge based on free-text. The experimental results indicate that the DEM can achieve excellent editing performance.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Are LLMs Good Annotators for Discourse-level Event Relation Extraction?

Kangda Wei, Aayush Gautam, Ruihong Huang

Large Language Models (LLMs) have demonstrated proficiency in a wide array of natural language processing tasks. However, its effectiveness over discourse-level event relation extraction (ERE) tasks remains unexplored. In this paper, we assess the effectiveness of LLMs in addressing discourse-level ERE tasks characterized by lengthy documents and intricate relations encompassing coreference, temporal, causal, and subevent types. Evaluation is conducted using a commercial model, GPT-3.5, and an open-source model, LLaMA-2. Our study reveals a notable underperformance of LLMs compared to the baseline established through supervised learning. Although Supervised Fine-Tuning (SFT) can improve LLMs performance, it does not scale well compared to the smaller supervised baseline model. Our quantitative and qualitative analysis shows that LLMs have several weaknesses when applied for extracting event relations, including a tendency to fabricate event mentions, and failures to capture transitivity rules among relations, detect long distance relations, or comprehend contexts with dense event mentions.

Language Modeling

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Knowledge Editing in Language Models via Adapted Direct Preference Optimization

Amit Rozner, Barak Battash, Lior Wolf, Ofir Lindenbaum

Large Language Models (LLMs) can become outdated over time as they may lack updated world knowledge, leading to factual knowledge errors and gaps. Knowledge Editing (KE) aims to overcome this challenge using weight updates that do not require expensive retraining. We propose treating KE as an LLM alignment problem. Toward this goal, we introduce Knowledge Direct Preference Optimization (KDPO), a variation of the Direct Preference Optimization (DPO) that is more effective for knowledge modifications. Our method is based on an online approach that continually updates the knowledge stored in the model. We use the current knowledge as a negative sample and the new knowledge we want to introduce as a positive sample in a process called DPO. We also use teacher-forcing for negative sample generation and optimize using the positive sample, which helps maintain localized changes. We tested our KE method on various datasets and models, comparing it to several cutting-edge methods, with 100 and 500 sequential edits. Additionally, we conducted an ablation study comparing our method to the standard DPO approach. Our experimental results show that our modified DPO method allows for more refined KE, achieving similar or better performance compared to previous methods.

Linguistic Theories, Cognitive Modeling and Psycholinguistics

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions

Hakyung Sung, Kristopher Kyle

This study evaluates the effectiveness of pre-trained language models in identifying argument structure constructions, important for modeling both first and second language learning. We examine three methodologies: (1) supervised training with RoBERTa using a gold-standard ASC treebank, including by-tag accuracy evaluation for sentences from both native and non-native English speakers, (2) prompt-guided annotation with GPT-4, and (3) generating training data through prompts with GPT-4, followed by RoBERTa training. Our findings indicate that RoBERTa trained on gold-standard data shows the best performance. While data generated through GPT-4 enhances training, it does not exceed the benchmarks set by gold-standard data.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Multi-Loss Fusion: Angular and Contrastive Integration for Machine-Generated Text Detection*Iqra Zahid, Yue Chang, Youcheng Sun, Riza Batista-Navarro*

Modern natural language generation (NLG) systems have led to the development of synthetic human-like open-ended texts, posing concerns as to who the original author of a text is. To address such concerns, we introduce DeB-Ang: the utilisation of a custom DeBERTa model with angular loss and contrastive loss functions for effective class separation in neural text classification tasks. We expand the application of this model on binary machine-generated text detection and multi-class neural authorship attribution. We demonstrate improved performance on many benchmark datasets whereby the accuracy for machine-generated text detection was increased by as much as 38.04% across all datasets.

Low-resource Methods for NLP

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Cost-Performance Optimization for Processing Low-Resource Language Tasks Using Commercial LLMs*Arijit Nag, Animesh Mukherjee, Niloy Ganguly, Soumen Chakrabarti*

Large Language Models (LLMs) exhibit impressive zero/few-shot inference and generation quality for high-resource languages (HRLs). A few of them have been trained on low-resource languages (LRLs) and give decent performance. Owing to the prohibitive costs of training LLMs, they are usually used as a network service, with the client charged by the count of input and output tokens. The number of tokens strongly depends on the script and language, as well as the LLM's subword vocabulary. We show that LRLs are at a pricing disadvantage, because the well-known LLMs produce more tokens for LRLs than HRLs. This is because most currently popular LLMs are optimized for HRL vocabularies. Our objective is to level the playing field: reduce the cost of processing LRLs in contemporary LLMs while ensuring that predictive and generative qualities are not compromised. As means to reduce the number of tokens processed by the LLM, we consider code-mixing, translation, and transliteration of LRLs to HRLs. We perform an extensive study using the IndicXTREME classification and six generative tasks dataset, covering 15 Indic and 3 other languages, while using GPT-4 (one of the costliest LLM services released so far¹²) as a commercial LLM. We observe and analyze interesting patterns involving token count, cost, and quality across a multitude of languages and tasks. We show that choosing the best policy to interact with the LLM can reduce cost by 90% while giving better or comparable performance, compared to communicating with the LLM in the original LRL.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Householder Pseudo-Rotation: A Novel Approach to Activation Editing in LLMs with Direction-Magnitude Perspective*Van-Cuong Pham, Thien Huu Nguyen*

Activation Editing, which involves directly editing the internal representations of large language models (LLMs) to alter their behavior and achieve desired properties, has emerged as a promising area of research. Existing works primarily treat LLMs' activations as points in space and modify them by adding steering vectors. We show that doing so would break the magnitude consistency of the activation vectors in LLMs. To overcome this shortcoming, we propose a novel editing method that views activations in terms of their directions and magnitudes. Our method, which we name Householder Pseudo-Rotation (HPR), mimics the rotation transformation, thus preserving activation norm and resulting in an improved performance on various safety benchmarks.

Machine Learning for NLP

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

A Training Data Recipe to Accelerate A* Search with Language Models*Devaansh Gupta, Boyang Li*

Combining Large Language Models (LLMs) with heuristic search algorithms like A* holds the promise of enhanced LLM reasoning and scalable inference. To accelerate training and reduce computational demands, we investigate the coreset selection problem for the training data of LLM heuristic learning. Few methods to learn the heuristic functions consider the interaction between the search algorithm and the machine learning model. In this work, we empirically disentangle the requirements of A* search algorithm from the requirements of the LLM to generalise on this task. Surprisingly, we find an overlap between their requirements; A* requires more accurate predictions on search nodes near the goal, and LLMs need the same set of nodes for effective generalisation. With these insights, we derive a data-selection distribution for learning LLM-based heuristics. On three classical planning domains, maze navigation, Sokoban, and sliding tile puzzles, our technique reduces the number of iterations required to find the solutions by up to 15x, with a wall-clock speed-up of search up to 5x. The code has been made available at https://github.com/devaansh100/a_star.

Machine Translation

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Reconsidering Sentence-Level Sign Language Translation*Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, David Uithas*

Historically, sign language machine translation has been posed as a sentence-level task: datasets consisting of continuous narratives are chopped up and presented to the model as isolated clips. In this work, we explore the limitations of this task framing. First, we survey a number of linguistic phenomena in sign languages that depend on discourse-level context. Then as a case study, we perform the first human

¹²<http://tinyurl.com/llm-costing>

baseline for sign language translation that actually substitutes a human into the machine learning task framing, rather than provide the human with the entire document as context. This human baseline—for ASL to English translation on the How2Sign dataset—shows that for 33% of sentences in our sample, our fluent Deaf signer annotators were only able to understand key parts of the clip in light of additional discourse-level context. These results underscore the importance of understanding and sanity checking examples when adapting machine learning to new domains.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Dual-teacher Knowledge Distillation for Low-frequency Word Translation

yifan guo, Hongying ZAN, Hongfei Xu

Neural Machine Translation (NMT) models are trained on parallel corpora with unbalanced word frequency distribution. As a result, NMT models are likely to prefer high-frequency words than low-frequency ones despite low-frequency word may carry the crucial semantic information, which may hamper the translation quality once they are neglected. The objective of this study is to enhance the translation of meaningful but low-frequency words. Our general idea is to optimize the translation of low-frequency words through knowledge distillation. Specifically, we employ a low-frequency teacher model that excels in translating low-frequency words to guide the learning of the student model. To remain the translation quality of high-frequency words, we further introduce a dual-teacher distillation framework, leveraging both the low-frequency and high-frequency teacher models to guide the student model's training. Our single-teacher distillation method already achieves a +0.64 BLEU improvements over the state-of-the-art method on the WMT 16 English-to-German translation task on the low-frequency test set. Whilst our dual-teacher framework leads to +0.87, +1.24, +0.47, +0.87 and +0.86 BLEU improvements on the IWSLT 14 German-to-English, WMT 16 English-to-German, WMT 15 English-to-Czech, WMT 14 English-to-French and WMT 18 Chinese-to-English tasks respectively compared to the baseline, while maintaining the translation performance of high-frequency words.

Multimodality and Language Grounding to Vision, Robotics and Beyond

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

RAI: Injecting Implicit Bias for Text-To-Image Prompt Refinement Models

Ziyi Kou, Shichao Pei, Meng Jiang, Xiangliang Zhang

Text-to-image prompt refinement (T2I-Refine) aims to rephrase or extend an input prompt with more descriptive details that can be leveraged to generate images with higher quality. In this paper, we study an adversarial prompt attacking problem for T2I-Refine, where the goal is to implicitly inject specific concept bias to the input prompts during the refinement process so that the generated images, still with higher quality, are explicitly biased to the target group. Our study is motivated by the limitation of current T2I-Refine research that lacks of explorations on the potential capacity of T2I-Refine models to provide prompt refinement service in a biased or advertising manner. To address the limitations, we develop RAI, a prompt refinement and attacking framework that attacks input prompts with intentionally selected adversarial replacements by optimizing a token distribution matrix based on the text-to-image finetuning strategy with a token-level bias obfuscation loss as regularization. We evaluate RAI on a large-scale text-to-image dataset with various concepts as target in both in-domain and transfer-domain scenarios. The evaluation results demonstrate that, compared to other T2I-Refine schemes, RAI is well capable of implicitly attacking input prompts to generate images with higher quality and explicit visual bias towards specific concept group.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Android in the Zoo: Chain-of-Action-Thought for GUI Agents

Jiwen Zhang, Jihao Wu, Teng Yihua, Minghui Liao, Nuo Xu, Xiao Xiao, zhongyu wei, Duyu Tang

Large language model (LLM) leads to a surge of autonomous GUI agents for smartphone, which completes a task triggered by natural language through predicting a sequence of actions of API. Even though the task highly relies on past actions and visual observations, existing studies typically consider little semantic information carried out by intermediate screenshots and screen operations. To address this, this work presents Chain-of-Action-Thought (dubbed CoAT), which takes the description of the previous actions, the current screen, and more importantly the action thinking of what actions should be performed and the outcomes led by the chosen action. We demonstrate that, in a zero-shot setting upon three off-the-shelf LLMs, CoAT significantly improves the action prediction compared to previous proposed context modeling. To further facilitate the research in this line, we construct a dataset Android-In-The-Zoo (AitZ), which contains 18,643 screen-action pairs together with chain-of-action-thought annotations. Experiments show that fine-tuning a 1B model (i.e. AUTO-UI-base) on our AitZ dataset achieves on-par performance with CogAgent-Chat-18B.

NLP Applications

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

DocHiNet: A Large and Diverse Dataset for Document Hierarchy Parsing

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, Cong Yao

Parsing documents from pixels, such as pictures and scanned PDFs, into hierarchical structures is extensively demanded in the daily routines of data storage, retrieval and understanding. However, previously the research on this topic has been largely hindered since most existing datasets are small-scale, or contain documents of only a single type, which are characterized by a lack of document diversity. Moreover, there is a significant discrepancy in the annotation standards across datasets. In this paper, we introduce a large and diverse document hierarchy parsing (DHP) dataset to compensate for the data scarcity and inconsistency problem. We aim to set a new standard as a more practical, long-standing benchmark. Meanwhile, we present a new DHP framework designed to grasp both fine-grained text content and coarse-grained pattern at layout element level, enhancing the capacity of pre-trained text-layout models in handling the multi-page and multi-level challenges in DHP. Through exhaustive experiments, we validate the effectiveness of our proposed dataset and method.

(Nov 14): 13:0014:00 (Afternoon) - Gather

How Do Humans Write Code? Large Models Do It the Same Way Too

Long Li, Xuzheng He, Haozhe Wang, Linlin Wang, Liang He

Program-of-Thought (PoT) replaces natural language-based Chain-of-Thought (CoT) as the most popular method in Large Language Models (LLMs) mathematical reasoning tasks by utilizing external tool calls to circumvent computational errors. However, our evaluation of the

GPT-4 and Llama series reveals that using PoT introduces more reasoning errors, such as incorrect formulas or flawed logic, compared to CoT. To address this issue, we propose Human-Think Language (HTL), which leverages a suite of strategies that help integrate PoT and CoT, encompassing: (1) a new generation paradigm that uses full CoT reasoning to control code generation. (2) Focus Attention, that directs model attention to the CoT reasoning during PoT to generate more logical code. (3) reinforcement learning that utilizes the accuracy of both CoT and PoT responses as rewards to prevent repetitive reasoning steps in LLMs when solving difficult math problems. Our method achieves an average improvement of 6.5% on the Llama-Base model and 4.3% on the Mistral-Base model across 8 mathematical calculation datasets. It also shows significant effectiveness on five out-of-domain datasets by controlling the model's information flow, exhibiting strong transferability. Additionally, HTL shows the most significant improvement in non-mathematical natural language inference task, contributing to a unified reasoning task framework.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

C-LLM: Learn to Check Chinese Spelling Errors Character by Character

Kunting Li, Yong Hu, Liang He, Fandong Meng, Jie Zhou

Chinese Spell Checking (CSC) aims to detect and correct spelling errors in sentences. Despite Large Language Models (LLMs) exhibit robust capabilities and are widely applied in various tasks, their performance on CSC is often unsatisfactory. We find that LLMs fail to meet the Chinese character-level constraints of the CSC task, namely equal length and phonetic similarity, leading to a performance bottleneck. Further analysis reveals that this issue stems from the granularity of tokenization, as current mixed character-word tokenization struggles to satisfy these character-level constraints. To address this issue, we propose C-LLM, a Large Language Model-based Chinese Spell Checking method that learns to check errors Character by Character. Character-level tokenization enables the model to learn character-level alignment, effectively mitigating issues related to character-level constraints. Furthermore, CSC is simplified to replication-dominated and substitution-supplemented tasks. Experiments on two CSC benchmarks demonstrate that C-LLM achieves a 2.1% enhancement in general scenarios and a significant 12% improvement in vertical domain scenarios compared to existing methods, establishing state-of-the-art performance.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

Joint Pre-Encoding Representation and Structure Embedding for Efficient and Low-Resource Knowledge Graph Completion

Chenyu Qiu, Pengjiang Qian, Chuang Wang, Jian Yao, Li Liu, Fang wei, Eddie Y.K. Eddie

Knowledge graph completion (KGC) aims to infer missing or incomplete parts in knowledge graph. The existing models are generally divided into structure-based and description-based models, among description-based models often require longer training and inference times as well as increased memory usage. In this paper, we propose Pre-Encoded Masked Language Model (PEMLM) to efficiently solve KGC problem. By encoding textual descriptions into semantic representations before training, the necessary resources are significantly reduced. Furthermore, we introduce a straightforward but effective fusion framework to integrate structural embedding with pre-encoded semantic description, which enhances the model's prediction performance on the WN18RR (MRR+5.4% and Hits@1+6.4%) and UMLS datasets. Compared to existing models, we have increased inference speed by 30x and reduced training memory by approximately 60%.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

PKAD: Pre-trained Knowledge is All You Need to Detect and Mitigate Textual Backdoor Attacks

Yu Chen, Qi Cao, Kaike Zhang, Xuchao Liu, Huawei Shen

In textual backdoor attacks, attackers insert poisoned samples with triggered inputs and target labels into training datasets to manipulate model behavior, threatening the model's security and reliability. Current defense methods can generally be categorized into inference-time and training-time ones. The former often requires a part of clean samples to set detection thresholds, which may be hard to obtain in practical application scenarios, while the latter usually requires an additional retraining or unlearning process to get a clean model, significantly increasing training costs. To avoid these drawbacks, we focus on developing a practical defense method before model training without using any clean samples. Our analysis reveals that with the help of a pre-trained language model (PLM), poisoned samples, different from clean ones, exhibit mismatched relationship and shared characteristics. Based on these observations, we further propose a two-stage poison detection strategy solely leveraging insights from PLM before model training. Extensive experiments confirm our approach's effectiveness, achieving better performance than current leading methods more swiftly. Our code is available at <https://github.com/Ascan/PKAD>.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

HyperBERT: Mixing Hypergraph-Aware Layers with Language Models for Node Classification on Text-Attributed Hypergraphs

Adrián Basagaña, Pietro Lio, Gos Micklem

Hypergraphs are characterized by complex topological structure, representing higher-order interactions among multiple entities through hyperedges. Lately, hypergraph-based deep learning methods to learn informative data representations for the problem of node classification on text-attributed hypergraphs have garnered increasing research attention. However, existing methods struggle to simultaneously capture the full extent of hypergraph structural information and the rich linguistic attributes inherent in the nodes' attributes, which largely hampers their effectiveness and generalizability. To overcome these challenges, we explore ways to further augment a pretrained BERT model with specialized hypergraph-aware layers for the task of node classification. Such layers introduce higher-order structural inductive bias into the language model, thus improving the model's capacity to harness both higher-order context information from the hypergraph structure and semantic information present in text. In this paper, we propose a new architecture, HyperBERT, a mixed text-hypergraph model which simultaneously models hypergraph relational structure while maintaining the high-quality text encoding capabilities of a pre-trained BERT. Notably, HyperBERT presents results that achieve a new state-of-the-art on five challenging text-attributed hypergraph node classification benchmarks.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

How Do Your Code LLMs perform? Empowering Code Instruction Tuning with Really Good Data

Yejie Wang, Kegang He, Dayuan Fu, Zhuoma GongQue, Heyang Xu, Yanxu Chen, Zhuxu Wang, Yujia Fu, Guanting Dong, Muxi Diao, Jingang Wang, Mengdi Zhang, XunLiang Cai, Weiran Xu

Recently, there has been a growing interest in studying how to construct better code instruction tuning data. However, we observe Code models trained with these datasets exhibit high performance on HumanEval but perform worse on other benchmarks such as LiveCodeBench. Upon further investigation, we find that many datasets suffer from severe data leakage. After cleaning up most of the leaked data, some well-known high-quality datasets perform poorly. This discovery reveals a new challenge: identifying which dataset genuinely qualify as high-quality code instruction data. To address this, we propose an efficient code data pruning strategy for selecting good samples. Our approach is based on three dimensions: instruction complexity, response quality, and instruction diversity. Based on our selected data, we present XCoder, a family of models finetuned from LLaMA3. Our experiments show XCoder achieves new state-of-the-art performance using fewer training data, which verify the effectiveness of our data strategy. Moreover, we perform a comprehensive analysis on the data composition and find existing code datasets have different characteristics according to their construction methods, which provide new insights for future code LLMs.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

RoBERT2VecTM: A Novel Approach for Topic Extraction in Islamic Studies

Sania Aftar, Amina El Ganadi, Luca Gagliardelli, Sonia Bergamaschi

Investigating "Hadith" texts, crucial for theological studies and Islamic jurisprudence, presents challenges due to the linguistic complexity of Arabic, such as its complex morphology. In this paper, we propose an innovative approach to address the challenges of topic modeling in Hadith studies by utilizing the Contextualized Topic Model (CTM). Our study introduces RoBERT2VecTM, a novel neural-based approach that combines the RoBERTa transformer model with Doc2Vec, specifically targeting the semantic analysis of "Matn" (the actual content). The methodology outperforms many traditional state-of-the-art NLP models by generating more coherent and diverse Arabic topics. The diversity of the generated topics allows for further categorization, deepening the understanding of discussed concepts. Notably, our research highlights the critical impact of lemmatization and stopwords in enhancing topic modeling. This breakthrough marks a significant stride in applying NLP to non-Latin languages and opens new avenues for the nuanced analysis of complex religious texts.

(Nov 14): 13:0014:00 (Afternoon) - Gather

TRoTR: A Framework for Evaluating the Re-contextualization of Text Reuse

Francesco Periti, Pierluigi Cassotti, Stefano Montanelli, Nina Tahmasebi, Dominik Schlechtweg

Current approaches for detecting text reuse do not focus on recontextualization, i.e., how the new context(s) of a reused text differs from its original context(s). In this paper, we propose a novel framework called TRoTR that relies on the notion of topic relatedness for evaluating the diachronic change of context in which text is reused. TRoTR includes two NLP tasks: TRiC and TRaC. TRiC is designed to evaluate the topic relatedness between a pair of recontextualizations. TRaC is designed to evaluate the overall topic variation within a set of recontextualizations. We also provide a curated TRoTR benchmark of biblical text reuse, human-annotated with topic relatedness. The benchmark exhibits an inter-annotator agreement of .811. We evaluate multiple, established SBERT models on the TRoTR tasks and find that they exhibit greater sensitivity to textual similarity than topic relatedness. Our experiments show that fine-tuning these models can mitigate such a kind of sensitivity.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Sing it, Narrate it: Quality Musical Lyrics Translation

Zhuorui Ye, Jinhan Li, Rongwu Xu

Translating lyrics for musicals presents unique challenges due to the need to ensure high translation quality while adhering to singability requirements such as length and rhyme. Existing song translation approaches often prioritize these singability constraints at the expense of translation quality, which is crucial for musicals. This paper aims to enhance translation quality while maintaining key singability features. Our method consists of three main components. First, we create a dataset to train reward models for the automatic evaluation of translation quality. Second, to enhance both singability and translation quality, we implement a two-stage training process with filtering techniques. Finally, we introduce an inference-time optimization framework for translating entire songs. Extensive experiments, including both automatic and human evaluations, demonstrate significant improvements over baseline methods and validate the effectiveness of each component in our approach.

Question Answering

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

CSLM: A Framework for Question Answering Dataset Generation through Collaborative Small Language Models

Yiming Wang, Yang Liu, Lingchen Wang, An Xiao

Collecting high-quality question-answer (QA) pairs is vital for the training of large language models (LLMs), yet this process is traditionally laborious and time-intensive. With the rapid evolution of LLMs, the potential for leveraging these models to autonomously generate QA pairs has become apparent, particularly through the use of large-scale models like GPT-4. However, the computational demands and associated costs often render such approaches prohibitive for the average researcher. Addressing this gap, we introduce the Collaborative Small Language Model Framework (CSLM), an innovative solution that combines a group of small-scaled, open-source LLMs to collaboratively produce QA pairs. Experiments on datasets of various domains show that CSLM unleashes the full potential of diverse small models to generate high-quality QA pairs, making it accessible to a broader range of researchers.

(Nov 14): 13:0014:00 (Afternoon) - Gather

ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, Xiaofeng Zhu

Uncertainty quantification (UQ) in natural language generation (NLG) tasks remains an open challenge, exacerbated by the closed-source nature of the latest large language models (LLMs). This study investigates applying conformal prediction (CP), which can transform any heuristic uncertainty notion into rigorous prediction sets, to black-box LLMs in open-ended NLG tasks. We introduce a novel uncertainty measure based on self-consistency theory, and then develop a conformal uncertainty criterion by integrating the uncertainty condition aligned with correctness into the CP algorithm. Empirical evaluations indicate that our uncertainty measure outperforms prior state-of-the-art methods. Furthermore, we achieve strict control over the correctness coverage rate utilizing 7 popular LLMs on 4 free-form NLG datasets, spanning general-purpose and medical scenarios. Additionally, the calibrated prediction sets with small size further highlights the efficiency of our method in providing trustworthy guarantees for practical open-ended NLG applications.

Resources and Evaluation

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

LawBench: Benchmarking Legal Knowledge of Large Language Models

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, Vincent Ng

We present LawBench, the first evaluation benchmark composed of 20 tasks aimed to assess the ability of Large Language Models (LLMs) to perform Chinese legal-related tasks. LawBench is meticulously crafted to enable precise assessment of LLMs' legal capabilities from three cognitive levels that correspond to the widely accepted Bloom's cognitive taxonomy. Using LawBench, we present a comprehensive evaluation of 21 popular LLMs and the first comparative analysis of the empirical results in order to reveal their relative strengths and weaknesses.

All data, model predictions and evaluation code are accessible from <https://github.com/open-compass/LawBench>.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

xCOMET-lite: Bridging the Gap Between Efficiency and Quality in Learned MT Evaluation Metrics

Daniil Larionov, Mikhail Seleznyov, Vasily Viskov, Alexander Panchenko, Steffen Eger

State-of-the-art trainable machine translation evaluation metrics like xCOMET achieve high correlation with human judgment but rely on large encoders (up to 10.7B parameters), making them computationally expensive and inaccessible to researchers with limited resources. To address this issue, we investigate whether the knowledge stored in these large encoders can be compressed while maintaining quality. We employ distillation, quantization, and pruning techniques to create efficient xCOMET alternatives and introduce a novel data collection pipeline for efficient black-box distillation. Our experiments show that, using quantization, xCOMET can be compressed up to three times with no quality degradation. Additionally, through distillation, we create an 278M-sized xCOMET-lite metric, which has only 2.6% of xCOMET-XXL parameters, but retains 92.1% of its quality. Besides, it surpasses strong small-scale metrics like COMET-22 and BLEURT-20 on the WMT22 metrics challenge dataset by 6.4%, despite using 50% fewer parameters. All code, dataset, and models are available online.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

Exploring the Capability of Multimodal LLMs with Yonkoma Manga: The YManga Dataset and Its Challenging Tasks

Qi Yang, Liang Yang, Jingji Zeng, Zhihao Yang, Hongfei Lin

Yonkoma Manga, characterized by its four-panel structure, presents unique challenges due to its rich contextual information and strong sequential features. To address the limitations of current multimodal large language models (MLLMs) in understanding this type of data, we create a novel dataset named YManga from the Internet. After filtering out low-quality content, we collect a dataset of 1,015 yonkoma strips, containing 10,150 human annotations. We then define three challenging tasks for this dataset: panel sequence detection, generation of the author's creative intention, and description generation for masked panels. These tasks progressively introduce the complexity of understanding and utilizing such image-text data. To the best of our knowledge, YManga is the first dataset specifically designed for yonkoma manga strips understanding. Extensive experiments conducted on this dataset reveal significant challenges faced by current multimodal large language models. Our results show a substantial performance gap between models and humans across all three tasks.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

TuringQ: Benchmarking AI Comprehension in Theory of Computation

Pardis Sadat Zahraei, Ehsaneddin Asgari

We present TuringQ, the first benchmark designed to evaluate the reasoning capabilities of large language models (LLMs) in the theory of computation. TuringQ consists of 4,006 undergraduate and graduate-level question-answer pairs, categorized into four difficulty levels and covering seven core theoretical areas. We evaluate several open-source LLMs, as well as GPT-4, using Chain of Thought prompting and expert human assessment. Additionally, we propose an automated LLM-based evaluation system that demonstrates competitive accuracy when compared to human evaluation. Fine-tuning a Llama3-8B model on TuringQ shows measurable improvements in reasoning ability and out-of-domain tasks such as algebra. TuringQ serves as both a benchmark and a resource for enhancing LLM performance in complex computational reasoning tasks. Our analysis offers insights into LLM capabilities and advances in AI comprehension of theoretical computer science.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

FRoG: Evaluating Fuzzy Reasoning of Generalized Quantifiers in LLMs

Yiyuan Li, Shichao Sun, Pengfei Liu

Fuzzy reasoning is vital due to the frequent use of imprecise information in daily contexts. However, the ability of current large language models (LLMs) to handle such reasoning remains largely uncharted. In this paper, we introduce a new benchmark, FRoG, for fuzzy reasoning, featuring real-world mathematical word problems that incorporate generalized quantifiers. Our experimental findings reveal that fuzzy reasoning continues to pose significant challenges for LLMs. Moreover, we find that existing methods designed to enhance reasoning do not consistently improve performance in tasks involving fuzzy logic. Additionally, our results show an inverse scaling effect in the performance of LLMs on FRoG. Interestingly, we also demonstrate that strong mathematical reasoning skills are not necessarily indicative of success on our benchmark.

(Nov 14): 13:00-14:00 (Afternoon) - Gather

Analyzing Dataset Annotation Quality Management in the Wild

Data quality is crucial for training accurate, unbiased, and trustworthy machine learning models and their correct evaluation. Recent works, however, have shown that even popular datasets used to train and evaluate state-of-the-art models contain a non-negligible amount of erroneous annotations, bias, or artifacts. There exist best practices and guidelines regarding dataset creation projects. Nevertheless, to the best of our knowledge, no large-scale analysis has been performed as of yet on how quality management is conducted when creating natural language datasets and whether these recommendations are followed. Therefore, we first survey and summarize recommended quality management practices for dataset creation as described in the literature and provide suggestions for applying them. Then, we compile a corpus of 591 scientific publications introducing text datasets and annotate it for quality-related aspects, such as annotator management, agreement, adjudication, or data validation. Using these annotations, we then analyze how quality management is conducted in practice. We find that a majority of the annotated publications apply good or very good quality management. However, we deem the effort of 30% of the works as only subpar. Our analysis also shows common errors, especially when using inter-annotator agreement and computing annotation error rates.

Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other Areas

(Nov 14): 13:00-14:00 (Afternoon) - Room: Gather

(Nov 14): 13:00-14:00 (Afternoon) - Gather

Should Cross-Lingual AMR Parsing go Meta? An Empirical Assessment of Meta-Learning and Joint Learning AMR Parsing

Jeongwoo Kang, Maximin Coavoux, Cédric Lopez, Didier Schwab

Cross-lingual AMR parsing is the task of predicting AMR graphs in a target language when training data is available only in a source language. Due to the small size of AMR training data and evaluation data, cross-lingual AMR parsing has only been explored in a small set of languages such as English, Spanish, German, Chinese, and Italian. Taking inspiration from Langedijk et al. (2022), who apply meta-learning to tackle cross-lingual syntactic parsing, we investigate the use of meta-learning for cross-lingual AMR parsing. We evaluate our models in k -shot scenarios (including 0-shot) and assess their effectiveness in Croatian, Farsi, Korean, Chinese, and French. Notably, Korean and Croatian test sets are developed as part of our work, based on the existing The Little Prince English AMR corpus, and made publicly available. We empirically study our method by comparing it to classical joint learning. Our findings suggest that while the meta-learning model performs

slightly better in 0-shot evaluation for certain languages, the performance gain is minimal or absent when k is higher than 0.

Sentiment Analysis, Stylistic Analysis, and Argument Mining

(Nov 14): 13:00 14:00 (Afternoon) - Room: Gather

(Nov 14): 13:00 14:00 (Afternoon) - Gather

D2R: Dual-Branch Dynamic Routing Network for Multimodal Sentiment Detection

Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, Fenghuan Li

Multimodal sentiment detection aims to classify the sentiment polarity of a given image-text pair. Existing approaches apply the same fixed framework to all input samples, lacking the flexibility to adapt to different image-text pairs. Furthermore, the interaction patterns of these methods are overly homogenized, limiting the model's capacity to extract multimodal sentiment information effectively. In this paper, we develop a Dual-Branch Dynamic Routing Network (D²R), which is the first multimodal dynamic interaction model towards multimodal sentiment detection. Specifically, we design six independent units to simulate inter- and intra-modal information interactions without depending on any existing fixed frameworks. Additionally, we configure a soft router in each unit to guide path generation and introduce the path regularization term to optimize these inference paths. Comprehensive experiments on three publicly available datasets demonstrate the superiority of our proposed model over state-of-the-art methods.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

External Knowledge-Driven Argument Mining: Leveraging Attention-Enhanced Multi-Network Models

Debela Gemechu, Chris Reed

Argument mining (AM) involves the identification of argument relations (AR) between Argumentative Discourse Units (ADUs). The essence of ARs among ADUs is context-dependent and lies in maintaining a coherent flow of ideas, often centered around the relations between discussed entities, topics, themes or concepts. However, these relations are not always explicitly stated; rather, inferred from implicit chains of reasoning connecting the concepts addressed in the ADUs. While humans can infer such background knowledge, machines face challenges where the contextual cues are not explicitly provided. This paper leverages external resources, including WordNet, ConceptNet, and Wikipedia to identify semantic paths (knowledge paths) connecting the concepts discussed in the ADUs to obtain the implicit chains of reasoning. To effectively leverage these paths for AR prediction, we propose attention-based Multi-Network architectures. Various architecture are evaluated on the external resources, and the Wikipedia based configuration attains F-scores of 0.85, 0.84, 0.70, and 0.87, respectively, on four diverse datasets, showing strong performance over the baselines.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

Are LLMs Good Zero-Shot Fallacy Classifiers?

Fengjun Pan, Xiaobao Wu, Zongrui Li, Anh Tuan Luu

Fallacies are defective arguments with faulty reasoning. Detecting and classifying them is a crucial NLP task to prevent misinformation, manipulative claims, and biased decisions. However, existing fallacy classifiers are limited by the requirement for sufficient labeled data for training, which hinders their out-of-distribution (OOD) generalization abilities. In this paper, we focus on leveraging Large Language Models (LLMs) for zero-shot fallacy classification. To elicit fallacy-related knowledge and reasoning abilities of LLMs, we propose diverse single-round and multi-round prompting schemes, applying different task-specific instructions such as extraction, summarization, and Chain-of-Thought reasoning. With comprehensive experiments on benchmark datasets, we suggest that LLMs could be potential zero-shot fallacy classifiers. In general, LLMs under single-round prompting schemes have achieved acceptable zero-shot performances compared to the best fullshot baselines and can outperform them in all OOD inference scenarios and some open-domain tasks. Our novel multi-round prompting schemes can effectively bring about more improvements, especially for small LLMs. Our analysis further underlines the future research on zero-shot fallacy classification. Codes and data are available at: https://github.com/panFJCharlotte98/Fallacy_Detection.

(Nov 14): 13:00 14:00 (Afternoon) - Gather

PFA-ERC Pseudo-Future Augmented Dynamic Emotion Recognition in Conversations

Tannay Khule, Rishabh Agrawal, Apurva Narayan

AI systems' ability to interpret human emotions and adapt to variations is becoming more crucial as AI gets embedded into everyone's daily lives. Emotion Recognition in Conversations (ERC) is based on this fundamental challenge. Current state-of-the-art technologies in ERC are limited due to the need for future information. We introduce High-Dimensional Temporal Fusion Transformer (HiTFT), a time-series forecasting transformer that predicts pseudo-future information to overcome this constraint. This retains the models' dynamic nature and provides future information more efficiently than other methods. Our proposed method combines pseudo future embeddings with an encoder that models the speaker's emotional state using past and pseudo-future information as well as inter and intra speaker interactions; these speaker states are then passed through a decoder block that predicts the inferred emotion of that utterance. We further evaluate our method and show that it achieves state of the art performance on three ERC datasets - MELD, EmoryNLP, and IEMOCAP.

Special Theme: Efficiency in Model Algorithms, Training, and Inference

(Nov 14): 13:00 14:00 (Afternoon) - Room: Gather

(Nov 14): 13:00 14:00 (Afternoon) - Gather

GRASS: Compute Efficient Low-Memory LLM Training with Structured Sparse Gradients

Aashiq Muhammed, Oscar Li, David Woodruff, Mona T. Diab, Virginia Smith

Large language model (LLM) training and finetuning are often bottlenecked by limited GPU memory. While existing projection-based optimization methods address this by projecting gradients into a lower-dimensional subspace to reduce optimizer state memory, they typically rely on *dense* projection matrices, which can introduce computational and memory overheads. In this work, we propose Grass (GRAident Stuctured Sparsification), a novel approach that leverages *sparse* projections to transform gradients into structured sparse updates. This design not only significantly reduces memory usage for optimizer states but also minimizes gradient memory footprint, computation, and communication costs, leading to substantial throughput improvements. Extensive experiments on pretraining and finetuning tasks demonstrate that Grass achieves comparable performance to full-rank training and existing projection-based methods. Notably, Grass enables half-precision pretraining of a 13B parameter LLaMA model on a single 40GB A100 GPU—a feat infeasible for previous methods—and yields up to a $2 \times$

throughput improvement on an 8-GPU system.

(Nov 14): 13:0014:00 (Afternoon) - Gather

Stochastic Fine-Tuning of Language Models Using Masked Gradients

Mohammad Akbar-Tajari, Mohammad Taher Pilehvar

Large Language Models (LLMs) have emerged as the dominant paradigm in Natural Language Processing owing to their remarkable performance across various target tasks. However, naively fine-tuning them for specific downstream tasks often requires updating a vast number of parameters, resulting in high computational costs and overfitting when training data is limited. In this paper, we propose a novel approach, called "Stochastic Tuning", that addresses these challenges by selectively updating a small subset of parameters in each step of the tuning process. Our approach is characterized by its customization of updates based on task-specific partial gradients with respect to stochastic sub-networks. The advantage of Stochastic Tuning over existing solutions lies in its ability to consider both parameter weights as well as forward values which guarantees a context-sensitive fine-tuning. Our experiments demonstrate that Stochastic Tuning outperforms existing lightweight fine-tuning methods, improving average performance by over two points on RoBERTa across several tasks in the GLUE benchmark while updating merely $\approx 0.08\%$ of the models parameters. The code for our implementation can be found at https://github.com/m-Tajari/StocTuning_LLMs.

(Nov 14): 13:0014:00 (Afternoon) - Gather

AutoPEFT: Automatic Configuration Search for Parameter-Efficient Fine-Tuning

Han Zhou, Xingchen Wan, Ivan Vuli, Anna Korhonen

Large pretrained language models are widely used in downstream NLP tasks via task-specific fine-tuning, but such procedures can be costly. Recently, Parameter-Efficient Fine-Tuning (PEFT) methods have achieved strong task performance while updating much fewer parameters than full model fine-tuning (FFT). However, it is non-trivial to make informed design choices on the PEFT configurations, such as their architecture, the number of tunable parameters, and even the layers in which the PEFT modules are inserted. Consequently, it is highly likely that the current, manually designed configurations are suboptimal in terms of their performance-efficiency trade-off. Inspired by advances in neural architecture search, we propose AutoPEFT for automatic PEFT configuration selection: we first design an expressive configuration search space with multiple representative PEFT modules as building blocks. Using multi-objective Bayesian optimisation in a low-cost setup, we then discover a Pareto-optimal set of configurations with strong performance-cost trade-offs across different numbers of parameters that are also highly transferable across different tasks. Empirically, on GLUE and SuperGLUE tasks, we show that AutoPEFT-discovered configurations significantly outperform existing PEFT methods and are on par or better than FFT without incurring substantial training efficiency costs.

(Nov 14): 13:0014:00 (Afternoon) - Gather

In Defense of Structural Sparse Adapters for Concurrent LLM Serving

Junda Su, Zirui Liu, Zetu Qiu, Weiyang Liu, Zhaozhuo Xu

Adapting large language models (LLMs) to specific tasks remains challenging due to the extensive retraining required, prompting the need for efficient adapter techniques. Despite this, the concurrent serving of multiple adapters, each with unique matrix shapes, poses significant system-level challenges. To address these issues, we identify an opportunity in structurally sparse adapters, which, unlike low-rank adapters, maintain consistent matrix shapes while varying in sparsity patterns. Leveraging this characteristic, we introduce SpartanServe, a system designed for efficient concurrent serving of LLMs using multiple structurally sparse adapters. SpartanServe employs a unified matrix multiplication operation and a novel memory management technique to enable effective batching. Furthermore, the incorporation of Triton kernels enhances the acceleration of matrix multiplication in the serving process. Experimental results demonstrate that SpartanServe achieves 2.12x speedup over S-LoRA when serving 96 adapters using a single NVIDIA A100 GPU (40GB), showcasing its efficacy in concurrent LLM serving.

Speech Processing and Spoken Language Understanding

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

Task Arithmetic can Mitigate Synthetic-to-Real Gap in Automatic Speech Recognition

Hsuan Su, Hua Fan, Fan-Yun Sun, Shang-Tse Chen, Hung-yi Lee

Synthetic data is widely used in speech recognition due to the availability of text-to-speech models, which facilitate adapting models to previously unseen text domains. However, existing methods suffer in performance when they fine-tune an automatic speech recognition (ASR) model on synthetic data as they suffer from the distributional shift commonly referred to as the synthetic-to-real gap. In this paper, we find that task arithmetic is effective at mitigating this gap. Our proposed method, *SYN2REAL* task vector, shows an average improvement of 10.03% improvement in word error rate over baselines on the SLURP dataset. Additionally, we show that an average of *SYN2REAL* task vectors, when we have real speeches from multiple different domains, can further adapt the original ASR model to perform better on the target text domain.

Syntax: Tagging, Chunking and Parsing

(Nov 14): 13:0014:00 (Afternoon) - Room: Gather

(Nov 14): 13:0014:00 (Afternoon) - Gather

A Fast and Sound Tagging Method for Discontinuous Named-Entity Recognition

Caio Filippo Corro

We introduce a novel tagging scheme for discontinuous named entity recognition based on an explicit description of the inner structure of discontinuous mentions. We rely on a weighted finite state automaton for both marginal and maximum a posteriori inference. As such, our method is sound in the sense that (1) well-formedness of predicted tag sequences is ensured via the automaton structure and (2) there is an unambiguous mapping between well-formed sequences of tags and (discontinuous) mentions. We evaluate our approach on three English datasets in the biomedical domain, and report comparable results to state-of-the-art while having a way simpler and faster model.

12

Tutorials: Friday, November 15, 2024

Overview

08:00 - 16:00

09:00 - 12:30

Registration

Morning tutorials

Countering Hateful and Offensive Speech Online - Open Challenges - Leon Derczynski, Marco Guerini, Debora Nozza, Flor Miriam Plaza-del-Arco, Jeffrey Sorensen and Marcos Zampieri

Monroe Ballroom
Terrace Level

12:30 - 14:00

14:00 - 17:30

Lunch Break

Afternoon tutorials

Reasoning with Natural Language Explanation - Marco Valentino and André Freitas

Brickell/Flagler
Ballrooms Terrace
Level

Language Agents: Foundations, Prospects, and Risks - Yu Su, Diyi Yang, Shunyu Yao and Tao Yu

Monroe Ballroom
Terrace Level

Brickell/Flagler
Ballrooms Terrace
Level

13

Tutorials: Saturday, November 16, 2024

Overview

08:00 - 16:00

Registration

09:00 - 12:30

Morning tutorial

AI for Science in the Era of Large Language Models - Zhenyu Bi, Minghao Xu, Jian Tang and Xuan Wang

Monroe Ballroom
Terrace Level

12:30 - 14:00

Lunch Break

14:00 - 17:30

Afternoon tutorial

Human-Centered Evaluation of Language Technologies - Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao and Ziang Xiao

Monroe Ballroom
Terrace Level

Please refer to the official program web page for last-minute updates: <https://2024.emnlp.org/program/tutorials/>

14

Tutorials Details

Tutorial Message

Welcome to the Tutorial Session of EMNLP 2024!

As the field of NLP continues to evolve, this year's tutorials at EMNLP 2024 will give the audience comprehensive introductions of six exciting topics by experts in these areas: natural language explanations, offensive speech, human-centered evaluation, AI for science, agents, and enhancing capabilities of LLMs.

As in recent years, the process of calling for, submitting, reviewing, and selecting tutorials was a collaborative effort across ACL, EACL, NAACL, and EMNLP. Each tutorial proposal was meticulously reviewed by a panel of three reviewers, who assessed them based on criteria such as clarity, preparedness, novelty, timeliness, instructors experience, potential audience, open access to teaching materials, and diversity (including multilingualism, gender, age, and geolocation). A total of six tutorials covering the aforementioned topics were selected for EMNLP.

We would like to thank the tutorial authors for their contributions, the tutorial chairs across conferences for this coordinated effort, as well as the EMNLP conference organizers, especially the general chair Thamar Solorio.

EMNLP 2024 Tutorial Co-chairs
Junyi Jessy Li
Fei Liu

T1 - Countering Hateful and Offensive Speech Online

- Open Challenges



Flor Miriam Plaza-del-Arco, Debora Nozza, Marco Guerini, Jeffrey Sorensen and Marcos Zampieri

<https://nlp-for-countering-hate-speech-tutorial.github.io/>

Room: "Monroe Ballroom - Terrace Level"
Friday, November 15, 2024 - from 09:00 to 12:30
Introductory

In today's digital age, hate speech and offensive speech online pose a significant challenge to maintaining respectful and inclusive online environments.

This tutorial aims to provide attendees with a comprehensive understanding of the field by delving into essential dimensions such as multilingualism, counter-narrative generation, a hands-on session with one of the most popular APIs for detecting hate speech, fairness, and ethics in AI, and the use of recent advanced approaches. In addition, the tutorial aims to foster collaboration and inspire participants to create safer online spaces by detecting and mitigating hate speech.

Time	Section	Presenter
9:00 - 9:10	Section 1: Introduction	Debora Nozza
9:10 - 9:45	Section 2: Multilingualism	Marcos Zampieri
9:45 - 10:20	Section 3: Counter-narrative Generation	Marco Guerini
10:20 - 10:30	Q&A	
10:30 - 11:00	Coffee Break	
11:00 - 11:25	Section 4: Hands-on Session (Perspective API)	Jeffrey Sorensen
11:25 - 11:55	Section 5: Fairness & Ethics	Debora Nozza
11:55 - 12:20	Section 6: How to use recent LLMs?	Flor Miriam Plaza-del-Arco
12:20 - 12:30	Q&A and Discussion	

Flor Miriam Plaza-del-Arco, Bocconi University, Italy

email: flor.plaza@unibocconi.it

website: <https://fmp Plaza.github.io/>

Flor Miriam Plaza-del-Arco is a Postdoctoral Research Fellow in Dirk Hovy's MilaNLP lab at Bocconi University. Her work centers on the intersection of language, computation, and society, with a focus on emotions and harmful language within NLP for social good. She earned her Ph.D. with highest honors (summa cum laude) in January 2023 from the SINAI lab at the University of Jaén, where she contributed to advancements in hate speech detection and emotion identification in Spanish. Active in the field, she has co-organized the 8th Workshop on Online Abuse and Harms (NAACL 2024), several IberLEF shared tasks, and SEPLN conferences. She is currently co-organizing a tutorial on countering online hate for EMNLP 2024.

Debora Nozza, Bocconi University, Italy

email: debora.nozza@unibocconi.it

website: <https://www.deboranozza.com/>

Debora Nozza is an Assistant Professor in Computing Sciences at Bocconi University and a recipient of a 1.5 million ERC Starting Grant (2023) for her research on personalized and subjective approaches to Natural Language Processing. She previously secured a 120,000 grant from Fondazione Cariplo for her project MONICA, which monitors coverage, attitudes, and accessibility of Italian COVID-19 response measures. Her research focuses on NLP, especially in detecting and countering hate speech and algorithmic bias in multilingual social media data. She has organized the 7th Workshop on Online Abuse and Harms (ACL 2023), the ICWSM 2023 Data Challenge, and tasks on misogyny identification and multilingual hate speech detection at Evalita and SemEval.

Marco Guerini, FBK, Italy

email: info@marcoguerini.eu

website: <https://www.marcoguerini.eu/>

Marco Guerini is a researcher in Computational Linguistics and leads the Language and Dialogue Technologies group at Fondazione Bruno Kessler (FBK). His work focuses on persuasive communication, sentiment analysis, and social media, with a recent emphasis on AI technologies for counter-narrative generation to combat online hate speech. He holds a Ph.D. in Information and Communication Technologies from the University of Trento and has published extensively in top conferences and international journals. His research has gained international media attention, including features in the Wall Street Journal, MIT Technology Review, and Harvard Business Review. He has worked with FBKs NLP group, Trento-Rise, and received a Google Research Award (2011) and a sponsorship from eBay (2016). Additionally, he consults for startups and companies and writes a technology blog for Corriere della Sera.

Jeffrey Sorenson, Jigsaw, USA

email: sorenj@google.com

website: <https://research.google/people/author14753/>

Jeffrey Sorenson is a researcher at Jigsaw, a unit within Google that explores threats to open societies, and builds technology that inspires scalable solutions. His work spans the development of scalable algorithms for threat detection, countering online abuse, and fostering resilience against cyber threats. An experienced contributor to the fields of computational social science and ethical AI, Jeffrey has presented his research at leading conferences, showcasing solutions that help safeguard online communities and enhance digital well-being on a global scale.

Marcos Zampieri, George Mason University, USA

email: mzampier@gmu.edu

website: <https://www.gmu.edu/profiles/mzampier>

Marcos Zampieri is an assistant professor in the Department of Information Sciences and Technology, School of Computing, at George Mason University. He received his PhD from Saarland University where he was affiliated with the German Research Center in Artificial Intelligence (DFKI). His research interests are in computational linguistics and natural language processing (NLP). His research deals with the collection and processing of large bodies of texts from various sources (e.g. social media, newspapers) with the goal of training robust NLP systems. He has published over 100 peer-reviewed papers in journals and conference proceedings and co-edited a dozen edited volumes, special issues, and workshop proceedings.

T2 - Enhancing LLM Capabilities Beyond Scaling Up



Wenpeng Yin, Muhao Chen, Rui Zhang, Ben Zhou, Fei Wang and Dan Roth

<https://www.wenpengyin.org/publications/beyond-llm-scaling-emnlp24>

Room: "Brickell/Flagler Ballrooms - Terrace Level"

Friday, November 15, 2024 - from 09:00 to 12:30

Cutting-edge

General-purpose large language models (LLMs) are progressively expanding both in scale and access to unpublic training data. This has led to notable progress in a variety of AI problems. Nevertheless, two questions exist: i) Is scaling up the sole avenue of extending the capabilities of LLMs? ii) Instead of developing general-purpose LLMs, how to endow LLMs with specific knowledge? This tutorial targets researchers and practitioners who are interested in capability extension of LLMs that go beyond scaling up. To this end, we will discuss several lines of research that follow that direction, including: (i) optimizing input prompts to fully exploit LLM potential, (ii) enabling LLMs to self-improve responses through various feedback signals, (iii) updating or editing the internal knowledge of LLMs when necessary, (iv) leveraging incidental structural supervision from target tasks, and (v) defending against potential attacks and threats from malicious users. At last, we will conclude the tutorial by outlining directions for further investigation.

Introduction - Dan Roth

- Progress of current LLMs
- Scaling trend of LLMs
- From general-purpose LLMs to domain-specific LLMs

Prompt Optimization for LLMs - Rui Zhang

- Search-based prompt optimization
- Text gradientbased prompt optimization
- Gradient-based prompt optimization

LLM Self-improvement & LLM-LLM Collaboration - Wenpeng Yin

- LLM self-improvement (w/ self-feedback, self-discriminative abilities)
- LLM-LLM collaboration
- LLM-LLM merging

Knowledge Update of LLMs - Fei Wang

- Examine the issues caused by unreliable knowledge, such as hallucinations
- Remedy LLMs internal knowledge by integrating external information in a training-free manner
- LLM knowledge editing with lightweight tuning

Aligning with Structures of Target Problems - Ben Zhou

- Symbolic constraints as structures (e.g., human-written, mathematical constraints, and compiler constraints)
- Structures from decomposing the target problem
- Procedural structures that come from cognitive or problem-solving processes, such as DSP, ReAct, and RAP.

Safety Enhancement for LLMs - Muhaoo Chen

- Introducing inference-time threats (e.g., prompt injection, malicious task instructions, jailbreaking attacks, adversarial demonstrations, and training-free backdoor attacks)
- Defense techniques (e.g., prompt robustness estimation, demonstration-based defense, and ensemble debiasing)

Conclusion & Future Directions - Dan Roth

Wenpeng Yin, Penn State University, USA

email: wenpeng@psu.edu

website: <https://www.wenpengyin.org/>

He is a tenure-track Assistant Professor in the Department of Computer Science and Engineering at Penn State, an affiliated faculty member of the Center for Socially Responsible Artificial Intelligence, and an Associate at the Institute for Computational and Data Sciences. He leads the AI4Research lab, which focuses on advancing responsible and innovative AI applications across disciplines.

Muhaao Chen, University of California, Davis, USA

email: muuchen@ucdavis.edu

website: <https://muhaochen.github.io/>

Muhaao Chen is an Assistant Professor in the Department of Computer Science at UC Davis and leads the Language Understanding and Knowledge Acquisition (LUKA) Lab. His research focuses on robust, minimally supervised machine learning for natural language processing, with recent work addressing accountability and security issues in large language and multi-modal models. Previously, he served as an Assistant Research Professor at USC (2020/2023) and as a Postdoctoral Fellow with Dan Roth at UPenn. He earned his Ph.D. in Computer Science at UCLA in 2019, following his bachelors from Fudan University in 2014.

Rui Zhang, Penn State University, USA

email: rmz5227@psu.edu

website: <http://ryanzhumich.github.io>

Rui Zhang is an Assistant Professor in the Computer Science and Engineering Department of Penn State University. He is a co-director of the PSU Natural Language Processing Lab. His research interest lies in Trustworthy Human-Centered AI, LLM Agents, and AI for Science. He received an NSF CAREER Award, a Microsoft Research Award, an Amazon Research Award, an eBay Research Award, and a Cisco Research Award. He received B.S. degrees from both Shanghai Jiao Tong University and the University of Michigan in 2015 and received his Ph.D. from the Computer Science Department at Yale University in 2020. He has done industry research internships at IBM Thomas J. Watson Research Center, Grammarly Research, and Google AI.

Ben Zhou, Arizona State University, USA

email: benzhou@asu.edu

website: <http://xuanyu.me/>

He is an assistant professor at Arizona State University, where he directs the ARC Lab. His primary research interest is in controllable and trustworthy NLP/AI, often leveraging neural-symbolic approaches that apply to complex reasoning tasks. His methods are inspired by human cognitive processes, such as analogy and experiential knowledge, aiming for generalizable solutions.

Fei Wang, University of Southern California, USA

email: fwang1412@gmail.com

website: <https://feiwang96.github.io/>

He is a Ph.D. candidate in computer science at the University of Southern California, co-advised by Muhaao Chen and Aram Galstyan, with close collaboration with Kai-Wei Chang at UCLA and Dan Roth at UPenn. His research, supported by the Amazon ML and Annenberg PhD Fellowships, focuses on NLP and ML, specifically on developing robust, controllable multimodal LLMs to improve reliability and generalizability in alignment and inference. He has interned at Google Cloud AI, AWS AI Labs, Amazon Alexa AI, and Tencent AI Lab (Seattle).

Dan Roth, University of Pennsylvania & Oracle, USA

email: danroth@seas.upenn.edu

website: <https://www.cis.upenn.edu/~danroth/>

His research centers on the computational foundations of intelligent behavior, particularly in machine learning and inference for natural language understanding. He has pioneered constrained conditional models, using integer linear programming to enhance learning and inference for NLP. His work has yielded leading systems in semantic role labeling, co-reference resolution, and textual entailment, with a recent emphasis on incidental supervision to handle complex problems. Additionally, he has advanced declarative learning languages, including LBJava and Saul, designed for rapid development in software with learned components.

T3 - Reasoning with Natural Language Explanation



Marco Valentino and André Freitas

<https://sites.google.com/view/reasoning-with-explanations>

Room: "Monroe Ballroom - Terrace Level"
Friday, November 15, 2024 - from 14:00 to 17:30
Introductory

Explanation constitutes an archetypal feature of human rationality, underpinning learning and generalisation, and representing one of the media supporting scientific discovery and communication.

Due to the importance of explanations in human reasoning, an increasing amount of research in Natural Language Inference (NLI) has started reconsidering the role that explanations play in learning and inference, attempting to build explanation-based NLI models that can effectively encode and use natural language explanations on downstream tasks.

Research in explanation-based NLI, however, presents specific challenges and opportunities, as explanatory reasoning reflects aspects of both material and formal inference, making it a particularly rich setting to model and deliver complex reasoning.

In this tutorial, we provide a comprehensive introduction to the field of explanation-based NLI, grounding this discussion on the epistemological-linguistic foundations of explanations, systematically describing the main architectural trends and evaluation methodologies that can be used to build systems capable of explanatory reasoning.

Tutorial paper: <https://arxiv.org/abs/2410.04148>

Marco Valentino, Idiap Research Institute, Swiss
 email: marco.valentino@idiap.ch

website: <https://www.marcovalentino.net/>

Marco is a postdoc at the Idiap Research Institute. His research activity lies at the intersection of natural language processing, reasoning, and explanation, investigating the development of AI systems that can support explanatory natural language reasoning in complex domains (e.g., mathematics, science, biomedical and clinical applications).

André Freitas, Idiap Research Institute, Swiss & University of Manchester, UK
 email: andre.freitas@manchester.ac.uk

website: <https://www.andrefreitas.net/>

André leads the Neuro-symbolic AI Lab at the University of Manchester and Idiap Research Institute. His main research interests are on enabling the development of AI methods to support abstract, flexible and controlled reasoning in order to support AI-augmented scientific discovery.

T4 - Language Agents: Foundations, Prospects, and Risks



Yu Su, Diyi Yang, Shunyu Yao and Tao Yu

<https://language-agent-tutorial.github.io/>

Room: "Brickell/Flagler Ballrooms - Terrace Level"

Friday, November 15, 2024 - from 14:00 to 17:30

Cutting-edge

A key topic in AI and NLP today is autonomous agents, typically powered by large language models (LLMs), which can follow language instructions to complete complex tasks in real or simulated environments. While the concept of AI agents has existed since the beginning of AI, the recent shift lies in their ability to use language as a vehicle for both thought and communication, a trait previously unique to humans. This expands their ability to tackle diverse problems autonomously and sets them apart from earlier AI agents. To reflect this, the term "language agents" has been suggested, emphasizing language as their defining feature. Language played a crucial role in human cognitive evolution, and AI might be following a similar path. However, there has been little systematic discussion on the definition, theoretical foundations, risks, and future directions of language agents. This tutorial aims to fill that gap by providing a comprehensive exploration of language agents, linking insights from both modern and classic AI research.

Yu Su, The Ohio State University, USA

email: su.809@osu.edu

website: <https://ysu1989.github.io/>

He is a Distinguished Assistant Professor of Engineering Inclusive Excellence in Computer Science and Engineering at The Ohio State University, where he co-directs the OSU NLP group, co-leads the Foundational AI team in the ICICLE AI Institute, and leads the Machine Learning Foundations team in the Imageomics Institute. His research focuses on artificial intelligence, particularly foundation models like large language models (LLMs) and multimodal models, emphasizing generalizability, interpretability, efficiency, and robustness for real-world applications.

Diyi Yang, Stanford University, USA

email: diiyi@cs.stanford.edu

website: <https://cs.stanford.edu/~diyiy/>

She is an assistant professor in the Computer Science Department at Stanford, affiliated with the Stanford NLP Group, Stanford HCI Group, Stanford AI Lab (SAIL), and Stanford Human-Centered Artificial Intelligence (HAI). Her research focuses on Socially Aware Natural Language Processing, aiming to deepen the understanding of human communication within social contexts. She seeks to develop socially aware language technologies that enhance both human-human and human-computer interaction.

Shunyu Yao, Princeton University, Open-AI, USA

email: shunuy@princeton.edu

website: <https://ysymyth.github.io/>

He is a researcher at OpenAI. He studies agents.

Tao Yu, Penn State University, USA

email: tao.yu.nlp@gmail.com

website: <https://taoyds.github.io/>

Tao Yu is an Assistant Professor of Computer Science at The University of Hong Kong and directs the XLANG Lab within the HKU NLP Group. His research focuses on Natural Language Processing, aiming to develop language model agents that translate language instructions into executable actions across databases, web applications, and real-world environments. This work seeks to advance natural language interfaces that interact with and learn from their environments, enhancing human engagement with data analysis, web tools, and robotics through conversational interfaces.

T5 - AI for Science in the Era of Large Language Models



Zhenyu Bi, Minghao Xu, Jian Tang and Xuan Wang

<https://xuanwang91.github.io/2024-11-12-emnlp24-tutorial>

Room: "Monroe Ballroom - Terrace Level"
Saturday, November 16, 2024 - from 09:00 to 12:30
Cutting-edge

The capabilities of AI in the realm of science span a wide spectrum, from the atomic level, where it solves partial differential equations for quantum systems, to the molecular level, predicting chemical or protein structures, and even extending to societal predictions like infectious disease outbreaks. Recent advancements in large language models (LLMs), exemplified by models like ChatGPT, have showcased significant prowess in tasks involving natural language, such as translating languages, constructing chatbots, and answering questions. When we consider scientific data, we notice a resemblance to natural language in terms of sequences: scientific literature and health records presented as text, bio-omics data arranged in sequences, or sensor data like brain signals. The question arises: Can we harness the potential of these recent LLMs to drive scientific progress? In this tutorial, we will explore the application of large language models to three crucial categories of scientific data: 1) textual data, 2) biomedical sequences, and 3) brain signals. Furthermore, we will delve into LLMs challenges in scientific research, including ensuring trustworthiness, achieving personalization, and adapting to multi-modal data representation.

Zhenyu Bi, Virginia Polytechnic Institute and State University, USA
 email: zhenyub@vt.edu

website: <https://bzzyzz.github.io/>
 Zhenyu Bi is a Ph.D. student in the Computer Science Department at Virginia Tech. His research area lies in the field of natural language processing, emphasizing real-world applications of Large Language Models. He is mainly interested in information extraction with weak supervision, especially text mining and event extraction; as well as fact-checking and trustworthy NLP. He received an M.S. degree in Intelligent Information Systems from Carnegie Mellon University in 2023, a B.S. degree in Cognitive Science, and a B.S. Degree in Computer Science from the University of California, San Diego in 2021.

Minghao Xu, Quebec AI Institute, Canada
 email: minghao.xu@umontreal.ca

website: <https://chrisallenming.github.io/>
 Minghao Xu is a Ph.D. student at Mila - Quebec AI Institute, Canada. His research interests mainly lie in protein function understanding and protein design. He aims to understand diverse protein functions with joint guidance from protein sequences, structures, and biomedical text, especially boosted by large-scale multi-modal pre-training. He is also pursuing structure- and sequence-based protein design via generative AI, geometric deep learning and dry-wet experiment closed looping. He has given an Oral presentation at the main conference of ICML23.

Jian Tang, Montreal Institute for Learning, Canada
 email: jian.tang@hec.ca

website: <https://jian-tang.com/>
 Jian Tang is an Associate Professor at Mila - Quebec AI Institute, Canada. His long-term interests focus on understanding the language of life (DNA, RNAs, and Proteins) with generative AI and geometric deep learning, with applications in biomedicine and synthetic biology. His group has developed one of the first open-source machine learning frameworks on drug discovery, TorchDrug (for small molecules) and TorchProtein (for proteins), and developed the first diffusion models for 3D molecular structure generation, GeoDiff (among the 50 most cited AI paper in 2022). He has given a few tutorials at international AI and data mining conferences including KDD 2017, AAAI 2019, AAAI 2022.

Xuan Wang, Virginia Tech University, USA
 email: xuanw@vt.edu

website: <https://xuanwang91.github.io/>
 Xuan Xuan Wang is an Assistant Professor in the Computer Science Department at Virginia Tech. Her research focuses on natural language processing and text mining, emphasizing applications to science and healthcare domains. Her current projects include NLP and text mining with extremely weak supervision; text-augmented knowledge graph reasoning; fact-checking and trustworthy NLP, AI for science; and AI for healthcare. She received a Ph.D. degree in Computer Science, an M.S. degree in Statistics, and an M.S. degree in Biochemistry from the University of Illinois Urbana-Champaign in 2022, 2017, and 2015, respectively, and a B.S. degree in Biological Science from Tsinghua University in 2013. She has delivered tutorials in IEEE-BigData 2019, WWW 2022, and KDD 2022.

T6 - Human-Centered Evaluation of Language Technologies



Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao and Ziang Xiao

<https://human-centered-eval.github.io/>

Room: "Monroe Ballroom - Terrace Level"

Saturday, November 16, 2024 - from 14:00 to 17:30

Introductory

Evaluation is a cornerstone topic in NLP. However, many criticisms have been raised about the community's evaluation practices, including a lack of human-centered considerations about people's needs for language technologies and technologies' actual impact on people. This evaluation crisis is exacerbated by the recent development of large generative models with diverse and uncertain capabilities. This tutorial aims to inspire more human-centered evaluation in NLP by introducing perspectives and methodologies from the social sciences and human-computer interaction (HCI), a field concerned primarily with the design and evaluation of technologies. The tutorial will start with an overview of current NLP evaluation practices and their limitations, then introduce complementary perspectives from the social sciences and a toolbox of evaluation methods from HCI, accompanied by discussions of considerations such as what to evaluate for, how generalizable the results are to the real-world contexts, and pragmatic costs of conducting the evaluation. The tutorial will also encourage reflection on how these HCI perspectives and methodologies can complement NLP evaluation through Q A discussions and a hands-on exercise.

Su Lin Blodgett, FATE, Microsoft Research Montréal, Canada

email: sulin.blodgett@microsoft.com

website: <https://sblodgett.github.io/>

She is a senior researcher in the Fairness, Accountability, Transparency, and Ethics in AI (FATE) group at Microsoft Research Montréal, focusing on the social and ethical implications of natural language processing (NLP) technologies. Her work includes developing methods to anticipate, measure, and mitigate harms associated with language technologies, emphasizing languages social complexities and supporting NLP practitioners in ethical practice. Previously a postdoctoral researcher at MSR Montréal, she earned her Ph.D. in computer science at the University of Massachusetts Amherst under the guidance of Brendan OConnor, supported by the NSF Graduate Research Fellowship, and holds a B.A. in mathematics from Wellesley College.

Jackie Chi Kit Cheung, McGill University, Canada

email: jackie.cheung@mcgill.ca

website: <https://www.cs.mcgill.ca/~jcheung/>

Jackie Chi Kit Cheung is an Associate Professor in Computer Science at McGill University, where he holds a Canada CIFAR AI Chair with Mila - Quebec AI Institute. His research focuses on natural language processing (NLP), particularly in evaluation, natural language generation, automatic summarization, computational semantics, and commonsense reasoning. His lab aims to develop NLP models that support complex tasks, incorporating insights from linguistics and psychology to create systems for various applications, including education and health.

Q. Vera Liao, University of Michigan CSE, USA

email: veraalao@microsoft.com

website: <https://qveraliao.com/>

She is a Principal Researcher at Microsoft Research, actively involved in the FATE (Fairness, Accountability, Transparency, and Ethics of AI) group. Her research primarily examines and aims to mitigate the risks associated with emerging technologies, with a recent focus on the transparency of AI systems, including aspects like explainability, evaluation, and uncertainty communication, and how these intersect with human experiences such as trust and control. Before joining Microsoft, she worked at IBM's T.J. Watson Research Center, contributing to key products like AI Explainability 360 and Watson Assistant, and she studied at the University of Illinois at Urbana-Champaign and Tsinghua University.

Ziang Xiao, Johns Hopkins University, USA

email: ziang.xiao@jhu.edu

website: <https://www.ziangxiao.com/>

He is an Assistant Professor in Computer Science at Johns Hopkins University, having completed his Ph.D. in Computer Science at the University of Illinois Urbana-Champaign, where he was co-advised by Professors Hari Sundaram and Karrie Karahalios. He earned his B.S. in Psychology and Statistics & Computer Science at the University of Illinois under the guidance of Professor Dov Cohen. His research focuses on understanding human behavior at scale through human-computer interaction, exploring topics such as AI for social science, human-centered model evaluation, and information seeking, while integrating insights from natural language processing and social and personality psychology.

A large, stylized number '15' is positioned in the upper right corner of the slide.

Workshops

Overview

Friday, November 15, 2024

Jasmine	W1 - BlackboxNLP 2024: Analyzing and interpreting neural networks for NLP	p.423
Merrick 1	W2 - Seventh Workshop on Computational Models of Reference, Anaphora and Coreference	p.424
Miami Lecture Hall	W3 - Seventh Workshop on Fact Extraction and VERification (FEVER)	p.426
Johnson	W4 - Workshop on the Future of Event Detection	p.430
Pearson	W5 - The Sixth Workshop on Narrative Understanding	p.431
Foster	W6 - Third Workshop on NLP for Positive Impact	p.432
Merrick 2	W7 - The Third Workshop on Text Simplification, Accessibility and Readability	p.433
Hibiscus	W8 - The Eighth Widening NLP Workshop (WiNLP 2024)	p.435

Friday, November 15 and Saturday, November 16, 2024

Tuttle	W9 - The SIGNLL Conference on Computational Natural Language Learning (CoNLL)	p.436
Ashe Auditorium	W10 - Ninth Conference on Machine Translation (WMT24)	p.439

Saturday, November 16, 2024

Merrick 1	W11 - Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)	p.449
Merrick 2	W12 - The 4th International Workshop on Natural Language Processing for Digital Humanities (NLP4DH)	p.451
Brickell	W13 - GenBench: The second workshop on generalisation (benchmarking) in NLP	p.455
Hibiscus A	W14 - Natural Legal Language Processing (NLLP) Workshop 2024	p.457
Jasmine	W15 - The 4th Workshop on Multilingual Representation Learning	p.460
Hibiscus B	W16 - NLP4Science: The First Workshop on Natural Language Processing for Science	p.461
Pearson	W17 - The Second Workshop on Social Influence in Conversations (SICon 2024)	p.462
Miami Lecture Hall	W18 - The 11th Workshop on Asian Translation (WAT2024)	p.463
Johnson	W19 - The First Workshop on Advancing Natural Language Processing for Wikipedia (NLP for Wikipedia)	p.465

W1 - BlackboxNLP 2024: Analyzing and interpreting neural networks for NLP

Organizers:

Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller,
Hanjie Chen

<https://blackboxnlp.github.io/>

Room: Jasmine

Friday, November 15, 2024

Many recent performance improvements in NLP have come at the cost of understanding of the systems. How do we assess what representations and computations models learn? How do we formalize desirable properties of interpretable models, and measure the extent to which existing models achieve them? How can we build models that better encode these properties? What can new or existing tools tell us about these systems inductive biases? The goal of this workshop is to bring together researchers focused on interpreting and explaining NLP models by taking inspiration from fields such as machine learning, psychology, linguistics, and neuroscience. We hope the workshop will serve as an interdisciplinary meetup that allows for cross-collaboration.

Time	Session
09:00 09:10	Opening remarks
09:10 10:00	Invited talk by Jack Merullo
10:00 10:30	Oral presentations: <ul style="list-style-type: none">• Routing in Sparsely-gated Language Models responds to Context. <i>Stefan Arnold, Marian Fietta, Dilara Yesilbas</i>• Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models. <i>Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, Anna A Ivanova</i>
10:30 11:00	Break
11:00 12:30	In-person & virtual poster session 1
12:30 14:00	Lunch
14:00 15:00	Invited talk by Himabindu Lakkaraju
15:00 15:30	Oral presentations: <ul style="list-style-type: none">• Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. <i>Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, Neel Nanda</i>• Mechanistic? <i>Naomi Saphra, Sarah Wiegreffe</i>
15:30 16:00	Break
16:30 16:40	Closing remarks and awards
17:00 18:00	Panel discussion on Interpretability with: Dieuwke Hupkes, Vera Liao, Asma Ghandeharioun, Marius Mosbach and Jack Merullo

W2 - Seventh Workshop on Computational Models of Reference, Anaphora and Coreference

Organizers:

Maciej Ograniczuk, Sameer Pradhan, Anna Nedoluzhko, Massimo Poesio,
Vincent Ng

<https://sites.google.com/view/crac2024/>

Room: Merrick 1
Friday, November 15, 2024

Since 2016, the yearly CRAC (and its predecessor, CORBON) workshop has become the primary forum for researchers interested in the computational modeling of reference, anaphora, and coreference to discuss and publish their results. Over the years, this workshop series has successfully organized five shared tasks, which stimulated interest in new problems in this area of research, facilitated the discussion and dissemination of results on new problems/directions (e.g., multimodal reference resolution), and helped expand the coreference community that used to be dominated by European researchers to include young researchers from the Americas. The aim of the workshop is to provide a forum where work on all aspects of computational work on anaphora resolution and annotation can be presented.

Time	Session
09:00 09:15	Opening and welcome (<i>Vincent Ng and Maciej Ograniczuk</i>)
09:15 10:30	Invited talk: Reference at the Heart of Natural Language Processing. <i>Jackie Chi Kit Cheung</i>
	Findings Paper Session
11:00 11:20	<ul style="list-style-type: none">• Challenges to Evaluating the Generalization of Coreference Resolution Models: A Measurement Modeling Perspective. <i>Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung</i>
11:20 11:40	<ul style="list-style-type: none">• Any Other Thoughts, Hedgehog? Linking Deliberation Chains in Collaborative Dialogues. <i>Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Collin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy</i>
11:40 11:50	<ul style="list-style-type: none">• MMAR: Multilingual and Multimodal Anaphora Resolution in Instructional Videos. <i>Cennet Oguz, Pascal Denis, Simon Ostermann, Emmanuel Vincent, Natalia Skachkova, and Josep van Genabith</i>
	EMNLP 2024 Paper
11:50 - 12:10	<ul style="list-style-type: none">• Major Entity Identification: A Generalizable Alternative to Coreference Resolution. <i>Kawshik S. Manikantan, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi</i>
	Research Paper Session
13:50 14:00	<ul style="list-style-type: none">• Enriching Conceptual Knowledge in Language Models through Metaphorical Reference Explanation. <i>Zixuan Zhang and Heng Ji</i>
14:00 14:10	<ul style="list-style-type: none">• Polish Coreference Corpus as an LLM Testbed: Evaluating Coreference Resolution within Instruction-Following Language Models by Instruction-Answer Alignment. <i>Karol Saputa, Angelika Peljak-apiska, and Maciej Ograniczuk</i>

14:10	14:30	• MSCAW-coref: Multilingual, Singleton and Conjunction-Aware Word-Level Coreference Resolution. <i>Houjun Liu, John Bauer, Karel D’Oosterlinck, Christopher Potts, and Christopher D. Manning</i>
14:30	14:50	• Unifying the Scope of Bridging Anaphora Types in English: Bridging Annotations in ARRAU and GUM. <i>Lauren Levine and Amir Zeldes</i>
14:50	15:10	• Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case. <i>Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow</i>
15:10	15:30	• DeepHCoref: A Deep Neural Coreference Resolution for Hindi Text. <i>Kusum Lata, Kamlesh Dutta, Pardeep Singh, and Abhishek Kanwar</i>
Shared Task Paper Session		
16:00	16:30	• Findings of the Third Shared Task on Multilingual Coreference Resolution. <i>Michal Novák, Barbora Dohnalová, Miloslav Konopík, Anna Nedoluzhko, Martin Popel, Ondej Praák, Jakub Sido, Milan Straka, Zdenk abokrtský, and Daniel Zeman</i>
16:30	16:50	• CorPipe at CRAC 2024: Predicting Zero Mentions from Raw Text. <i>Milan Straka</i>
16:50	17:10	• End-to-end Multilingual Coreference Resolution with Headword Mention Representation. <i>Ondej Praák and Miloslav Konopík</i>
17:10	17:20	• Multilingual coreference resolution as text generation. <i>Natalia Skachkova</i>
Panel Discussion		
17:20	17:50	• The future of coreference resolution in the era of LLMs. <i>Michal Novák, Ondej Praák, and Martin Popel</i>
17:50	18:00	Closing of the workshop (<i>Maciej Ogrodniczuk</i>)

W3 - Seventh Workshop on Fact Extraction and VERification (FEVER)

Organizers:

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng,
Mubashara Akhtar, Rami Aly, Rui Cao, Zhijiang Guo, Christos Christodoulouopoulos,
Oana Cocarascu, Arpit Mittal, James Thorne, Andreas Vlachos

<https://fever.ai/workshop.html>

Room: Miami Lecture Hall
Friday, November 15, 2024

With billions of individual pages on the web providing information on almost every conceivable topic, we should have the ability to reason about a wide range of domains. However, in order to do so, we need to ensure that we trust the accuracy of the sources of information that we use. Handling false information coming from unreliable sources has become the focus of a lot of recent research and media coverage. In an effort to jointly address these problems, we are organizing the 7th instalment of the Fact Extraction and VERification (FEVER) workshop (<http://fever.ai/>) to promote research in this area.

In this years workshop, we are also organising a new fact checking shared task AVeriTec: Automated Verification of Textual Claims. The aim is to fact-check real-world claims using evidence from the web. For each claim, systems must return a label (Supported, Refuted, Not Enough Evidence, Conflicting Evidence/Cherry-picking) and appropriate evidence. The evidence must be retrieved from the document collection provided by the organisers or from the Web (e.g. using a search API).

Time	Session
9:00-9:45	Opening Remarks & Shared Task Overview <i>FEVER Organizers</i>
9:45-10:30	Keynote Talk 1: Omar Khattab
10:30-11:00	Coffee break
11:00-12:00	Poster Session <ul style="list-style-type: none">• Multi-hop Evidence Pursuit Meets the Web: Team Papelo at FEVER 2024. <i>Christopher Malon</i>• Retrieving Semantics for Fact-Checking: A Comparative Approach using CQ (Claim to Question) & AQ (Answer to Question). <i>Nicolò Urbani, Sandip Modha and Gabriella Pasi</i>• RAG-Fusion Based Information Retrieval for Fact-Checking. <i>Yuki Momii, Tetsuya Takiguchi and Yasuo Ariki</i>• UHH at AVeriTec: RAG for Fact-Checking with Real-World Claims. <i>Ozge Sevgili, Irina Nikishina, Seid Muhie Yimam, Martin Semmann and Chris Bieman</i>• Improving Evidence Retrieval on Claim Verification Pipeline through Question Enrichment. <i>Svetlana Churina, Anab Maulana Barik and Saisamarth Rajesh Phaye</i>• Dunamu-mls Submissions on AVERITEC Shared Task. <i>Heesoo Park, Dongjun Lee, Jaehyuk Kim, ChoongWon Park and Changwha Park</i>• FZI-WIM at AVeriTec Shared Task: Real-World Fact-Checking with Question Answering. <i>Jin Liu, Steffen Thoma and Achim Rettinger</i>

-
- Zero-Shot Learning and Key Points Are All You Need for Automated Fact-Checking. *Mohammad Ghiasvand Mohammadkhani, Ali Ghiasvand Mohammadkhani and Hamid Beigy*
 - Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. *Ronit Singal, Pranshu Patwa, Parth Patwa, Aman Chadha and Amitava Das*
 - SKäDU Team: Cross-Encoder based Evidence Retrieval and Question Generation with Improved Prompt for the AVeriTeC Shared Task. *Shrikant Malviya and Stamos Katsigiannis*
 - InFact: A Strong Baseline for Automated Fact-Checking. *Mark Rothermel, Tobias Braun, Marcus Rohrbach and Anna Rohrbach*
 - Exploring Retrieval Augmented Generation For Real-world Claim Verification. *Adjali Omar*
 - GProofT: A Multi-dimension Multi-round Fact Checking Framework Based on Claim Fact Extraction. *Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang and Yangqiu Song*
 - HerO at AVeriTeC: The Herd of Open Large Language Models for Verifying Real-World Claims. *Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon and Kunwoo Park*
 - AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. *Herbert Ullrich, Tomás Mlynář and Jan Drchal*
 - Enhancing Fact Verification with Causal Knowledge Graphs and Transformer-Based Retrieval for Deductive Reasoning. *Fiona Anting Tan, Jay Desai and Srinivasan H. Sengamedu*
 - Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis. *Agam Shah, Arnav Hiray, Prativi Shah, Arkaprabha Banerjee, Anushka Singh, Dheeraj Deepak Eidnani, Sahasra Chava, Bhaskar Chaudhury and Sudheer Chava*
 - Streamlining Conformal Information Retrieval via Score Refinement. *Yotam Intrator, Regev Cohen, Ori Kelner, Roman Goldenberg, Ehud Rivlin and Daniel Freedman*
 - Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning. *Francielle Vargas, Isadora Salles, Diego Alves, Ameeta Agrawal, Thiago A. S. Pardo and Fabrício Benevenuto*
 - Fast Evidence Extraction for Grounded Language Model Outputs. *Pranav Mani, Davis Liang and Zachary Chase Lipton*
 - Question-Based Retrieval using Atomic Units for Enterprise RAG. *Vatsal Raina and Mark Gales*
 - AMREx: AMR for Explainable Fact Verification. *Chathuri Jayaweera, Sangpil Youm and Bonnie J Dorr*
 - Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? *Laura Majer and Jan Snajder*
 - Contrastive Learning to Improve Retrieval for Real-World Fact Checking. *Aniruddh Sriram, Fangyuan Xu, Eunsol Choi and Greg Durrett*
 - RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. *Mohammed Abdul Khalig, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pfleugfelder and Filip Miletic*

-
- FactGenius: Combining Zero-Shot Prompting and Fuzzy Relation Mining to Improve Fact Verification with Knowledge Graphs. *Sushant Gautam and Roxana Pop*
 - Fact or Fiction? Improving Fact Verification with Knowledge Graphs through Simplified Subgraph Retrievals. *Tobias Aanderaa Opsahl*
 - ProTrix: Building Models for Planning and Reasoning over Tables with Sentence Context. *Zirui Wu and Yansong Feng*
 - SparseCL: Sparse Contrastive Learning for Contradiction Retrieval. *Haike Xu, Zongyu Lin, Yizhou Sun, Kai-Wei Chang and Piotr Indyk*
 - Learning to Verify Summary Facts with Fine-Grained LLM Feedback. *Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yeon Kim, Taewon Yun, Ryan Donghan Kwon and Hwanjun Song*
 - DAHL: Domain-specific Automated Hallucination Evaluation of Long-Form Text through a Benchmark Dataset in Biomedicine. *Jean Seo, Jongwon Lim, Dongjun Jang and Hyopil Shin*
 - Detecting Misleading News Representations on Social Media Posts. *Satoshi Tohda, Naoki Yoshinaga, Masashi Toyoda, Sho Cho and Ryota Kitabayashi*
 - Evidence Retrieval for Fact Verification using Multi-stage Reranking. *Shrikant Malviya and Stamos Katsigianis*
 - Generating Media Background Checks for Automated Source Critical Reasoning. *Michael Schlichtkrull*
 - DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. *Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lio-ma and Isabelle Augenstein*
 - Zero-Shot Fact Verification via Natural Logic and Large Language Models. *Marek Strong, Rami Aly and Andreas Vlachos*
 - Do We Need Language-Specific Fact-Checking Models? The Case of Chinese. *Caiqi Zhang, Zhijiang Guo and Andreas Vlachos*
-

12:00-12:35

Contributed Shared Task Talks

- InFact: A Strong Baseline for Automated Fact-Checking. *Mark Rothermel, Tobias Braun, Marcus Rohrbach and Anna Rohrbach*
 - HerO at AVeriTec: The Herd of Open Large Language Models for Verifying Real-World Claims. *Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon and Kunwoo Park*
 - AIC CTU system at AVeriTec: Re-framing automated fact-checking as a simple RAG task. *Heribert Ullrich, Tomás Mlynár and Jan Drchal*
 - Dunamu-mls Submissions on AVERITEC Shared Task. *Heesoo Park, Dongjun Lee, Jaehyuk Kim, ChoongWon Park and Changhwa Park*
 - Multi-hop Evidence Pursuit Meets the Web: Team Papelo at FEVER 2024. *Christopher Malon*
-

12:35-14:00

Lunch Break

14:00-14:45

Keynote Talk 2: Rada Mihalcea

14:45-15:30

Keynote Talk 3: Peter Cunliffe-Jones

15:30-16:00

Coffee Break

16:00-16:30

Contributed Shared Task Talks

- Enhancing Fact Verification with Causal Knowledge Graphs and Transformer-Based Retrieval for Deductive Reasoning. *Fiona Anting Tan, Jay Desai and Srinivasan H. Sengamedu*
-

-
- Contrastive Learning to Improve Retrieval for Real-World Fact Checking.
Aniruddh Sriram, Fangyuan Xu, Eunsol Choi and Greg Durrett
 - FactGenius: Combining Zero-Shot Prompting and Fuzzy Relation Mining to Improve Fact Verification with Knowledge Graphs. *Sushant Gautam and Roxana Pop*

16:30-17:15

Keynote Talk 4: *Peter Cunliffe-Jones*

17:15-17:30

Closing Remarks

FEVER Organizers

W4 - Workshop on the Future of Event Detection

Organizers:

Joel Tetreault, Thien Huu Nguyen, Hemank Lamba, Amanda Hughes

<https://future-of-event-detection.github.io/>

Room: Johnson

Friday, November 15, 2024

In recent years, there has been a significant increase in the amount of publicly-generated digital data. One prominent category of this data, and arguably the largest in terms of daily generation, pertains to various real-world events, ranging from natural disasters to political occurrences to sports events. Detecting these events serves various crucial purposes, including early warning systems, emergency response, situational awareness, tracking public health trends, and understanding societal shifts, among others. However, automatic real-time event detection presents intriguing challenges, primarily stemming from the characteristics of the data. These challenges include the diversity of public online data (multimodal nature), the rapid pace at which data is produced (velocity), the sheer volume of data generated, and the reliability of the data (veracity). Moreover, the recent advancements in powerful Large Language Models (LLMs) and Generative AI Systems offer new opportunities to revise event detection pipelines, enabling novel approaches and applications across various domains. The workshop focuses on: The workshop focuses on: Looking forward and looking back: The workshop will solicit ideas on how the field of event detection should evolve over the next twenty years, as well as solicit papers reflecting on what has worked and not worked in the field thus far. Expanding Beyond NLP: As noted above, there are many sibling areas that actively research event detection. Many of these areas have remained siloed and there is not much cross communication though they are working on similar problem areas. This workshop seeks to address this by actively soliciting research and invited speakers from these areas. Theory to Application: Finally, this workshop will emphasize how event detection technology can be used in real-world applications.

Time	Session
09:00 09:15	Opening Remarks
09:15 10:00	Keynote: Heng Ji (UIUC)
10:00 10:30	DEGREE2: Efficient Extraction of Multiple Events Using Language Models, An Incremental Clustering Baseline for Event Detection on Twitter
10:30 11:00	Coffee Break
11:00 12:30	BERTrend: Neural Topic Modeling for Emerging Trends Detection, MUMOSA, Interactive Dashboard for MUlti-MODal Situation Awareness, A Comprehensive Survey on Document-Level Information Extraction, Generative Approaches to Event Extraction: Survey and Outlook
12:30 14:15	Lunch
14:15 15:00	Keynote: Lise St. Denis (University of Colorado, Boulder)
15:00 15:30	When and Where Did it Happen? An Encoder-Decoder Model to Identify Scenario Context (Findings), Reasoning and Tools for Human-Level Forecasting
15:30 16:00	Coffee Break
16:00 16:20	Grounding Partially-Defined Events in Multimodal Data (Findings)
16:20 17:00	Panel
17:00 17:15	Concluding Remarks

W5 - The Sixth Workshop on Narrative Understanding

Organizers:

Faeze Brahman, Anneliese Brei, Khyathi Raghavi Chandu, Snigdha Chaturvedi,
Elizabeth Clark, Yash Kumar Lal, Mohit Iyyer

<https://sites.google.com/cs.stonybrook.edu/wnu2024>

Room: Pearson

Friday, November 15, 2024

This is the 6th iteration of the Narrative Understanding Workshop, which brings together an interdisciplinary group of researchers from AI, ML, NLP, Computer Vision and other related fields, as well as scholars from the humanities to discuss methods to improve automatic narrative understanding capabilities. The workshop will consist of talks from invited speakers, a panel of researchers and writers, and talks and posters from accepted papers.

Time	Session
09:00 - 10:00	<i>Virtual Poster Session</i>
10:00 - 10:10	Opening Remarks
10:10 - 10:50	Invited Talk: Mirella Lapata
10:50 - 11:10	Break
11:10 - 11:50	Invited Talk Lydia Chilton
11:50 - 12:30	Invited Talk David Mimno
12:30 - 14:00	Lunch
14:00 - 14:40	Invited Talk Shashank Srivastava
14:40 - 15:20	Invited Talk Maarten Sap
15:20 - 15:30	Break
15:30 - 16:30	<i>In-Person Poster Session</i>

W6 - Third Workshop on NLP for Positive Impact

Organizers:

Zhijing Jin, Rada Mihalcea, Joel Tetreault, Jieyu Zhao, Steven Wilson, Oana Ignat,
Daryna Dementieva, Giorgio Piatti

<https://sites.google.com/view/nlp4positiveimpact>

Room: Foster

Friday, November 15, 2024

The Third Workshop on Positive impact continues the trend of responsible NLP models and application development including fairness, sustainability, and inclusivity. We are connecting NLP with various socially important fields like healthcare, education, environment, etc. Specifically this year, we also have invited NGOs that will showcase their challenges together with insights from an NGO expert on digital violence. We welcome specialists from various perspectives to network, foster cross-disciplinary collaboration, and spark new ideas at our workshop.

Time	Session
09:00 - 09:05	Opening Remark
09:05 - 09:30	Opening Talk by Rada Mihalcea
09:30 - 10:30	Theme Session 1 (Two Invited Talks, and 5-min Q&A) <i>Prof Yulia Tsvetkov (UW)</i>
09:30 - 10:00	<i>Prof Anjalie Field (JHU)</i>
10:00 - 10:30	
10:30 - 11:00	NGO lightning talk
11:00 - 12:00	Poster Session (In-person posters; virtual presenters in Zoom session)
12:00 - 01:00	Lunch break
13:00 - 14:00	Theme Session 2: Education (Two Invited Talks, and 5-min Q&A) <i>Prof Mrinmaya Sachan (ETH)</i>
13:00 - 13:30	<i>Stephen Mayhew (Duolingo)</i>
13:30 - 14:00	
14:00 - 15:00	Theme Session 3: Healthcare (Two Invited Talks, and 5-min Q&A) <i>Prof Veronica Perez-Rosa</i>
14:00 - 14:30	<i>Prof Louis-Philippe Morency</i>
14:30 - 15:00	
15:00 - 15:30	Oral Talk Sessions (5 talks & 5 min each, Q&A in the last 5 mins)
15:30 - 15:45	Coffee Break by EMNLP
15:45 - 15:05	Special Theme 'Digital Violence': NGO Talk by Cordelia Moore
16:05 - 17:00	Panel: 'Encouraging collaborations to advance NLP for Positive Impact' Panelists: <i>Cordelia Moore, Stephen Mayhew, Anjalie Field, and Jieyu Zhao</i>
17:00 - 17:45	Research Brainstorming with attendees: Advancing NLP for Social Good Topics: <i>Community problems, AI solutions, collaborations</i>
17:45 - 18:00	Best Paper Announcement & Closing

W7 - The Third Workshop on Text Simplification, Accessibility and Readability

Organizers:

Matthew Shardlow, Fernando Alva-Manchego, Kai North, Regina Stodden, Sanja tajner, Marcos Zampieri, Horacio Saggion

<https://tsar-workshop.github.io/>

Room: Merrick 2

Friday, November 15, 2024

The Text Simplification, Accessibility, and Readability (TSAR) workshop aims at bringing together researchers, developers and industries of assistive technologies, public organizations representatives, and other parties interested in the problem of making information more accessible to all citizens. We will discuss recent trends and developments in the area of automatic text simplification, automatic readability assessment, language resources and evaluation for text simplification.

Time	Session
09:00 09:20	Welcome Presented by: <i>Matthew Shardlow</i>
09:20 10:00	Invited Talk 1: Easy-to-understand Writing with AI Assistance <i>Walburga Fröhlich</i> Session Chair: <i>Matthew Shardlow</i>
10:00 10:30	Poster Micro-pitches Session Chair: <i>Matthew Shardlow</i>
10:30 11:00	Coffee Break
11:00 12:00	Oral Session 1: Main Track Session Chair: <i>Fernando Alva-Manchego</i> <ul style="list-style-type: none">• Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts <i>Jan Bakker and Jaap Kamps</i>• Images Speak Volumes: User-Centric Assessment of Image Generation for Accessible Communication <i>Miriam Anschütz, Tringa Sylaj and Georg Groh</i>• Society of Medical Simplifiers <i>Chen Lyu and Gabriele Pergola</i>
12:00 13:00	Lunch Break
13:00 14:40	Poster Session Session Chair: <i>Matthew Shardlow</i> (in-person), <i>Kai North</i> (virtual), <i>Regina Stodden</i> (virtual) <ul style="list-style-type: none">• CompLex-ZH: A New Dataset for Lexical Complexity Prediction in Mandarin and Cantonese [poster] <i>Le Qiu, Shanyue Guo, Tak-sum Wong, Emmanuelle Chersoni, John Sie Yuen Lee and Chu-Ren Huang</i>• MultiLS: An End-to-End Lexical Simplification Framework <i>Kai North, Tharindu Ranasinghe, Matthew Shardlow and Marcos Zampieri</i>• Considering Human Interaction and Variability in Automatic Text Simplification [poster] <i>Jenia Kim, Stefan Leijnen and Liza Beinborn</i>• OtoBERT: Simplifying Suffixed Verbal Forms in Modern Hebrew Literature <i>Avi Shmidman and Shaltiel Shmidman</i>• Difficult for Whom? A Study of Japanese Lexical Complexity <i>Adam Nohejl, Akio Hayakawa, Yusuke Ide and Taro Watanabe</i>• EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages [poster] <i>Regina Stodden</i>

-
- Evaluating the Simplification of Brazilian Legal Rulings in LLMs Using Readability Scores as a Target [poster] *Antonio Flavio Castro Paula and Celso G. Camilo-Junior*
 - Measuring and Modifying the Readability of English Texts with GPT-4 [poster] *Sean Trott and Pamela Rivière*
 - SpeciaLex: A Benchmark for In-Context Specialized Lexicon Learning (Findings) *Joseph Marvin Imperial and Harish Tayyar Madabushi*
 - README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP (Findings) *Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, README annotation team, and Hong Yu*
 - Automating Easy Read Text Segmentation (Findings) *Jesús Calleja, Thierry Etchegeoyhen, and David Ponce*

14:40 15:30	Invited Talk 2: Artificial Intelligence and Plain Language <i>Iria da Cunha</i> Session Chair: <i>Horacio Saggion</i>
15:30 16:00	Coffee Break
16:00 16:45	Oral Session 2: Special Track Session Chair: <i>Marcos Zampieri</i> <ul style="list-style-type: none">• Lexical Complexity Prediction and Lexical Simplification for Catalan and Spanish <i>Horacio Saggion, Stefan Bott, Sandra Szasz, Nelson Pérez, Saúl Calderón and Martín Solís</i>• SciGisPy: A Novel Metric for Biomedical Text Simplification via Gist Inference Score <i>Chen Lyu and Gabriele Pergola</i>
16:45 17:30	Round Table Session Chair: <i>Horacio Saggion</i>
17:30 17:35	Closing Presented by: <i>Matthew Shardlow</i>

W8 - The Eighth Widening NLP Workshop (WiNLP 2024)

Organizers:

Atnafu Lambebo Tonja, Alfredo Gomez, Chanjun Park, Hellina Hailu Nigatu,
Santosh T.Y.S.S., Tanvi Anand, Wiem Ben Rim

<https://www.winlp.org/winlp-2024-workshop/>

Room: Hibiscus

Friday, November 15, 2024

The WiNLP workshop is open to all to foster an inclusive and welcoming ACL environment. It aims to promote diversity and highlight the work of underrepresented groups in NLP: anyone who self-identifies within an underrepresented group [based on gender, ethnicity, nationality, sexual orientation, disability status, or otherwise] is invited to submit a two-page abstract for a poster presentation. In our 2024 iteration, we hope to be more intentional about centering discussions of access and disability, as well as contributing to diversity in scientific background, discipline, training, obtained degrees, seniority, and communities from underrepresented languages. The full-day event includes invited talks, oral presentations, and poster sessions. The workshop provides an excellent opportunity for junior members in the community to showcase their work and connect with senior mentors for feedback and career advice. It also offers recruitment opportunities with leading industrial labs. Most importantly, the workshop will provide an inclusive and accepting space, and work to lower structural barriers to joining and collaborating with the NLP community at large.

Time	Session
9:00 - 9:10	Welcome (Opening Session)
9:10 - 10:10	Keynote A: Danish Pruthi
10:10 - 11:00	Poster Session A
11:00 - 12:00	Panel A: Global Voices
12:00 - 12:45	Virtual Poster Session
12:45 - 13:30	Lunch
13:30 - 14:10	Mentorship Session
14:10 - 15:10	Panel B: Sailing the NLP Seas
15:10 - 16:00	Poster Session B
16:00 - 17:00	Keynote B: Alham Fikri Aji
17:00 - 17:10	Closing Session

W9 - The SIGNLL Conference on Computational Natural Language Learning (CoNLL)

Organizers:

Libby Barak, Malihe Alikhani, Mert Inan and Julia Watson

<https://conll.org/2024>

Room: Tuttle

Friday, November 15, 2024 - Saturday, November 16, 2024

CoNLL is a yearly conference organized by SIGNLL (ACL's Special Interest Group on Natural Language Learning), focusing on theoretically, cognitively and scientifically motivated approaches to computational linguistics. This year, CoNLL will be colocated with EMNLP 2024. Registrations for CoNLL can be made through EMNLP (workshop 1). The focus of CoNLL is on theoretically, cognitively and scientifically motivated approaches to computational linguistics, rather than on work driven by particular engineering applications.

Day 1 (Friday, Nov 15, 2024)

Time	Session
09:00 - 09:10	Opening Remarks
09:10 - 10:30	Keynote 1: Lorna Quandt
10:30 - 11:00	Coffee Break
11:00 - 12:30	Oral Session 1: Psycholinguistic Session (Chair: Libby Barak) <ul style="list-style-type: none">• Leveraging a Cognitive Model to Measure Subjective Similarity of Human and GPT-4 Written Content. <i>Tyler Malloy, Maria José Ferreira, Fei Fang, Cleotilde Gonzalez</i>• SPAWNing Structural Priming Predictions from a Cognitively Motivated Parser. <i>Grusha Prasad, Tal Linzen</i>• Lossy Context Surprise Predicts Task-Dependent Patterns in Relative Clause Processing. <i>Kate McCurdy, Michael Hahn</i>• Multimodal Large Language Models Foresee Objects Based on Verb Information But Not Gender. <i>Shuqi Wang, Xufeng Duan, Zhenguang Cai</i>
12:30 - 13:45	Lunch
13:45 - 15:30	Poster Session 1
15:30 - 16:00	Coffee Break
16:00 - 17:30	Oral Session 2: Syntax and Structure Session (Chair: Omri Abend) <ul style="list-style-type: none">• Is Structure Dependence Shaped for Efficient Communication? A Case Study on Coordination. <i>Kohei Kajikawa, Yusuke Kubota, Yohei Oseki</i>• NeLLCom-X: A Comprehensive Neural-Agent Framework to Simulate Language Learning and Group Communication. <i>Yuchen Lian, Tessa Verhoeft, Arianna Bisazza</i>• Solving the Challenge Set without Solving the Task: On Winograd Schemas as a Test of Pronominal Coreference Resolution. <i>Ian Porada, Jackie CK Cheung</i>• Global Learning with Triplet Relations in Abstractive Summarization. <i>Jiaxin Duan, Fengyu Lu, Junfei Liu</i>

Day 2 (Saturday, Nov 16, 2024)

Time	Session
09:00 - 09:10	Best Paper Awards
09:10 - 10:30	Keynote 2: Thamar Solorio
10:30 - 10:45	Coffee Break
10:45 - 12:15	Oral Session 3: LLM Session (Chair: Malihe Alikhani) <ul style="list-style-type: none">• Global-Pruner: A Stable and Efficient Pruner for Retraining-Free Pruning of Encoder-Based Language Models. <i>Guangzhen Yao, Sandong Zhu, Long Zhang, MiaoQI</i>• Investigating large language models for their competence in extracting grammatically sound sentences from transcribed noisy utterances. <i>Alina Wróblewska</i>• The Effect of Word Predictability on Reading Times in Information Seeking and Repeated Reading. <i>Keren Gruteke Klein, Yoav Meiri, Omer Shubi, Yevgeni Berzak</i>• Multi-Cultural Norm Base: Frame-based Norm Discovery in Multi-Cultural Settings. <i>Viet Thanh Pham, Shilin Qu, Farhad Moghimifar, Suraj Sharma, Yuan-Fang Li, Weiqing Wang, Reza Haf</i>
12:15 - 13:45	Lunch
13:45 - 15:00	Poster Session 2
15:00 - 15:30	BabyLM Challenge (oral session)
15:30 - 16:00	Coffee Break
16:00 - 17:20	BabyLM Challenge (poster session)
17:20 - 17:30	Closing Remarks

Poster Sessions

- Text2Afford: Probing Object Affordance Prediction abilities of Language Models solely from Text. *Sayantan Adak, Daivik Agrawal, Animesh Mukherjee, Somak Aditya*
- Transformer verbatim in-context retrieval across time and scale. *Kristijan Armeni, Marko Pranji, Senja Pollak*
- Of Models and Men: Probing Neural Networks for Agreement Attraction with Psycholinguistic Data. *Maxim Bazhukov, Ekaterina Voloshina, Sergey Pletenev, Arseny Anisimov, Oleg Serikov, Svetlana Toldova*
- How Are Metaphors Processed by Language Models? The Case of Analogies. *Joanne Boisson*
- AIStorySimilarity: Quantifying Story Similarity Using Narrative for Search, IP Infringement, and Guided Creativity. *Jon Chun*
- Explaining the Hardest Errors of Contextual Embedding Based Classifiers. *Claudio Moisés Valiense de Andrade, Washington Cunha, Guilherme Fonseca, Ana Clara Souza Pagano, Luana de Castro Santos, Adriana Silvina Pagano, Leonardo Chaves Dutra da Rocha, Marcos André Gonçalves*
- EditEval: An Instruction-Based Benchmark for Text Improvements. *Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, Fabio Petroni*
- Advancing Arabic Sentiment Analysis: ArSen Benchmark and the Improved Fuzzy Deep Hybrid Network. *Yang Fang, Cheng Xu, Shuhao Guan, Nan Yan, Yuke Mei*

-
- Critical Questions Generation: Motivation and Challenges. *Blanca Calvo Figueras, Rodrigo Agerri*
 - Generalizations across filler-gap dependencies in neural language models. *Katherine Howitt, Sathvik Nair, Allison Dods, Robert Melvin Hopkins*
 - Continuous Attentive Multimodal Prompt Tuning for Few-Shot Multimodal Sarcasm Detection. *Soumyadeep Jana, Animesh Dey, Ranbir Singh Samasam*
 - Aligning Alignments: Do Colexification and Distributional Similarity Align as Measures of cross-lingual Lexical Alignment?. *Taelin Karidi, Eitan Grossman, Omri Abend*
 - On Functional Competence of LLMs for Linguistic Disambiguation. *Raihan Kibria, Sheikh Intiser Uddin Dipta, Muhammad Abdullah Adnan*
 - TpT-ADE: Transformer Based Two-Phase ADE Extraction. *Suryamukhi Kuchibhotla, Manish Singh*
 - PRACT: Optimizing Principled Reasoning and Acting of LLM Agent. *Zhiwei Liu, Weiran Yao, Jianguo Zhang, Zuxin Liu, Liangwei Yang, Rithesh R N, Tian Lan, Ming Zhu, Juntao Tan, Shirley Kokane, Thai Quoc Hoang, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Silvio Savarese, Caiming Xiong*
 - Mitigating Bias in Language Model Evaluators: A Causal ATE Approach. *Rahul Madhavan, Kahini Wadhawan*
 - Words That Stick: Using Keyword Cohesion to Improve Text Segmentation. *Amit Maraj, Miguel Vargas Martin, Masoud Makrehchi*
 - An Empirical Comparison of Vocabulary Expansion and Initialization Approaches For Language Models. *Nandini Mundra, Aditya Nanda Kishore Khadavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra*
 - Revisiting Hierarchical Text Classification: Inference and Metrics. *Roman Plaud, Matthieu Labeau, Antoine Saillenfest, Thomas Bonald*
 - Image-conditioned human language comprehension and psychometric benchmarking of visual language models. *Subha Nawer Pushpita, Roger P. Levy*
 - Large Language Model Recall Uncertainty is Modulated by the Fan Effect. *Jesse Roberts, Kyle Moore, Douglas Fisher, Oseremhen Ewaleifoh, Thao Pham*
 - Self-supervised speech representations display some human-like cross-linguistic perceptual abilities. *Joselyn Rodriguez, Kamala Sreepada, Ruolan Leslie Famularo, Sharon Goldwater, Naomi Feldman*
 - One-Vs-Rest Neural Network English Grapheme Segmentation: A Linguistic Perspective. *Samuel Rose, Nina Dethlefs, C. Kambhampati*
 - CrowdCounter: A benchmark type-specific multi-target counterspeech dataset. *Punyajoy Saha, Abhilash Datta, Abhik Jana, Animesh Mukherjee*
 - Translating Across Cultures: LLMs for Intralingual Cultural Adaptation. *Pushpdeep Singh, Mayur Patidar, Lovekesh Vig*
 - Making Distilled Language Models Even Smaller: Lightweight Reconstruction of Rare Token Embeddings. *Kohki Tamura, Naoki Yoshinaga, Masato Neishi*
 - A Novel Instruction Tuning Method for Vietnamese Math Reasoning using Trainable Open-Source Large Language Models. *Nguyen Quang Vinh, Thanh-Do Nguyen, Vinh Van Nguyen, Nam Khac-Hoai Bui*
 - Information Association for Language Model Updating by Mitigating LM-Logical Discrepancy. *Pengfei Yu, Heng Ji*
-

W10 - Ninth Conference on Machine Translation (WMT24)

Organizers:

Philipp Koehn, Barry Haddow, Christof Monz, Tom Kocmi

<https://www2.statmt.org/wmt24/>

Room: Ashe Auditorium

Friday, November 15, 2024 - Saturday, November 16, 2024

WMT24 focuses on advancing machine translation techniques, discussing progress, challenges, and future directions. It features shared tasks on translation quality evaluation, multilingual translation, and domain-specific translation.

Day 1 (Friday, Nov 15, 2024)

Time	Session
8:45 - 9:00	Opening Remarks
9:00 - 9:30	Session 1: Shared Task Overview Papers I Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. <i>Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenneth Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson and Vilém Zouhar</i>
9:30 - 9:45	Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. <i>Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-ku Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jia-yi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva and Alon Lavie</i>
9:45 - 9:55	Findings of the Quality Estimation Shared Task at WMT 2024: Are LLMs Closing the Gap in QE?. <i>Chrysoula Zerva, Frederic Blain, José G. C. de Souza, Diptesh Kanodia, Sourabh Dattatray Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag and André Martins</i>
9:55 - 10:10	Findings of the WMT 2024 Shared Task of the Open Language Data Initiative. <i>Jean Maillard, Laurie V. Burchell, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn and Skyler Wang</i>
10:10 - 10:20	Results of the WAT/WMT 2024 Shared Task on Patent Translation. <i>Shohei Higashiyama</i>
10:20 - 10:30	Findings of the WMT 2024 Biomedical Translation Shared Task: Test Sets on Abstract Level. <i>Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova and Antonio Jimeno Yepes</i>
10:30 - 11:00	Coffee Break

11:00 - 12:00

Session 2: Shared Task Posters I

General Translation Task

- MSLC24 Submissions to the General Machine Translation Task. *Samuel Larkin, Chi-kiu Lo and Rebecca Knowles*
- IOL Research Machine Translation Systems for WMT24 General Machine Translation Shared Task. *Wenbo Zhang*
- Choose the Final Translation from NMT and LLM Hypotheses Using MBR Decoding: HW-TSCs Submission to the WMT24 General MT Shared Task. *Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin GUO, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie and Hao Yang*
- CycleGN: A Cycle Consistent Approach for Neural Machine Translation. *Sören Dreano, Derek Molloy and Noel Murphy*
- UvA-MTs Participation in the WMT24 General Translation Shared Task. *Shaomu Tan, David Stap, Seth Aycock, Christof Monz and Di Wu*
- Tower v2: Unbabel-IST 2024 Submission for the General MT Shared Task. *Ricardo Rei, Jose Maria Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. de Souza and André Martins*
- TSU HITSS Submissions to the WMT 2024 General Machine Translation Shared Task. *Vladimir Mynka and Nikolay Mikhaylovskiy*
- Document-level Translation with LLM Reranking: Team-J at WMT 2024 General Translation Task. *Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai and Jun Suzuki*
- DLUT and GTCOMs Neural Machine Translation Systems for WMT24. *Hao Zong, Chao Bei, Huan Liu, Conghu Yuan, Wentao Chen and Degen Huang*
- CUNI at WMT24 General Translation Task: LLMs, (Q)LoRA, CPO and Model Merging. *Miroslav Hrabal, Josef Jon, Martin Popel, Nam Luu, Danil Semin and Ondej Bojar*
- From General LLM to Translation: How We Dramatically Improve Translation Quality Using Human Evaluation Data for LLM Finetuning. *Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev and Kirill Denisov*
- Cogs in a Machine, Doing What Theyre Meant to Do the AMI Submission to the WMT24 General Translation Task. *Atli Jasonarson, Hinrik Hafsteinsson, Bjarki Ármannsson and Steinþór Steingrímsson*
- IKUN for WMT24 General MT Task: LLMs Are Here for Multilingual Machine Translation. *Baohao Liao, Christian Herold, Shahram Khadivi and Christof Monz*
- NTTSU at WMT2024 General Translation Task. *Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomu Kano and Takehito Utsuro*
- SCIR-MTs Submission for WMT24 General Machine Translation Task. *Bao-hang Li, Zekai Ye, Yichong Huang, Xiaocheng Feng and Bing Qin*

	<ul style="list-style-type: none"> • AIST AIRC Systems for the WMT 2024 Shared Tasks. <i>Matiss Rikters and Makoto Miwa</i> • Occiglot at WMT24: European Open-source Large Language Models Evaluated on Translation. <i>Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm and Kristian Kersting</i> • Test Suites • CoST of breaking the LLMs. <i>Ananya Mukherjee, Saumitra Yadav and Manish Srivastava</i> • WMT24 Test Suite: Gender Resolution in Speaker-Listener Dialogue Roles. <i>Hillary Dawkins, Isar Nejadgholi and Chi-ku Lo</i> • The GenderQueer Test Suite. <i>Steinunn Rut Friðriksdóttir</i> • Domain Dynamics: Evaluating Large Language Models in English-Hindi Translation. <i>Soham Bhattacharjee, Baban Gain and Asif Ekbal</i> • Investigating the Linguistic Performance of Large Language Models in Machine Translation. <i>Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov and Sebastian Möller</i> • IsoChronoMeter: A Simple and Effective Isochronic Translation Evaluation Metric. <i>Nikolai Rozanov, Vikentiy Pankov, Dmitrii Mukhutdinov and Dima Vypirailenko</i> • A Test Suite of Prompt Injection Attacks for LLM-based Machine Translation. <i>Antonio Valerio Miceli Barone and Zhifan Sun</i> • Killing Two Flies with One Stone: An Attempt to Break LLMs Using English/Icelandic Idioms and Proper Names. <i>Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jónasson and Steinþor Steingrimsson</i>
12:30 - 13:30	<p>Session 3: Shared Task Posters II</p> <p>Metrics Task</p> <ul style="list-style-type: none"> • MetaMetrics-MT: Tuning Meta-Metrics for Machine Translation via Human Preference Calibration. <i>David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya and Genta Indra Winata</i> • chrF-S: Semantics Is All You Need. <i>Ananya Mukherjee and Manish Srivastava</i> • MSLC24: Further Challenges for Metrics on a Wide Landscape of Translation Quality. <i>Rebecca Knowles, Samuel Larkin and Chi-ku Lo</i> • MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. <i>Juraj Juraska, Daniel Deutsch, Mara Finkelstein and Markus Freitag</i> • Evaluating WMT 2024 Metrics Shared Task Submissions on AfriMTE (the African Challenge Set). <i>Jiayi Wang, David Ifeoluwa Adelani and Pontus Stenertorp</i> • Machine Translation Metrics Are Better in Evaluating Linguistic Errors on LLMs than on Encoder-Decoder Systems. <i>Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz and Sebastian Möller</i> • Quality Estimation Task • TMU-HITs Submission for the WMT24 Quality Estimation Shared Task: Is GPT-4 a Good Evaluator for Machine Translation?. <i>Ayako Sato, Kyotaro Nakajima, Hwican Kim, Zhousi Chen and Mamoru Komachi</i>

-
- HW-TSC 2024 Submission for the Quality Estimation Shared Task. *Weiqiao Shan, Ming Zhu, Yuang Li, Mengyao Piao, Xiaofeng Zhao, Chang Su, Min Zhang, Hao Yang and Yanfei Jiang*

- HW-TSCs Participation in the WMT 2024 QEAPE Task. *Jiawei Yu, Xiaofeng Zhao, Min Zhang, Zhao Yanqing, Yuang Li, Su Chang, Xiasong Qiao, Ma Miao-miao and Hao Yang*

- **Open Language Data Initiative**

- Expanding the FLORES+ Multilingual Benchmark with Translations for Aragonese, Aranese, Asturian, and Valencian. *Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Espà-Gomis, Aaron Galiano Jiménez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau and Juan Pablo Martínez*

- The Bangla/Bengali Seed Dataset Submission to the WMT24 Open Language Data Initiative Shared Task. *Firoz Ahmed, Nitin Venkateswaran and Sarah Moeller*

- A High-quality Seed Dataset for Italian Machine Translation. *Edoardo Fer-rante*

- Correcting FLORES Evaluation Dataset for Four African Languages. *Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo N. Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe and Yukosi Marivate*

- Expanding FLORES+ Benchmark for More Low-Resource Settings: Portuguese-Emakhuwa Machine Translation Evaluation. *Felermimo Dario Mario Ali, Henrique Lopes Cardoso and Rui Sousa-Silva*

- Enhancing Tuval Language Resources through the FLORES Dataset. *Ali Kuzhuget, Airana Mongush and Nachyn-Enkhedorzhoo Oorzhak*

- Machine Translation Evaluation Benchmark for Wu Chinese: Workflow and Analysis. *Hongjian Yu, Yiming Shi, Zherui Zhou and Christopher Haberland*

- Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak. *Mukhammadsaid Mamasaidov and Abror Shopulatov*

- FLORES+ Translation and Machine Translation Evaluation for the Erzya Lan-guage. *Isai Gordeev, Sergey Kuldin and David Dale*

- Spanish Corpus and Provenance with Computer-Aided Translation for the WMT24 OLDI Shared Task. *Jose Cols*

Patent Translation Task

- Efficient Terminology Integration for LLM-based Translation in Specialized Domains. *Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim and Jorge Froilan Gimenez Perez*

- Rakutens Participation in WMT 2024 Patent Translation Task. *Ohnmar Htun and Alberto Poncelas*

Biomedical Translation Task

- The SETU-ADAPT Submission for WMT 24 Biomedical Shared Task. *Anto-nio Castaldo, Maria Zafar, Prashanth Nayak, Rejwanul Haque, Andy Way and Johanna Monti*

14:00 - 15:00	Session 4: Invited Talk by Ricardo Rei and Nuno M. Guerreiro: 'What Makes MT Research Special in the LLM Age?'
---------------	---

15:00 - 15:30	Coffee Break
---------------	---------------------

Session 5: Featured Research Papers Oral Presentations

15:00 - 15:15	Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. <i>Eleftheria Briakou, Jiaming Luo, Colin Cherry and Markus Freitag</i>
15:15 - 15:30	Is Preference Alignment Always the Best Option to Enhance LLM-Based Translation? An Empirical Analysis. <i>Hippolyte Gisserot-Boukhlef, Ricardo Rei, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo and Nuno M. Guerreiro</i>
15:30 - 15:45	On Instruction-Finetuning Neural Machine Translation Models. <i>Vikas Raunak, Roman Grundkiewicz and Marcin Junczys-Dowmunt</i>
15:45 - 16:00	Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation. <i>Vivek Iyer, Bhavitya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow and Alexandra Birch</i>
16:00 - 16:15	Post-edits Are Preferences Too. <i>Nathaniel Berger, Stefan Riezler, Miriam Exel and Matthias Huck</i>
16:15 - 16:30	Benchmarking Visually-Situated Translation of Text in Natural Images. <i>Elizabeth Salesky, Philipp Koehn and Matt Post</i>

Day 2 (Saturday, Nov 16, 2024)

Time	Session
Session 6: Shared Task Overview Papers II	
9:00 - 9:15	Findings of WMT 2024 Shared Task on Low-Resource Indic Languages Translation. <i>Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Mohana Krishna, Arnab Kumar Maji, Sandeep Kumar Dash, Lenin Laitonjam, Lyngdoh Sarah and Riyanka Manna</i>
9:15 - 9:30	Findings of WMT 2024s MultiIndic22MT Shared Task for Machine Translation of 22 Indian Languages. <i>Raj Dabre and Anoop Kunchukuttan</i>
9:30 - 9:45	Findings of WMT2024 English-to-Low Resource Multimodal Translation Task. <i>Shantipriya Parida, Ondej Bojar, Idris Abdulkummin, Shamsuddeen Hassan Muhammad and Ibrahim Said Ahmad</i>
9:45 - 10:00	Findings of the WMT 2024 Shared Task Translation into Low-Resource Languages of Spain: Blending Rule-Based and Neural Systems. <i>Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aaron Galiano Jimenez and Antoni Oliver</i>
10:00 - 10:10	Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation. <i>Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jia-hao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way and Yulin Yuan</i>
10:10 - 10:20	Findings of the WMT 2024 Shared Task on Chat Translation. <i>Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha and José G. C. de Souza</i>
10:20 - 10:30	Findings of the WMT 2024 Shared Task on Non-Repetitive Translation. <i>Kazutaka Kinugawa, Hideya Mino, Isao Goto and Naoto Shirai</i>
10:30 - 11:00	Coffee Break
11:00 - 12:00	Session 7: Shared Task Posters III Low-Resource Indic Language Translation Task • A3-108 Controlling Token Generation in Low Resource Machine Translation Systems. <i>Saumitra Yadav, Ananya Mukherjee and Manish Srivastava</i>

-
- Samsung R&D Institute Philippines @ WMT 2024 Indic MT Task. *Matthew Theodore Roque, Carlos Rafael Catalan, Dan John A. Velasco, Manuel Antonio Rufino and Jan Christian Blaise Cruz*
 - DLUT-NLP Machine Translation Systems for WMT24 Low-Resource Indic Language Translation. *Chenfei Ju, Junpeng Liu, Kaiyu Huang and Degen Huang*
 - SRIB-NMTs Submission to the Indic MT Shared Task in WMT 2024. *Pranamya Ajay Patil, Raghavendra HR, Aditya Raghuwanshi and Kushal Verma*
 - MTNLP-IIITH: Machine Translation for Low-Resource Indic Languages. *Abinav P M, Ketaki Shetye and Parameswari Krishnamurthy*
 - Exploration of the CycleGN Framework for Low-Resource Languages. *Sören DREANO, Derek MOLLOY and Noel MURPHY*
 - The SETU-ADAPT Submissions to the WMT24 Low-Resource Indic Language Translation Task. *Neha Gajakos, Prashanth Nayak, Rejwanul Haque and Andy Way*
 - SPRING Lab IITMs Submission to Low Resource Indic Language Translation Shared Task. *Advait Joglekar, Hamees Ul Hasan Sayed and Srinivasan Umesh*
 - Machine Translation Advancements of Low-Resource Indian Languages by Transfer Learning. *Bin Wei, Zheng Jiawei, Zongyao Li, Zhanglin Wu, Jiaxin GUO, Daimeng Wei, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie and Hao Yang*
 - NLIP_Lab-IIITH Low-Resource MT System for WMT24 Indic MT Shared Task. *Pramit Sahoo, Maharaj Brahma and Maunendra Sankar Desarkar*
 - Yes-MTs Submission to the Low-Resource Indic Language Translation Shared Task in WMT 2024. *Yash Bhaskar and Parameswari Krishnamurthy*

MultiIndic22MT Task

- System Description of BV-SLP for Sindhi-English Machine Translation in MultiIndic22MT 2024 Shared Task. *Nisheeth Joshi, Pragya Katyayan, Palak Arora and Bharti Nathani*
- WMT24 System Description for the MultiIndic22MT Shared Task on Manipuri Language. *Ningthoujam Justwani Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam and Thoudam Doren Singh*
- NLIP_Lab-IIITH Multilingual MT System for WAT24 MT Shared Task. *Maharaj Brahma, Pramit Sahoo and Maunendra Sankar Desarkar*

English-to-Lowres Multi-Modal Translation Task

- DCU ADAPT at WMT24: English to Low-resource Multi-Modal Translation Task. *Sami Ul Haq, Rudali Huidrom and Sheila Castilho*
- English-to-Low-Resource Translation: A Multimodal Approach for Hindi, Malayalam, Bengali, and Hausa. *Ali Hatami, Shubhanker Banerjee, Mihael Ar- can, Bharathi Raja Chakravarthi, Paul Buitelaar and John Philip McCrae*
- OdiaGenAIs Participation in WMT2024 English-to-Low Resource Multimodal Translation Task. *Shantipriya Parida, Shashikanta Sahoo, Sambit Sekhar, Open- dra Kumar Jena, Sushovan Jena and Kusum Lata*
- Arewa NLPs Participation at WMT24. *Mahmoud Said Ahmad, Auwal Abubakar Khalid, Lukman Jibril Aliyu, Babangida Sani and Mariya Sunusi Abdulla*
- Multimodal Machine Translation for Low-Resource Indic Languages: A Chain-of-Thought Approach Using Large Language Models. *Pawan Kumar Ra- jpoot, Nagaraj N. Bhat and Ashish Shrivastava*

	<ul style="list-style-type: none"> • Chitranuvad: Adapting Multi-lingual LLMs for Multimodal Translation. <i>Shahrukh Khan, Ayush Tarun, Ali Faraz, Palash Kamble, Vivek Dahiya, Praveen Pokala, Ashish Anand Kulkarni, Chandra Khatri, Abhinav Ravi and Shubham Agarwal</i> • Brotherhood at WMT 2024: Leveraging LLM-Generated Contextual Conversations for Cross-Lingual Image Captioning. <i>Siddharth Betala and Ishan Chokshi</i>
12:30 - 13:30	<p>Session 8: Shared Task Posters IV</p> <p>Translation into Low-Resource Languages of Spain Task.</p> <ul style="list-style-type: none"> • TIM-UNIGE Translation into Low-Resource Languages of Spain for WMT24. <i>Jonathan Mutal and Lucía Ormaechea</i> • TAN-IBE Participation in the Shared Task: Translation into Low-Resource Languages of Spain. <i>Antoni Oliver</i> • Enhanced Apertium System: Translation into Low-Resource Languages of Spain Spanish - Asturian. <i>Sofia García</i> • Universitat dAlacants Submission to the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain. <i>Aaron Galiano Jimenez, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz and Felipe Sánchez-Martínez</i> • Samsung R&D Institute Philippines @ WMT 2024 Low-resource Languages of Spain Shared Task. <i>Dan John A. Velasco, Manuel Antonio Rufino and Jan Christian Blaise Cruz</i> • Back to the Stats: Rescuing Low Resource Neural Machine Translation with Statistical Methods. <i>Menan Velayuthan, Dilith Randinu Jayakody, Nisansa de Silva, Aloka Fernando and Surangika Dayani Ranathunga</i> • Hybrid Distillation from RBMT and NMT: Helsinki-NLPs Submission to the Shared Task on Translation into Low-Resource Languages of Spain. <i>Ona de Gibert, Mikko Aulamo, Yves Scherrer and Jörg Tiedemann</i> • Robustness of Fine-Tuned LLMs for Machine Translation with Varying Noise Levels: Insights for Asturian, Aragonese and Aranese. <i>Martin Bär, Elisa Forcada Rodríguez and María García-Abadillo Velasco</i> • Training and Fine-Tuning NMT Models for Low-Resource Languages Using Apertium-Based Synthetic Corpora. <i>Aleix Sant, Daniel Bardanca, José Ramon Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier García Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao and Maite Melero</i> • Vicomtech@WMT 2024: Shared Task on Translation into Low-Resource Languages of Spain. <i>David Ponce, Harritxu Gete and Thierry Etchegoyen</i> • SJTU System Description for the WMT24 Low-Resource Languages of Spain Task. <i>Tianxiang Hu, Haoxiang Sun, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang and Rui Wang</i> • Multilingual Transfer and Domain Adaptation for Low-Resource Languages of Spain. <i>Yuanchang Luo, Zhanglin Wu, Daimeng Wei, Hengchao Shang, Zongyao Li, Jiaxin GUO, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Yuhao Xie, Zheng Jiawei, Bin Wei and Hao Yang</i> • TRIBBLE - TRanslating IBERian languages Based on Limited E-resources. <i>Igor Kuzmin, Piotr Przybyla, Euan McGill and Horacio Saggion</i> <p>Discourse-Level Literary Translation Task</p> <ul style="list-style-type: none"> • CloudSheep System for WMT24 Discourse-Level Literary Translation. <i>Lisa Liu, Ryan Liu, Angela Tsai and Jingbo Shang</i>

-
- Final Submission of SJTULoveFiction to Literary Task. *Haoxiang Sun, Tianxiang Hu, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang and Rui Wang*
 - Context-aware and Style-related Incremental Decoding Framework for Discourse-Level Literary Translation. *Yuanchang Luo, Jiaxin GUO, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li, Jinlong Yang and Hao Yang*
 - NovelTrans: System for WMT24 Discourse-Level Literary Translation. *Yuchen Liu, Yutong Yao, Runzhe Zhan, Yuchu Lin and Derek F. Wong*
 - LinChanceENTU for Unconstrained WMT2024 Literary Translation. *Kechen Li, Yaotian Tao, Hongyi Huang and Tianbo Ji*

Chat Translation Task

- Improving Context Usage for Translating Bilingual Customer Support Chat with Large Language Models. *Jose Maria Pombal, Sweta Agrawal and André Martins*
- Optimising LLM-Driven Machine Translation with Context-Aware Sliding Windows. *Xinye Yang, Yida Mu, Kalina Bontcheva and Xingyi Song*
- Context-Aware LLM Translation System Using Conversation Summarization and Dialogue History. *Mingi Sung, Seungmin Lee, Jiwon Kim and Sejoon Kim*
- Enhancing Translation Quality: A Comparative Study of Fine-Tuning and Prompt Engineering in Dialog-Oriented Machine Translation Systems. Insights from the MULTITAN-GML Team. *Lichao Zhu, Maria Zimina, Behnoosh Namdarzadeh, Nicolas Ballier and Jean-Baptiste Yunès*
- The SETU-ADAPT Submissions to WMT 2024 Chat Translation Tasks. *Maria Zafar, Antonio Castaldo, Prashanth Nayak, Rejwanul Haque and Andy Way*
- Exploring the Traditional NMT Model and Large Language Model for Chat Translation. *Jinlong Yang, Hengchao Shang, Daimeng Wei, Jiaxin GUO, Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Shaojun Li, Yuhao Xie, Yuanchang Luo, Zheng Jiawei, Bin Wei and Hao Yang*

Non-Repetitive Translation Task

- Reducing Redundancy in Japanese-to-English Translation: A Multi-Pipeline Approach for Translating Repeated Elements in Japanese. *Qiao Wang, Yixuan Huang and Zheng Yuan*
- SYSTRAN @ WMT24 Non-Repetitive Translation Task. *Marko Avila and Josep Crego*

14:00 - 15:30	Session 9: Research Paper Boaster Session
15:30 - 16:30	Session 10: Research Paper Poster Session I <ul style="list-style-type: none">• Mitigating Metric Bias in Minimum Bayes Risk Decoding. <i>Geza Kovacs, Daniel Deutsch and Markus Freitag</i>• Beyond Human-Only: Evaluating Human-Machine Collaboration for Collecting High-Quality Translation Data. <i>Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah and Markus Freitag</i>• How Effective Are State Space Models for Machine Translation? <i>Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso and André Martins</i>• Evaluation and Large-scale Training for Contextual Machine Translation. <i>Matt Post and Marcin Junczys-Dowmunt</i>• A Multi-task Learning Framework for Evaluating Machine Translation of Emotion-loaded User-generated Content. <i>Shenbin Qian, Constantin Orasan, Diptesh Kanodia and Félix do Carmo</i>

-
- On Instruction-Finetuning Neural Machine Translation Models. *Vikas Raunak, Roman Grundkiewicz and Marcin Junczys-Dowmunt*
 - Benchmarking Visually-Situated Translation of Text in Natural Images. *Elizabeth Salesky, Philipp Koehn and Matt Post*
 - Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. *Fedor Sizov, Cristina España-Bonet, Josef van Genabith, Roy Xie and Koel Dutta Chowdhury*
 - How Grammatical Features Impact Machine Translation: A New Test Suite for Chinese-English MT Evaluation. *Huacheng Song, Yi Li, Yiwen Wu, Yu Liu, Jingxia Lin and Hongzhi Xu*
 - Improving Statistical Significance in Human Evaluation of Automatic Metrics via Soft Pairwise Accuracy. *Brian Thompson, Nitika Mathur, Daniel Deutsch and Huda Khayrallah*
 - Speech Is More than Words: Do Speech-to-Text Translation Systems Leverage Prosody? *Ioannis Tsiamas, Matthias Sperber, Andrew Finch and Sarthak Garg*
 - Cultural Adaptation of Menus: A Fine-Grained Approach. *Zhonghe Zhang, Xiaoyu He, Vivek Iyer and Alexandra Birch*
 - Pitfalls and Outlooks in Using COMET. *Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe and Barry Haddow*

16:30 - 17:00	Coffee Break
17:00 - 18:00	<p>Session 11: Research Paper Poster Session II</p> <ul style="list-style-type: none"> • Post-edits Are Preferences Too. <i>Nathaniel Berger, Stefan Riezler, Miriam Exel and Matthias Huck</i> • Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. <i>Eleftheria Briakou, Jiaming Luo, Colin Cherry and Markus Freitag</i> • Scaling Laws of Decoder-Only Models on the Multilingual Machine Translation Task. <i>Gaëtan Caillaut, Mariam Nakhlé, Raheel Qader, Jingshu Liu and Jean-Gabriel Barthélémy</i> • Shortcomings of LLMs for Low-Resource Translation: Retrieval and Understanding Are Both the Problem. <i>Sara Kay Court and Micha Elsner</i> • Introducing the NewsPaLM MBR and QE Dataset: LLM-Generated High-Quality Parallel Data Outperforms Traditional Web-Crawled Data. <i>Mara Finkelstein, David Vilar and Markus Freitag</i> • Is Preference Alignment Always the Best Option to Enhance LLM-Based Translation? An Empirical Analysis. <i>Hippolyte Gisserot-Boukhlef, Ricardo Rei, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo and Nuno M. Guerreiro</i> • Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation. <i>Vivek Iyer, Bhavitya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow and Alexandra Birch</i> • Efficient Technical Term Translation: A Knowledge Distillation Approach for Parenthetical Terminology Translation. <i>Myung Jiyoob, Jihyeon Park, Jungki Son, Kyungro Lee and Joohyung Han</i> • Assessing the Role of Imagery in Multimodal Machine Translation. <i>Nicholas Kashani Motlagh, Jim Davis, Jeremy Gwinnup, Grant Erdmann and Tim Anderson</i>

-
- Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. *Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grunkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan and Mariya Shmatova*
 - Neural Methods for Aligning Large-Scale Parallel Corpora from the Web for South and East Asian Languages. *Philipp Koehn*
 - Plug, Play, and Fuse: Zero-Shot Joint Decoding via Word-Level Re-ranking across Diverse Vocabularies. *Sai Koneru, Matthias Huck, Miriam Exel and Jan Niehues*
-

W11 - Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)

Organizers:

Chan Young Park, Vidhisha Balachandran, Weijia Shi, Shirley Anugrah Hayati,
Sachin Kumar

<https://customnlp4u-24.github.io/>

Room: Merrick I

Saturday, November 16, 2024

CustomNLP4U explores the customization of NLP systems for specific domains or user groups, focusing on the challenges of tailoring NLP for personalized applications. For NLP models to be usable in practice, particularly in emerging scenarios with widely varying use cases, situations, and user expectations, there is a need to develop models that can be tailored to different consumers (individuals, groups, or organizations) and easily controlled by them; models that can reason about their users (often private) knowledge and context to provide personalized responses. The topics of this workshop include (but not limited to): Data collection, processing, analysis, and annotation efforts to increase representation and aid customization; discussion and analysis of data sources not publicly available, and associated issues of privacy and copyright. Modeling: New pretraining, fine-tuning, inference methods for customizing NLP models; customizing reward models and model alignment to diverse consumers. New modeling paradigms aimed at customization such as model ensembles, model averaging, federated learning, nonparametric models, etc.; customizing models at inference time via prompting, in-context learning, chain-of-thought prompting, etc. Evaluation: Evaluation of existing generalist, non-customized models, identifying their shortcomings for varied use-cases; evaluation of customization techniques and customized models; interpretability and analysis of customization patterns across different kinds of consumers. Open Science: Best practices for open and reproducible science concerning customizable NLP: dataset release and licensing, open-sourcing models, related privacy, copyright, and policy issues. Applications: e.g., information seeking on sensitive data comprising legal, medical, or financial information; NLP models for communities reflecting sociolects, dialects, or other language varieties; personalized AI assistants, etc. Ethical Issues: privacy and copyright; personalization, intrusiveness, unintended biases; invisibility versus hypervisibility.

Time	Session
9:00 - 9:15	Opening Remarks
9:15 - 10:00	Invited Talk 1 - Diyi Yang (Stanford)
10:00 - 10:45	Invited Talk 2 - Hanna Kirk (Oxford)
10:45 - 11:00	Coffee Break
11:00 - 12:00	Poster Session
13:00 - 13:15	Lightening Slides
13:15 - 14:00	Invited Talk 3 - Jared Roesch
14:00 - 14:45	Invited Talk 4
14:45 - 15:30	Invited Talk 5 - Maartje Ter Hoeve (Apple)
15:30 - 16:00	Coffee Break
16:00 - 16:30	Outstanding Papers Oral Presentations (10 min each)

-
1. Constructing Domain-Specific Evaluation Sets for LLM-as-a-judge
 2. Trustful LLMs: Customizing and Grounding Text Generation with knowledge bases and Dual Decoders
 3. Customizing LLM Generation in Safety Scenarios with Active Learning for Enhanced Representativeness and Robustness
-

16:30 - 17:00

Best Paper Award + Closing Remarks

W12 - The 4th International Workshop on Natural Language Processing for Digital Humanities (NLP4DH)

Organizers:

Mika Hämäläinen, Emily Öhman, Khalid Alnajjar, So Miyagawa, Yuri Bizzoni

<https://www.nlp4dh.com/nlp4dh-2024>

Room: Merrick 2

Saturday, November 16, 2024

The 4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2024) will be organized together with EMNLP 2024. The proceedings of the conference will be published in the ACL anthology. The conference will take place in Miami, USA on November 16, 2024. The focus of the conference is on applying natural language processing techniques to digital humanities research. The topics can be anything of digital humanities interest with a natural language processing or generation aspect. A list of suitable topics includes but is not limited to: Text analysis and processing related to humanities using computational methods; Thorough error analysis of an NLP system using (digital) humanities methods; Dataset creation and curation for NLP (e.g. digitization, digitalization, datafication, and data preservation); Research on cultural heritage collections such as national archives and libraries using NLP; NLP for error detection, correction, normalization and denoising data; Generation and analysis of literary works such as poetry and novels; Analysis and detection of text genres.

Time	Session
9:00 9:10	Opening words
	Oral session 1
9:10 9:30	Lightning talks
9:30 9:50	Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits <i>Emily Sofi Öhman and Aatu Liimatta</i>
9:50 10:10	Tracing the Genealogies of Ideas with Sentence Embeddings <i>Lucian Li</i>
10:10 10:30	Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark <i>Funing Yang and Carolyn Jane Anderson</i>
10:30 11:00	Coffee break
	Oral session 2
11:00 11:20	Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis <i>Ray Umphrey, Jesse Roberts, Lindsey Roberts</i>
11:20 11:40	Extracting Relations from Ecclesiastical Cultural Heritage Texts <i>Giulia Cruciani</i>
11:40 12:00	Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience <i>Lucija Krusic</i>
12:00 12:20	It is a Truth Individually Acknowledged: Cross-references On Demand <i>Piper Vasicek, Courtney Byun, Kevin Seppi</i>
12:20 12:40	Extracting Position Titles from Unstructured Historical Job Advertisements <i>Klara Venglarová, Raven Adam, Georg Vogeler</i>
12:40 13:10	Lunch
	Oral session 3

13:10 13:30	Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs <i>Natalie Hervieux, Peiran Yao, Susan Brown, Denilson Barbosa</i>
13:30 13:50	NLP for Digital Humanities: Processing Chronological Text Corpora <i>Adam Pawowski, Tomasz Walkowiak</i>
13:50 14:10	A Multi-task Framework with Enhanced Hierarchical Attention for Sentiment Analysis on Classical Chinese Poetry: Utilizing Information from Short Lines <i>Quanqi Du and Veronique Hoste</i>
14:10 14:30	Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature <i>So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi, Kyoko Amano</i>
14:30 14:50	Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction <i>Laura Manrique-Gomez, Tony Montes, Arturo Rodriguez Herrera, Ruben Manrique</i>
14:50 15:10	Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870 - 1900) <i>Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig, Kristoffer Nielbo</i>
15:10 15:30	Deciphering Psycho-Social Effects of Eating Disorder: Analysis of Reddit Posts using Large Language Models and Topic Modeling <i>Medini Chopra, Anindita Chatterjee, Lipika Dey, Partha Pratim Das</i>
15:30 16:30	Posters and coffee <ul style="list-style-type: none"> • Topic-Aware Causal Intervention for Counterfactual Detection <i>Thong Thanh Nguyen, Truc-My Nguyen</i> • UD for German Poetry <i>Stefanie Dipper, Ronja Laermann-Quante</i> • Molyé: A Corpus-Based Approach to Language Contact in Colonial France <i>Rasul Dent, Juliette Janes, Thibault Clerice, Pedro Ortiz Suarez, Benoît Sagot</i> • Improving Latin Dependency Parsing by Combining Treebanks and Predictions <i>Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, Jenna Kanerva</i> • From N-Grams to Pre-Trained Multilingual Models for Language Identification <i>Thapelo Andrew Sindane, Vukosi Marivate</i> • Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time <i>Malak Rassem, Myrto Tsigkouli, Chris W Jenkins, Filip Mileti, Sabine Schulte im Walde</i> • SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek <i>Konstantin Schulz, Florian Deichsler</i> • A Two-Model Approach for Humour Style Recognition <i>Mary Ogbuka Kenneth, Foaad Khosmood, Abbas Edalat</i> • N-Gram-Based Preprocessing for Sandhi Reversion in Vedic Sanskrit <i>Yuzuki Tsukagoshi, Ikkai Ohmukai</i> • Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams <i>Roberts Daris, Gunīts Brzdi, Inguna Skadia, Baiba Saulīte</i> • Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew 'she-' and 'ki' <i>Avi Shmidman, Aynat Rubinstein</i>

-
- From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations *Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, Veronika Laippala*
 - Testing and Adapting the Representational Abilities of Large Language Models on Folktales in Low-Resource Languages *J. A. Meaney, Beatrice Alex, William Lamb*
 - Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus *Craig Messner, Thomas Lippincott*
 - Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts *Ayuki Katayama, Yusuke Sakai, Shohei Higashiyama, Hiroki Ouchi, Ayano Takeuchi, Ryo Bando, Yuta Hashimoto, Toshinobu Ogiso, Taro Watanabe*
 - Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction *Woori Jang, Seohyon Jung*
 - Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin *Svetlana Gorovaia, Gleb Schmidt, Ivan P. Yamshchikov*

16:30 17:30

Virtual posters (online on Gather)

- Classification of Buddhist Verses: The Efficacy and Limitations of Transformer-Based Models *Nikita Neveditsin, Ambuja Salgaonkar, Pawan Lingras, Vijay Mago*
- Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities *Shafqat Mumtaz Virk, Claes Ohlsson, Nina Tahmasebi, Henrik Björck, Leif Runefelt*
- Adapting Measures of Literality for Use with Historical Language Data *Adam Roussel*
- Vector Poetics: Parallel Couplet Detection in Classical Chinese Poetry *Maciej Kurzynski, Xiaotong Xu, Yu Feng*
- Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora *Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala, Erik Henriksson*
- Text vs. Transcription: A Study of Differences Between the Writing and Speeches of U.S. Presidents *Mina Rajaei Moghadam, Mosab Rezaei, Güllat Aygen, Reva Freedman*
- Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language *Xinmeng Hou*
- Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration *Ryo Igarashi, So Miyagawa*
- Exploring Large Language Models for Qualitative Data Analysis *Tim Fischer, Chris Biemann*
- Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs *Chahan Vidal-Gorène, Nadi Tomeh, Victoria Khurshudyan*
- Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference for Cost-Effective Cultural Heritage Dataset Generation *William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin, Diana Maynard*

-
- Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca *Audric-Charles Wannaz, So Miyagawa*
 - The Social Lives of Literary Characters: Combining Citizen Science and Language Models to Understand Narrative Social Networks *Andrew Piper, Michael Xu, Derek Ruths*
 - Multi-Word Expressions in Biomedical Abstracts and Their Plain English Adaptations *Sergei Bagdasarov, Elke Teich*
 - Assessing the Performance of ChatGPT-4, Fine-Tuned BERT and Traditional ML Models on Moroccan Arabic Sentiment Analysis *Mohamed Hannani, Abdellah Soudi, Kristof Van Laerhoven*
 - Analyzing Pokémon and Mario Streamers' Twitch Chat with LLM-Based User Embeddings *Mika Hämäläinen, Jack Rueter, Khalid Alnajjar*
 - Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification *Keito Inoshita*
 - Generating Interpretations of Policy Announcements *Andreas Marfurt, Ashley Thornton, David Sylvan, James Henderson*
 - Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses *Erkki Mervaala, Ilona Kousa*
 - CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration *Anton Eklund, Mona Forsman, Frank Drewes*
 - Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy *Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen, Mika Hämäläinen*
-

W13 - GenBench: The second workshop on generalisation (benchmarking) in NLP

Organizers:

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Amirhossein Kazemnejad,
Khuyagbaatar Batsuren, Ryan Cotterell, Christos Christodoulopoulos

<https://genbench.org/workshop/>

Room: Brickell

Saturday, November 16, 2024

The ability to generalise well is often mentioned as one of the primary desiderata for models of natural language processing (NLP). Yet, there are still many open questions related to what it means for an NLP model to generalise well, and how generalisation should be evaluated. LLMs, trained on gigantic training corpora that are - at best - hard to analyse or not publicly available at all, bring a new set of challenges to the topic. The second GenBench workshop aims to serve as a cornerstone to catalyse research on generalisation in the NLP community. The workshop has two concrete goals: Bring together different expert communities to discuss challenging questions relating to generalisation in NLP; Establish a shared platform for state-of-the-art generalisation testing in NLP. We started this last year, and this years collaborative benchmarking task (CBT) is solely LLM-focused! particular engineering applications.

Time	Session
09:00 09:15	Opening Remarks
09:15 10:00	Keynote 1, by Pascale Fung
10:00 10:30	Oral Presentations <ul style="list-style-type: none">• Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution. <i>Hayley Ross, Kathryn Davidson, Najoung Kim</i>• Investigating the Generalizability of Pretrained Language Models across Multiple Dimensions: A Case Study of NLI and MRC. <i>Ritam Dutt, Sagnik Ray Choudhury, Varun Venkat Rao, Carolyn Rose, V.G. Vinod Vydiswaran</i>
10:30 11:00	Coffee Break
11:00 11:45	Keynote 2, by Najoung Kim
11:45 12:30	Spotlight Presentations: <ul style="list-style-type: none">• MLissard: Multilingual Long and Simple Sequential Reasoning Benchmarks. <i>Mirelle Candida Bueno, Roberto Lotufo, Rodrigo Frassetto Nogueira</i>• OmniDialog: A Multimodal Benchmark for Generalization Across Text, Visual, and Audio Modalities. <i>Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov, Denis Dimitrov</i>• MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models. <i>Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, Sungeun Lee</i>• The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns. <i>Bastian Bunzeck, Sina ZarrieSS</i>

-
- MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models. *Wentian Wang, Sarthak Jain, Paul Kantor, Jacob Feldman, Lazaros Gallos, Hao Wang*

12:30 13:45	Lunch Break
13:45 15:00	Poster Session
15:00 15:45	Keynote 3, by Sameer Singh
15:45 16:00	Coffee Break
16:00 16:30	Panel
16:30 16:45	Closing Remarks and Best Paper Award

W14 - Natural Legal Language Processing (NLLP) Workshop 2024

Organizers:

Nikolaos Aletras, Leslie Barrett, Ilias Chalkidis, Catalina Goanta, Daniel Preotiuc-Pietro, Gerasimos Spanakis

<https://nllpw.org/workshop/>

Room: Hibiscus A

Saturday, November 16, 2024

The Natural Legal Language Processing (NLLP) 2024 workshop, now at its sixth edition, brings together researchers, practitioners, policy makers from around the world who develop NLP techniques within the legal domain. NLP technologies allow legal practitioners and decision-makers to make more informed decisions, optimize legal strategies and serve clients/consumers/citizens in a more cost-efficient way. The fast-paced, multi-jurisdictional world of law is a growing area of application for NLP, offering data sources which are often multilingual and multimodal. For example, evidentiary data sets used in private and public legal practice require in-depth image analysis and speech recognition technologies to complement text data (e.g., opinions and judgments) currently dominating the area. Legal NLP research can create societal impact by informing regulators how to best protect certain categories of citizens at risk (e.g. vulnerable consumers), or by enhancing citizen education and access to justice. This is an exciting opportunity to expand the boundaries of our field by identifying new problems and exploring new data as it interacts with the full inventory of NLP and machine learning approaches.

Time	Session
	Session 1
09:00 - 09:15	Workshop opening
09:15 - 09:20	Summarizing Long Regulatory Documents with a Multi-Step Pipeline. <i>Mika Sie, Ruby Beek, Michiel Bots, Sjaak Brinkkemper, Albert Gatt</i>
09:20 - 09:25	Towards an Automated Pointwise Evaluation Metric for Generated Long-Form Legal Summaries. <i>Shao Min Tan, Quentin Grail, Lee Quartey</i>
09:25 - 09:30	Cross Examine: An Ensemble-based Approach to Leverage Large Language Models for Legal Text Analytics. <i>Saurav Chowdhury, Lipika Dey, Suyog Joshi</i>
09:30 - 09:35	LexSumm and LexT5: Benchmarking and Modeling Legal Summarization Tasks in English. <i>Santosh T.Y.S.S, Cornelius Johannes Weiss, Matthias Grabmair</i>
09:35 - 09:40	Algorithm for Automatic Legislative Text Consolidation. <i>Matias Etcheverry, Thibaud Real-del-sarte, Pauline Chavallard</i>
09:40 - 09:50	Joint Q&A
09:50 - 09:55	LeGen: Complex Information Extraction from Legal Sentences using Generative Models. <i>Chaitra C R, Sankalp Kulkarni, Sai Rama Akash Varma Sagi, Shashank Pandey, Rohit Yalavarthy, Dipanjan Chakraborty, Prajna Devi Upadhyay</i>
09:55 - 10:00	Information Extraction for Planning Court Cases. <i>Drish Mali, Rubash Mali, Claire Barale</i>
10:00 - 10:05	Automated Anonymization of Parole Hearing Transcripts. <i>Abed El Rahman Itani, Wassiliki Siskou, Annette Hautli-Janisz</i>
10:05 - 10:10	BLT: Can Large Language Models Handle Basic Legal Text?. <i>Andrew Blair-Stanek, Nils Holzenberger, Benjamin Van Durme</i>

10:10 - 10:15	Classify First, and Then Extract: Prompt Chaining Technique for Information Extraction. <i>Alice Saebom Kwak, Clayton T. Morrison, Derek Bambauer, Mihai Surdeanu</i>
10:15 - 10:20	HiCuLR: Hierarchical Curriculum Learning for Rhetorical Role Labeling of Legal Documents. <i>Santosh T.Y.S.S, Apolline Isaia, Shiyu Hong, Matthias Grabmair</i>
10:20 - 10:30	Joint Q&A
10:30 - 11:00	Break
	Session 2
11:00 - 11:05	Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models. <i>Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, Arnab Bhattacharya</i>
11:05 - 11:10	The CLC-UKET Dataset: Benchmarking Case Outcome Prediction for the UK Employment Tribunal. <i>Huiyuan Xie, Felix Steffek, Joana Ribeiro de Faria, Christine Carter, Jonathan Rutherford</i>
11:10 - 11:15	Transductive Legal Judgment Prediction Combining BERT Embeddings with Delaunay-Based GNNs. <i>Hugo Attali, Nadi Tomeh</i>
11:15 - 11:20	Comparative Study of Explainability Methods for Legal Outcome Prediction. <i>Ieva Raminta Stalinaite, Josef Valvoda, Ken Satoh</i>
11:20 - 11:25	Incorporating Precedents for Legal Judgement Prediction on European Court of Human Rights Cases. <i>Santosh T.Y.S.S, Mohamed Hesham Elganayni, Stanisaw Sójka, Matthias Grabmair</i>
11:25 - 11:30	The Craft of Selective Prediction: Towards Reliable Case Outcome Classification An Empirical Study on European Court of Human Rights Cases. <i>Santosh T.Y.S.S, Irtiza Chowdhury, Shanshan Xu, Matthias Grabmair</i>
11:30 - 11:45	Joint Q&A
11:45 - 11:50	Quebec Automobile Insurance Question-Answering With Retrieval-Augmented Generation. <i>David Beauchemin, Richard Khoury, Zachary Gagnon</i>
11:50 - 11:55	Attributed Question Answering for Preconditions in the Dutch Law. <i>Felicia Re-delaar, Romy van Drie, Suzan Verberne, Maaike de Boer</i>
11:55 - 12:00	Measuring the Groundedness of Legal Question-Answering Systems. <i>Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Alan Pol, Guglielmo Bonifazi, Shang Gao, Martin Gajek</i>
12:00 - 12:10	Joint Q&A
12:10 - 14:00	Lunch & In-Person Poster Session (Lunch provided)
	Session 3
14:00 - 15:00	Keynote Talk: Omri Ben-Shahar. Privacy Protection, At What Cost?
15:00 - 15:15	Shared Task: Enhancing Legal Violation Identification with LLMs and Deep Learning Techniques: Achievements in the LegalLens 2024 Competition. <i>Ben Hagag, Gil Gil Semo, Dor Bernsohn, Liav Harpaz, Pashootan Vaeziipoor, Rohit Saha, Kyryl Truskovskyi, Gerasimos Spanakis</i>
15:15 - 15:30	Shared Task Winner Presentation
15:30 - 16:00	Break
	Session 4
16:00 - 16:05	LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services. <i>Marco Aspromonte, Andrea Filippo Ferraris, Federico Galli, Giuseppe Contissa</i>

16:05 - 16:10	Enhancing Contract Negotiations with LLM-Based Legal Document Comparison. <i>Savinay Narendra, Kaushal Shetty, Adwait Ratnaparkhi</i>
16:10 - 16:15	Multi-Property Multi-Label Documents Metadata Recommendation Based on Encoder Embeddings. <i>Nasredine Cheniki, Vidas Daudaravicius, Abdelfettah Feiliachi, Didier Hardy, Marc Wilhelm Küster</i>
16:15 - 16:20	CLERC: A Dataset for U. S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation. <i>Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, Benjamin Van Durme</i>
16:20 - 16:25	Empowering Air Travelers: A Chatbot for Canadian Air Passenger Rights. <i>Maksym Taranukhin, Sahithya Ravi, Gabor Lukacs, Evangelos Milios, Vered Schwartz</i>
16:25 - 16:30	The Impact of Formulaic Language in the Court of Justice of the European Union on the Performance of Lexical and Dense Retrieval Methods. <i>Larissa Mori, Carlos Sousa de Oliveira, Yuehwern Yih, Mario Ventresca</i>
16:30 - 16:40	Joint Q&A
16:40 - 16:45	Gaps or Hallucinations? Scrutinizing Machine-Generated Legal Analysis for Fine-Grained Text Evaluations. <i>Abe Bohan Hou, William Juraj, Nils Holzenberger, Andrew Blair-Stanek, Benjamin Van Durme</i>
16:45 - 16:50	How Many Van Goghs Does It Take to Van Gogh? Finding the Imitation Threshold. <i>Sahil Verma, Royi Rassin, Arnav Mohanty Das, Gantavya Bhatt, Preethi Seshadri, Chirag Shah, Jeff Bilmes, Hannaneh Hajishirzi, Yanai Elazar</i>
16:50 - 16:55	Towards Supporting Legal Argumentation with NLP: Is More Data Really All You Need?. <i>Santosh T.Y.S.S, Kevin Ashley, Katie Atkinson, Matthias Grabmair</i>
16:55 - 17:00	Misinformation with Legal Consequences (MisLC): A New Task Towards Harnessing Societal Harm of Misinformation. <i>Chu Fei Luo, Radin Shayanfar, Rohan V Bhamphoria, Samuel Dahan, Xiaodan Zhu</i>
17:00 - 17:10	Joint Q&A
17:10 - 17:15	LAR-ECHR: A New Legal Argument Reasoning Task and Dataset for Cases of the European Court of Human Rights. <i>Odysseas S. Chapanis, Dimitris Galanis, Ion Androutsopoulos</i>
17:15 - 17:20	Developing a Pragmatic Benchmark for Assessing Korean Legal Language Understanding in Large Language Models. <i>Kimyeon Choi, Youngrok Eunkyoung Choi, JinHwan Choi, Hai Jin Park, Wonseok Hwang</i>
17:20 - 17:25	Enhancing Legal Expertise in Large Language Models through Composite Model Integration: The Development and Evaluation of Law-Neo. <i>Zhihao Liu, Yanzhen Zhu, Mengyuan Lu</i>
17:25 - 17:30	Joint Q&A
17:30 - 17:40	Best Presentation Award

W15 - The 4th Workshop on Multilingual Representation Learning

Organizers:

David Ifeoluwa Adelani, Duygu Ataman, Mammad Hajili, Raghav Mantri, Abraham Owodunni, Jonne Saleva, David Strap, Francesco Tinner

<https://sigtyp.github.io/ws2024-mrl.html>

Room: Jasmine

Saturday, November 16, 2024

Multi-lingual representation learning methods have recently been found to be extremely efficient in learning features useful for transfer learning between languages and demonstrating potential in achieving successful adaptation of natural language processing (NLP) models into languages or tasks with little to no training resources. On the other hand, there are many aspects of such models which have the potential for further development and analysis in order to prove their applicability in various contexts. These contexts include different NLP tasks and also understudied language families, which face important obstacles in achieving practical advances that could improve the state-of-the-art in NLP of various low-resource or underrepresented languages.

Time	Session
09:00 09:10	Opening Remarks
09:10 09:50	Invited Talk by Karen Livescu
09:50 10:30	Invited Talk by Hila Gonen
10:30 11:00	Coffee Break
11:00 12:30	Poster Session
12:30 14:00	Lunch Break
14:00 14:30	Shared Task Session: <ul style="list-style-type: none">• Findings Paper• Winning Team Presentation
14:30 15:30	Best Paper Session: <ul style="list-style-type: none">• Best Paper• Honorable Mentions
15:30 16:00	Coffee Break
16:00 16:50	Invited Talk by Sebastian Ruder
16:50 17:00	Closing Remarks

W16 - NLP4Science: The First Workshop on Natural Language Processing for Science

Organizers:

Nitay Calderon, Alex Chapanin, Rotem Dror, Amir Feder, Ariel Goldstein, Anna Korhonen, Shir Lissak, Yaakov Ophir, Lotem Peled-Cohen, Roi Reichart, Ilanit Sobol, Rafael Tikochinski, Mor Ventura

<https://sites.google.com/view/nlp4science/home>

Room: Hibiscus B

Saturday, November 16, 2024

The NLP4Science workshop ventures into an important new frontier: leveraging NLP to better understand the human mind. Researchers are increasingly using NLP and LLMs in particular for the scientific modeling and understanding of human behavior. They apply NLP tools to gain invaluable insights into social science, psychology, psychiatry, health, neuroscience, behavioral economics, and beyond. In this workshop we will cover principles of NLP-driven scientific modeling, advanced methods for statistically robust evaluation of NLP models, experimental design, causal inference, and causality-based methods for text models in science.

Time	Session
08:45 09:00	Gathering and Welcome
09:00 09:45	Invited Speaker - Amit Sharma
09:45 10:30	Invited Speaker - Rita Goldstein
10:30 11:00	Coffee Break
11:00 12:00	Panel Discussion and Q&A
12:00 13:00	Lunch Break
13:00 13:45	Invited Speaker - Roger Levy
13:45 14:30	Invited Speaker - Hadas Raviv
14:30 16:00	Poster Session + Coffee Break
16:00 16:15	Best Paper Announcement + Short Oral
16:15 16:45	Invited Speaker - Nitay Calderon
16:45 17:00	Closing Remarks

W17 - The Second Workshop on Social Influence in Conversations (SICon 2024)

Organizers:

Muskan Garg, Kushal Chawla, Weiyan Shi, Ritam Dutt, Deuksin Brian Kwon,
James Hale, Daniel Hershcovich, Aina Gari Soler, Liang Qiu, Alexandros
Papangelis, Zhou Yu, Gale Lucas

<https://sites.google.com/view/sicon2024/home>

Room: Pearson

Saturday, November 16, 2024

Social influence (SI) is the change in an individual's thoughts, feelings, attitudes, or behaviors from interacting with another individual or a group. For example, a buyer uses SI skills to negotiate trade-offs and build rapport with the seller. SI is ubiquitous in everyday life, and hence, realistic human-machine conversations must reflect these dynamics, making it essential to model and understand SI in dialogue research systematically. This would improve SI systems' ability to understand users' utterances, tailor communication strategies, personalize responses, and actively lead conversations. These challenges draw on perspectives not only from NLP and AI research but also from Game Theory, Affective Computing, Communication, and Social Psychology. SI dialogue tasks like negotiation, persuasion, therapy, and argumentation have recently gained traction. Current conversational systems emphasize modeling system strategies using dialogue acts and strategy annotations or modeling users. Prior work also explored related tasks crucial for the eventual development of SI systems, namely outcome prediction, argument mining, and lie detection. However, these efforts are scattered, and only limited efforts focus on building useful systems exhibiting SI skills, such as chatbots. Ensuring AI-driven models safety, interpretability, and integration into real-time applications that simulate or analyze SI remains challenging.

Time	Session
09:00–09:10	Opening Remarks
09:10–09:40	Invited Talk: David Jurgens
09:40–10:10	Invited Talk: Kyriaki Kalimeri
10:10–10:40	Invited Talk: Maurice Schweitzer
10:40–11:00	Coffee Break
11:00–12:00	Panel Discussion 1
12:00–13:30	Lunch Break
13:30–14:30	Poster Session 1
14:30–15:30	Poster Session 2
15:30–16:00	Invited Talk: Viktoria Spaiser
16:00–16:30	Invited Talk: Yulia Tsvetkov
16:30–16:45	Lightning Talks
16:45–17:00	Coffee Break
17:00–17:30	Invited Talk: Yi-Chia Wang
17:30–18:00	Invited Talk: Henning Wachsmuth
18:00	Closing Remarks

W18 - The 11th Workshop on Asian Translation (WAT2024)

Organizers:

Toshiaki Nakazawa, Isao Goto, Hidaya Mino, Kazutaka Kinugawa, Haiyue Song, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Ondej Bojar, Sadao Kurohashi, Pushpak Bhattacharyya

<https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2024/index.html>

Room: Miami Lecture Hall
Saturday, November 16, 2024

Many Asian countries are rapidly growing these days and the importance of communicating and exchanging the information with these countries has intensified. To satisfy the demand for communication among these countries, machine translation technology is essential. Machine translation technology has rapidly evolved recently and it is seeing practical use especially between European languages. However, the translation quality of Asian languages is not that high compared to that of European languages, and machine translation technology for these languages has not reached a stage of proliferation yet. This is not only due to the lack of the language resources for Asian languages but also due to the lack of techniques to correctly transfer the meaning of sentences from/to Asian languages. Consequently, a place for gathering and sharing the resources and knowledge about Asian language translation is necessary to enhance machine translation research for Asian languages. The Workshop on Machine Translation (WMT), the world's largest machine translation workshop, mainly targets on European languages and does not include Asian languages. The International Workshop on Spoken Language Translation (IWSLT) has spoken language translation tasks for some Asian languages using TED talk data, but these is no task for written language. The Workshop on Asian Translation (WAT) is an open machine translation evaluation campaign focusing on Asian languages. WAT gathers and shares the resources and knowledge of Asian language translation to understand the problems to be solved for the practical use of machine translation technologies among all Asian countries. WAT is unique in that it is an "open innovation platform": the test data is fixed and open, so participants can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so participants can submit translation results at any time.

Time	Session
09:00 09:05	Welcome (Toshiaki Nakazawa)
09:05 10:30	Panel Discussion (Chair: Toshiaki Nakazawa) <i>Machine Translation of Asian Languages in the LLM Era</i> <i>Min Zhang, Thepchai Supnithi, Kozo Moriguchi, Fred Bane</i>
10:30 11:00	Break
11:00 12:40	Research Paper Presentations (20 mins. each) <ul style="list-style-type: none">• Machine Translation Of Marathi Dialects: A Case Study Of Kadodi <i>Raj Dabre, Mary Noel Dabre, Teresa Pereira</i>• AI-Tutor: Interactive Learning of Ancient Knowledge from Low-Resource Languages <i>Siddhartha R. Dalal, Rahul Aditya, Vethavikashini Chithrra Raghu-ram, Prahlad Koratamaddi</i>• An Empirical Study of Multilingual Vocabulary for Neural Machine Translation Models <i>Kenji Imamura, Masao Utiyama</i>

- Are Large Language Models State-of-the-art Quality Estimators for Machine Translation of User-generated Content? *Shenbin Qian, Constantin Orasan, Diptesh Kanodia, Félix do Carmo*

- Creative and Context-Aware Translation of East Asian Idioms with GPT-4 *Kenan Tang, Peiyang Song, Yao Qin, Xifeng Yan*

12:40 12:45**Closing Remarks** (Toshiaki Nakazawa)

W19 - The First Workshop on Advancing Natural Language Processing for Wikipedia (NLP for Wikipedia)

Organizers:

Lucie-Aimée Kaffee, Isaac Johnson, Angela Fan, Tajuddeen Gwadabe, Fabio Petroni, Daniel van Strien

[https://meta.wikimedia.org/wiki/NLP_for_Wikipedia_\(EMNLP_2024\)](https://meta.wikimedia.org/wiki/NLP_for_Wikipedia_(EMNLP_2024))

Room: Johnson

Saturday, November 16, 2024

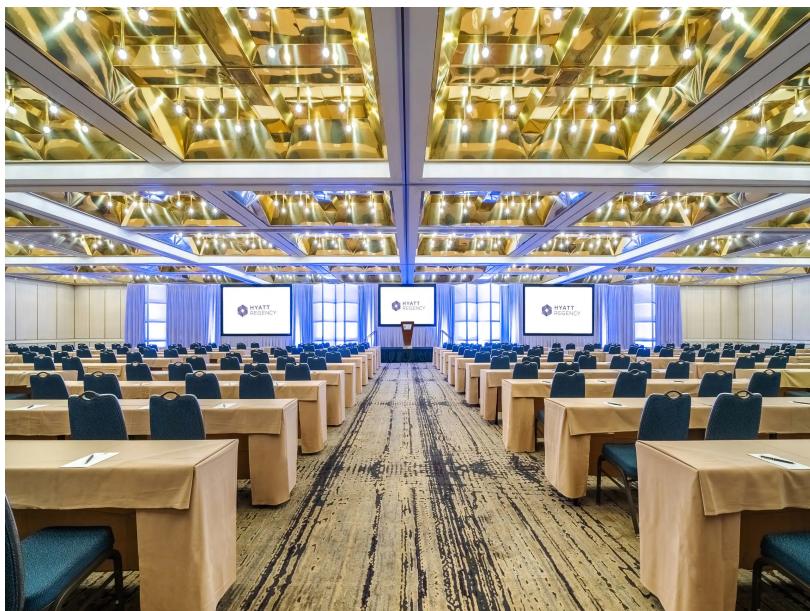
A space both to celebrate Wikimedia's contributions to the NLP community and highlight approaches to ensuring the sustainability of this relationship for years to come. Wikipedia is a uniquely important resource for the NLP community; it is multilingual, can be freely reused under its open license, and is edited and maintained by a dedicated community of editors who have earned its status as a very high-quality dataset for many applications. With this value comes many tensions however: Despite Wikipedia's presence in over 300 language editions, much focus in language modeling remains on the high-resource languages; Despite the openness of Wikipedia and its role in many advances in natural language modeling, there are concerns that some of these advances such as generative text models could undermine Wikipedia and threaten its sustainability as a community and ultimately data resource; Despite the heavy usage of Wikimedia data among the NLP community, few researchers work on developing tools that can contribute back to the Wikimedia community. We will invite researchers to contribute novel uses of Wikimedia data or studies of the impact of Wikimedia data within the NLP community. We will also discuss successful approaches to developing tooling that can assist the Wikimedia community in maintaining and improving the breadth of the Wikimedia projects.

Time	Session
09:00 - 09:05	Opening remarks - Isaac/Lucie introducing the workshop
09:05 - 09:50	Keynote by Jess Wade - 30-min talk (remote) + Joint Q&A
09:50 - 10:30	2-minute paper lightning talks - Pre-recorded
10:30 - 11:00	Coffee break
11:00 - 12:00	Poster session - In-person or virtual on Gather
12:00 - 12:45	Lunch
12:45 - 13:30	Keynote by Scott A. Hale - 30-min talk (remote) + Joint Q&A
13:30 - 14:15	Panel: Misinformation + Wikipedia - Isabelle Augenstein, Andreas Vlachos
14:15 - 15:00	Panel: Impact of LLMs on Wikipedia - Ilyas Lebleu, David Adelani
15:00 - 15:30	Closing (30-minutes) - Statements by Leila Zia

16

Local Guide

Conference Venue



Venue: **Hyatt Regency Miami** - 400 South East Second Ave - Miami, FL 33131

<https://www.hyatt.com/>

Phone: (305) 358-1234

This year venue for **EMNLP 2024** is the **Hyatt Regency Miami**, located at 400 SE 2nd Ave, Miami, FL 33131, USA. Our Downtown Miami Hotel is next to Brickell, one of the trendiest neighborhoods in Miami. Our hotel overlooks the Miami River and our ideal downtown location puts you steps from the Miami Riverwalk and Bayfront Park. The hotel is also close to the Port of Miami and the Kaseya Center (formerly FTX Arena). For a day of shopping, our hotel is near Brickell City Centre. Or, explore Little Havana and tour the Phillip & Patricia Frost Museum of Science. Rooms and suites at our downtown Miami hotel offer stunning city and Biscayne Bay views with all the comforts you expect from a luxury, urban destination. Hyatt Regency Miamis 615 rooms and suites are outfitted with beautiful wood furnishings, generous wardrobe space and functional work areas.

Hotel Parking

Self-parking: \$53.00

Valet parking: Standard vehicles \$53.00; Oversize vehicles: \$78.00

General Parking Info

- Self Parking is also available in the downtown area at Miami Tower, 100 SE 2nd Street - 20.75
- 225 SE. 2nd St. Garage: Located 0.1 mi away and costs \$15
- 200 SE. 2nd St. Garage - P2531: Located 0.1 mi away and costs \$25
- 250 SE. 3rd Ave. Garage - P2530: Located 0.1 mi away and costs \$25
- 200 S. Biscayne Blvd. SE Financial Center Garage: Located 0.2 mi away and costs \$20
- 81 SE. 5th St. 444 Brickell Ave. Garage: Located 0.3 mi away and costs \$20
- 29 NW. 1st St. Cindy Lot: Located 0.3 mi away and costs \$5

Market

In need of a late-night treat or early morning breakfast? Stop by the Market on the Lobby Level. The Markets inviting, airy feel makes it easy to dash in and grab a quick bite to go. Choose from a steady rotation of bakery items, pizzas, sandwiches, and healthy choices, as well as a full specialty coffee menu. Choose from a steady rotation of bakery items, pizzas, sandwiches, and fresh options plus a full specialty coffee menu.

Hours: Mon Sun 5:00 AM - 2:00 AM

Need it To Go

Enjoy a restaurant experience in the comfort of your room. The Market offers freshly prepared food in eco-friendly containers for pickup or in-room delivery. Order from your in-room phone or use our mobile app.

Hours: Mon Sun 7:00 AM - 10:30 PM

Riverview Bar & Grill

Located on the Lobby Level, Riverview Bar & Grill offers a wide variety of menu items for breakfast,

lunch, and dinner. Enjoy views of the Miami River for breakfast or lunch. Stop by for a signature mojito as you take in the latest sporting event and explore our lounge menu or full à la carte dinner offerings for the perfect bite.

Breakfast

Mon Sun 7:00 AM - 11:00 AM

Lunch

Mon Sun 12:00 PM - 4:00 PM

Dinner

Mon Thu 4:00 PM - 11:00 PM

Fri & Sat 4:00 PM - 12:00 AM

Sun 4:00 PM - 11:00 PM

Amenities

- **Fitness Center:** Daily 24-hour access to our StayFit fitness center.
- **Pool:** The pool is open daily from dawn to dusk.
- **Valet Parking:** From \$55
- **Business Services**
- **Concierge**
- **Digital Check-In**
- **Laundry**

Outlets

For USA there are two associated plug types, types A and B. Plug type A is the plug which has two flat parallel pins and plug type B is the plug which has two flat parallel pins and a grounding pin. USA operates on a 120V supply voltage and 60Hz.

Accessibility

We are committed to providing equal access and opportunity for individuals with disabilities. The features also make this hotel more accessible for older individuals with changing abilities to ensure a seamless experience. Our overall goal is to improve usability throughout the hotel for all guests.

About Miami

Explore Miami: <https://www.miamiandbeaches.com/>

Image credits: The Official Website of Greater Miami & Miami Beach

Miami, Florida is a renowned tourist destination located in the southeastern United States. Known for its stunning coastal beauty and vibrant cultural diversity, the city offers visitors an eclectic mix of art, history, and outdoor activities. Miami's year-round tropical climate makes it an ideal location for both relaxation and adventure. Below are some key highlights of the city:



Beaches and Outdoor Activities

Miami's beaches, particularly the famous South Beach, are a primary draw for visitors. South Beach offers white sands, turquoise waters, and a lively atmosphere, ideal for sunbathing, swimming, and water sports. Miami's location also provides access to nature reserves like the Everglades National Park, where tourists can explore unique ecosystems, take airboat tours, and observe wildlife such as alligators and exotic birds. The Biscayne Bay area allows for snorkeling, boating, and paddleboarding.

Cultural Attractions

Miami is celebrated for its multicultural flair, heavily influenced by Latin American, Caribbean, and European cultures. Visitors can explore Little Havana, the heart of Miami's Cuban culture, where they can enjoy authentic Cuban cuisine, music, and art. The Wynwood Arts District is another cultural hotspot, famous for its street art and galleries. Art enthusiasts can visit the Pérez Art Museum Miami (PAMM) and the Art Deco Historic District for architectural inspiration.

Dining and Nightlife

Miami's food scene is a blend of Latin, Caribbean, and international influences, reflecting the city's rich cultural diversity. Visitors can enjoy classic Cuban dishes like ropa vieja and Cuban sandwiches in Little Havana, while South Beach offers an array of upscale seafood restaurants. Miami is also famous for tropical fruits like mangos and passion fruit, which can be found in smoothies, desserts, and local markets. Nightlife in Miami is vibrant, with world-renowned clubs, rooftop bars, and live music venues, especially in districts like Wynwood, South Beach, and Brickell.

Shopping and Entertainment

Shopping enthusiasts will enjoy Miami's luxury retail destinations, such as Brickell City Centre, Bal Harbour Shops, and Bayside Marketplace, which combine high-end stores with dining and entertainment. Miami's thriving entertainment scene includes festivals, concerts, and sporting events. The annual Art Basel Miami Beach, a premier art show, attracts international artists and art lovers alike.

Miamis blend of sun-soaked beaches, rich culture, diverse dining options, and vibrant nightlife makes it a top destination for tourists from around the world. Whether visitors seek adventure, relaxation, or cultural enrichment, Miami offers an array of experiences that cater to all interests. For more details, visit the official tourism website at: <https://www.miamilandbeaches.com/>

Things to do in Miami

Miami, Florida, is bursting with activities and attractions that cater to all interests.

Explore the Beaches

Relax at iconic beaches like South Beach and Key Biscayne, known for their beautiful sands and vibrant atmosphere. Engage in water sports, sunbathing, and beach volleyball.

Visit Cultural Neighborhoods

Explore Little Havana to experience Cuban culture, cuisine, and art. The Wynwood Arts District showcases colorful murals and galleries, perfect for art enthusiasts.

Discover Nature

Take a trip to the Everglades National Park for unique wildlife encounters, airboat tours, and hiking trails. Enjoy outdoor activities such as kayaking in Biscayne Bay.

Experience Nightlife

Miamis nightlife is legendary, with numerous nightclubs, bars, and live music venues. Popular areas include South Beach, Wynwood, and Brickell, offering a range of entertainment options.

Shop and Dine

Shop at luxury destinations like Brickell City Centre and Bal Harbour Shops. Savor diverse cuisine from around the world, including fresh seafood and local delicacies.

Attend Events

Miami hosts various events throughout the year, including Art Basel, food festivals, and concerts, making it an exciting destination for visitors.

Shopping, Dining, & Entertainment

- Brickell City Centre
- Mary Brickell Village
- Bayside Marketplace
- James L. Knight Concert Venue
- Calle Ocho
- South Beach

Museums & Parks

- History Miami Museum
- Frost Science Museum
- Pérez Art Museum Miami
- Miami Seaquarium
- Jungle Island
- Historic Virginia Key Beach Park
- Cape Florida Lighthouse
- Fairchild Tropical Botanical Garden
- Biscayne Bay Park Biscayne National Park
- Island Queen Cruises & Tours

Sports & Performing Arts

- Adrienne Arsht Center
- Kaseya Center (formerly known as FTX Arena)
- LoanDepot Park (formerly known as Marlins Park)

Beaches

Beach Guide [here](#).

Budget Friendly

For free things to do in Miami, click [here](#).

Useful Information

The currency in Miami, FL is the U.S. Dollar. ATMs are readily accessible, and credit/debit cards as well as Apple Pay and Google Pay are widely accepted.

Tipping

- Bartending: \$1-\$2 per drink
- Restaurant: 15-20%
- Bellhop: \$1-\$3 per bag
- Housekeeper: \$2-\$3 per night
- Taxis/Rideshare: 15-20%
- Shuttle Driver: \$1-\$2 per person

Weather

During November, Miamis weather is still very summer-like with temperatures in the mid to low 80s F (about 30C). Skies have a tendency to be clear or partly sunny; with occasional rainfall. Summer attire works well most days.

Power Outlets and Adapters

The standard voltage in the United States is 120 V and the standard frequency is 60 Hz. The plug has two flat parallel pins.

Local Customs

Drinking

The federal legal age for buying and drinking alcohol is 21 years old. Its illegal to drink in public spaces, including the beach.

Beach Etiquette

Locals take the beach seriously, so avoid making faux pas by respecting peoples space on the beach.

Beach Safety

Dont go to a beach that is displaying a purple flag. This indicates that jellyfish, stingrays, or other dangerous critters are in the water.

Start Late

Miami is a late-night city with much nightlife not in full swing until after midnight. Take your time: stretch out dinner and enjoy some cocktails before heading out for the night.

Respect the Doormen

Dress to impress and be polite, not pushy, when waiting in line to get inside a club. Note that clubs rarely admit large groups of men (with the exception of gay clubs).

Language

Spanish is used in day-to-day life in Miami. It pays to learn at least a few words. Here are a few!

- Hola - Hello
- Por favor - Please
- Gracias - Thank you
- De nada - Youre welcome
- Sí - Yes
- No - No

Walking

Walk to the right of the sidewalk and step off to the side of the sidewalk if you want to stop to check your phone, look up directions, or want to take in a view.

Driving

Americans drive on the right-hand side of the road. Traffic laws can vary between states, so its worth finding out about any local differences if you plan on driving. It is legal to turn right on a red light if its safe to do so unless there are signs stating otherwise.

Public Transport

Allow others to disembark before boarding, dont take up more than one seat, and stand to offer seating to pregnant women or someone with a disability.

Get Used to Supersizing

Many travelers comment that the portion sizes in the U.S. are larger than they are used to back home. Dont worry; it is not considered impolite to leave a meal unfinished, and often you can ask for your leftover food to be boxed up to go.

Spitting

Spitting is considered rude in any public setting. Find more information about local customs and etiquette in the United States generally [here](#).

Food Options

Hyatt Regency Dining

Click [here](#).

- Market to Go
- Market
- Riverview Bar & Grill

Restaurants near Hyatt Regency Miami

Various options at Brickell City Centre. Click [here](#).
Plus tons more...

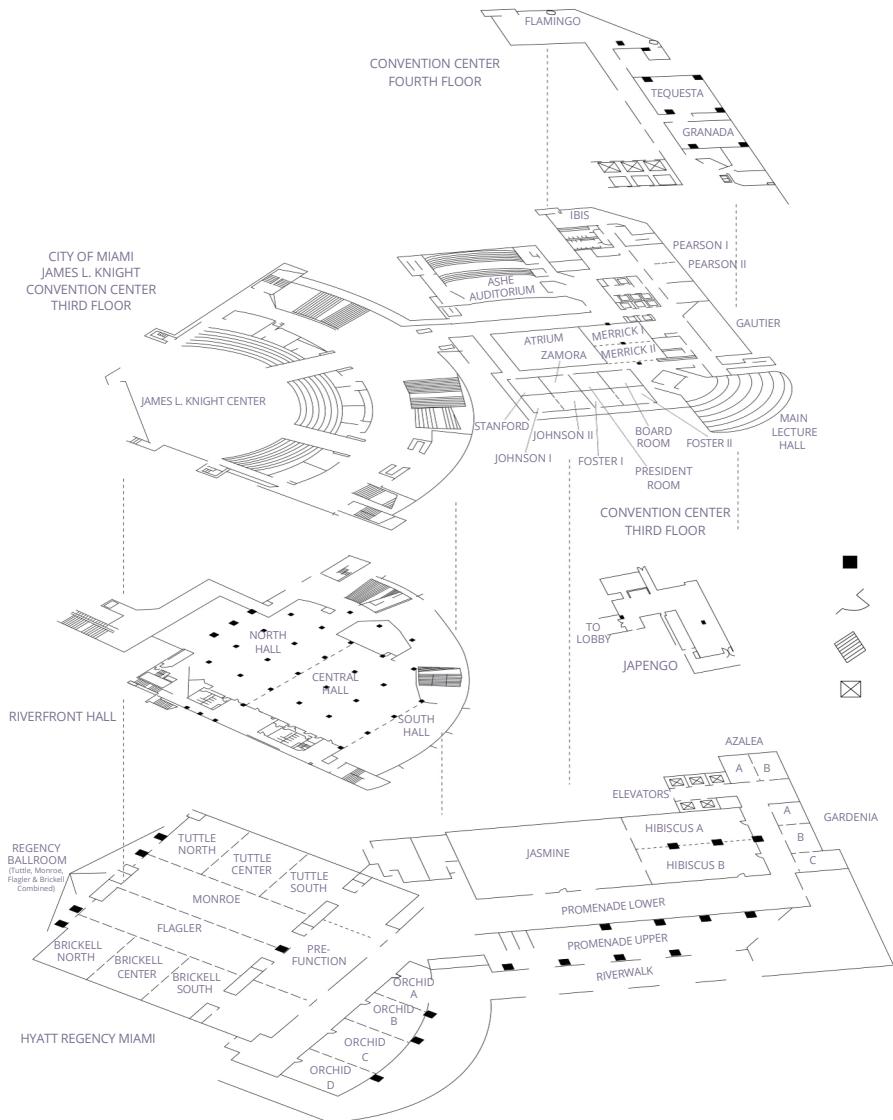
- Friends Market & Bistro
- Moxies
- Crazy About You
- Coyo Taco
- Novecento

- Hibachi Grill & Noodle Bar
- Bubba Gump Shrimp Co
- Bali Cafe
- Tacology

17

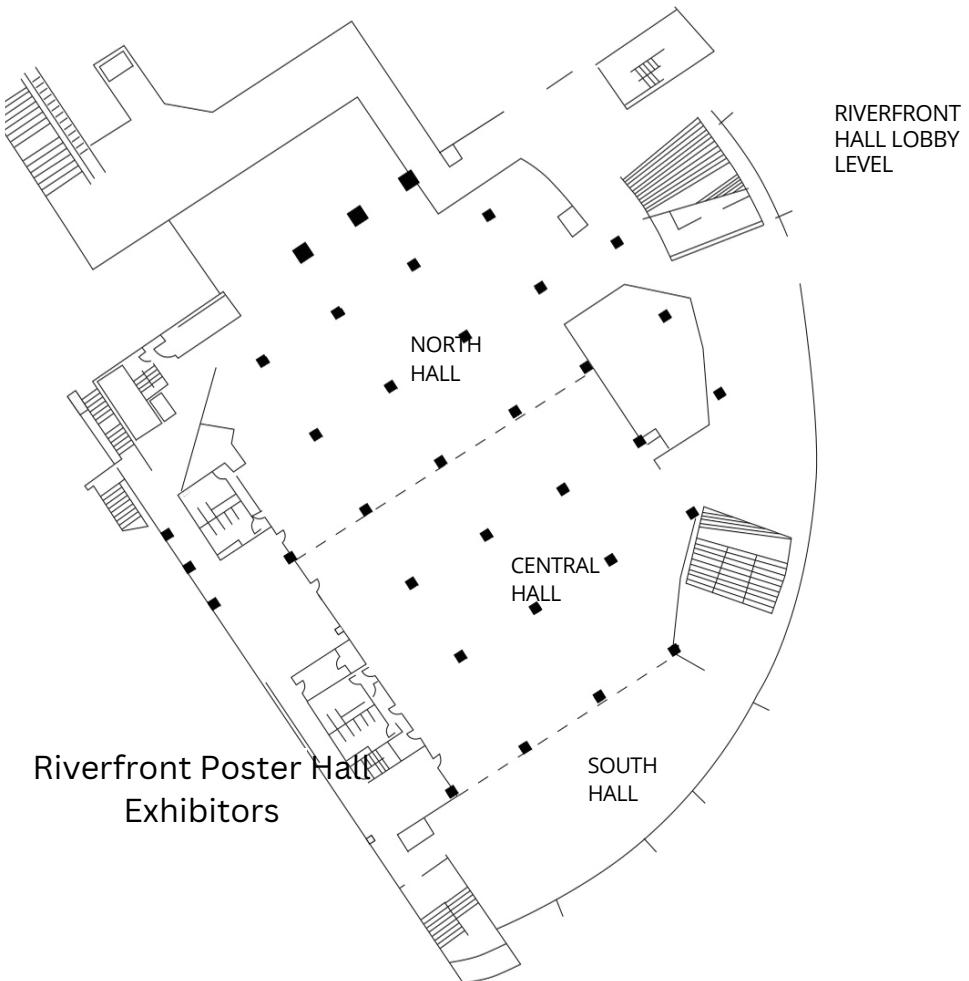
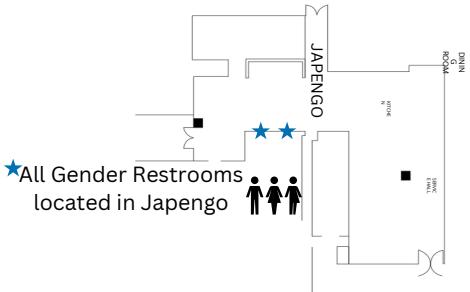
Venue Map

VENUE MAP

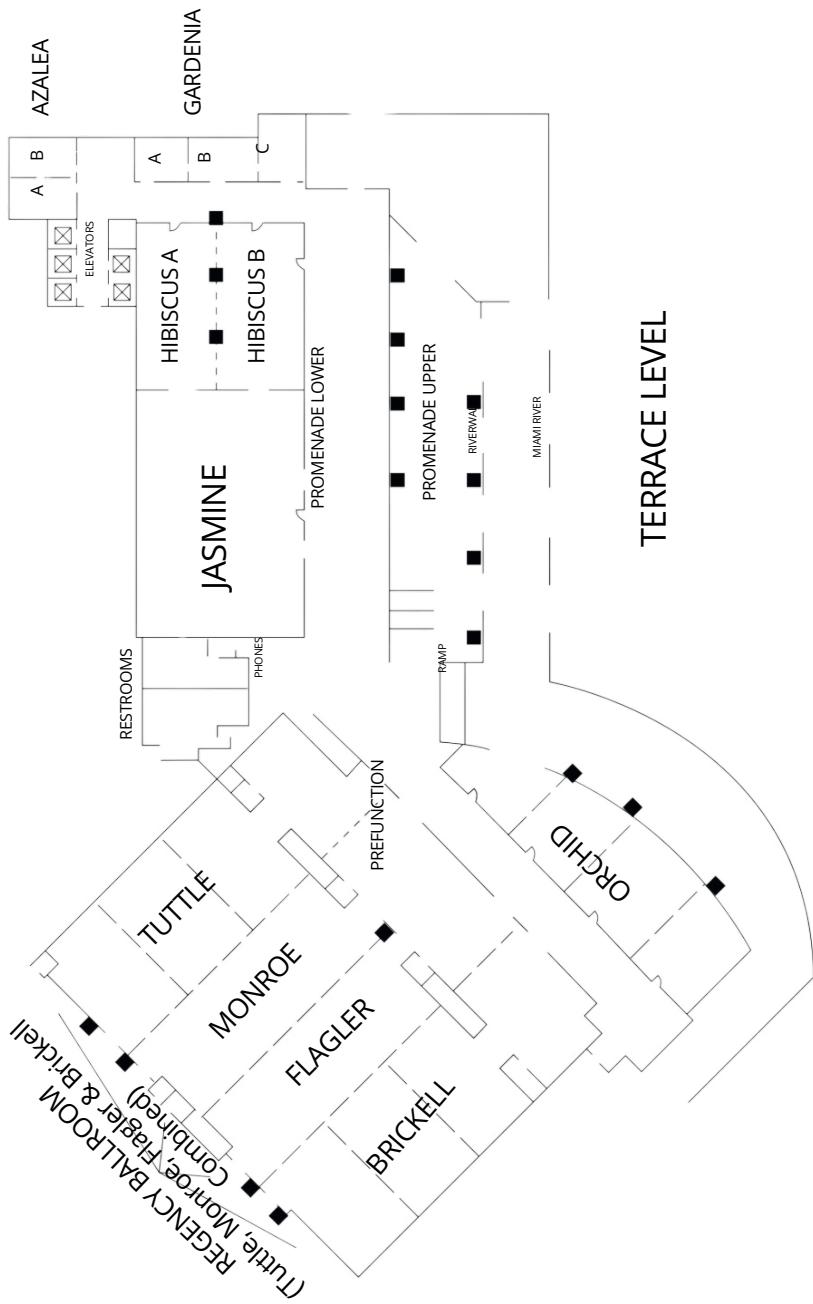


Lobby Level

Hyatt Regency
Check-in



Lower Terrace Level



Author Index

- Aakanksha, 143
Abagyan, 142
Abassy, 82
Abbasi, 361
Abdel-Salam, 287
Abdelaziz, 57
Abdelfattah, 290
Abdelmoneim, 253
Abdul-Mageed, 206, 221, 287
Abdulhai, 152, 397
Abdullah, 133
Abe, 140
Abend, 193, 269, 366
Abercrombie, 261
Abilov, 171
Abolade, 142
AbouElhamayed, 290
Abu-Ghazaleh, 105
Abzaliev, 145
Accomazzi, 307
Acharya, 184
Ackerman, 389
Aco, 142
Acquaye, 295
Acuna, 245
Adams, 47
Adauto, 174
Adel, 240
Adelani, 142, 146
Adhikari, 82
Adhistha, 169
Adi, 129
Adilazuarda, 142, 148, 151
Aditya, 88, 205, 248
Adlaon, 371
Adler, 123, 260
Adorno, 301
Aerts, 217
Afshar, 322
Aftar, 406
AFZAL, 82, 171
Agafonova, 397
Agarwal, 57, 63, 168, 266, 279, 302
Agerri, 174, 285
Aggarwal, 56, 296
Aghakhani, 150
Agostinelli, 155
Agostini, 161
Agrawal, 54, 70, 91, 138, 147, 155, 158, 191, 243, 314, 354, 408
Agro, 180
Agussurja, 278
Aharoni, 40, 126
Ahia, 142, 159
Ahmad, 74, 169, 282, 298
Ahmadian, 75, 143, 276

- Ahmed, 69, 82, 106, 279, 321
Ahn, 310
Ai, 85, 110, 216, 220, 262, 312, 325
Aizawa, 365, 401
Aji, 82, 113, 142, 145, 148, 151
Ajith, 94, 273
Akasaki, 353
Akash, 235
Akbar, 142, 193
Akbar-Tajari, 409
Akvik, 198, 214
Akbulut, 259
Akella, 178, 179, 352
Akhauri, 290
Akhtar, 47
Akkerman, 390
Akpinar, 165
Akter, 188
Akula, 304
Al-Dhabyani, 221
Al-Shaibani, 74
Alacam, 172, 173
Alakraa, 78
Alam, 169, 286, 298
Alamdari, 238
Alameddine, 61
Alastruey, 220
Alberti, 70
Aldarmaki, 170, 180, 291
Alemayehu, 156
Alemzadeh, 110
Alessa, 189
Aletras, 145, 256, 396
Alexandrov, 147
Alfonso-Hermelo, 285
Alfter, 213, 340
Algayres, 77
Alghisi, 216
Alhama, 390
Alhamouri, 221
Alhzami, 79
Ali, 145, 170, 213
Aliannejadi, 160
Alikhani, 258
Alishahi, 256
Alistarh, 168, 368
Alizadeh, 286
AlKhamissi, 398
ALLAITH, 294
Allan, 53, 94
Allauzen, 179
Allen, 64, 307
Allison, 215
Almasian, 234
almohaimeed, 123
Aloisi, 113
Alper, 246
Alqahtani, 73
AlQuabeh, 171
Alraeesi, 221
Alsayadi, 221
Alshomary, 176
Althoff, 286, 288, 337
Alvarez, 193
Alwajih, 221, 287
Aly, 209
Alzubi, 87
Amayuelas, 130
Amershi, 330
Amin, 239
Amini, 97
Amir, 125
Amiri, 260, 267
Amos, 154
Amosy, 318
Amouyal, 212
Amplayo, 312
An, 112, 119, 127, 177, 242, 295, 367, 373, 399
Anand, 165, 338
Ananiadou, 66, 190
Anastasopoulos, 133, 156, 188, 202, 261
Anderson, 62
Andreas, 138, 150
Andrews, 122, 193, 258
Angelov, 235
Anh, 310
Anikina, 162
Anish, 229
Ans dell, 307
Antoine, 129
Anton, 397
Antoniades, 130
Antoniak, 71
Antypas, 145
Anumanchipalli, 223, 270
Anwer, 81, 327
Ao, 281
Aoki, 364
Ao yama, 60, 97, 116
Apidianaki, 66
Appalaraju, 378
Arabzadeh, 282
Aradelli, 324
Araujo, 66, 272
Arbelle, 131

- Arefiyan, 231
 Aremu, 142
 Arias, 267
 Arif, 284
 Arik, 93
 Ariu, 140
 Arkin, 40
 Arnab, 316
 Arneja, 90
 Arnett, 141, 275
 Arora, 171, 207, 294, 299, 356
 Arppe, 147
 Artemova, 82, 387
 Artzi, 104
 Arulraj, 375
 Arya, 305
 Asai, 213
 Asfour, 166
 Asgari, 407
 Ash, 71, 321
 Ashby, 170
 Asher, 117, 271, 301
 Ashfaq, 135
 Ashkboos, 368
 Ashton, 313
 Asif, 105
 Asperti, 163
 Assi, 221
 Asthana, 46
 Aswani, 240
 Atanasova, 273
 Atari, 294
 Athar, 284
 Athiwaratkun, 264
 Atkinson, 127
 Attanasio, 59, 259
 Attari, 270
 Attia, 233
 Atwell, 258
 Augenstein, 60, 72, 171, 273, 299
 August, 212, 215
 Aumiller, 286, 293
 Aussavavirojekul, 388
 Averbuch-Elor, 246
 Averchenkova, 397
 Avestimehr, 305, 308
 Avram, 288
 Awadallah, 282
 Awal, 54
 Ayan, 100
 Aynetdinov, 214
 Ayyubi, 316
 Azar, 276
 Azeemi, 284
 Azime, 340
 Aziz, 82
 Azizi, 200
 Azizov, 171
 ASSenmacher, 267
 Ba, 131, 194
 Babkin, 232
 Babl, 312
 Bach, 88, 103, 228
 Bachem, 279
 Backes, 121, 188, 224
 Bae, 207, 263, 276, 301
 Baek, 143, 269
 Baff, 176
 Bafna, 144, 217
 Bagavan, 232
 Bai, 42, 106, 135, 172, 217, 229, 234, 244,
 258, 275, 283, 293, 320, 334, 335,
 339, 350, 361, 388
 Baig, 63
 Baik, 327
 Bailleux, 254
 Bajaj, 262
 Bajpai, 81, 200, 213
 Bak, 221
 Baker, 152
 Balachandran, 159
 Balashankar, 131
 Balasubramanian, 65
 Balawender, 303
 Balch, 169
 Balde, 135
 Baldridge, 100
 Baldwin, 327, 329
 Balepur, 86, 249, 288
 Balestrucci, 266
 Bali, 151
 Balivada, 322
 Balloccu, 216
 Ballout, 253
 bambroo, 195
 Bandari, 180
 Bandyopadhyay, 215, 249, 267
 Banerjee, 258, 357, 365
 Bang, 301
 Banijamali, 177
 Bannulmath, 394
 Bansal, 49, 88, 168, 217, 254, 316, 330
 Bao, 139, 185, 192, 202, 309, 347, 353, 360,
 381
 Bar, 172

- Baral, 78, 282
Baraniuk, 241, 287, 325
Barannikov, 380
Barba, 156
Barbieri, 145
Barez, 192, 194
Bari, 73
Barik, 229
Bari, 52
Barrett, 279
Barrie, 84
Barriere, 258
Barrow, 400
Barrón-Cedeño, 296
Barsoum, 357
Bartie, 317
Bartolo, 80
Bartoszcze, 277
Barua, 142, 201
Baryshnikov, 354
BASAIC, 67
Basaldella, 268
Basharat, 110
Bashivan, 319
Basile, 266
Basirat, 395
Baskaran, 270
Bassani, 169
Bastani, 136
Bastings, 398
Basu, 57, 101, 304, 315
Bateni, 236
Bathaie, 399
Batista, 171
Batista-Navarro, 403
Battash, 402
Bau, 127
Bauer, 261
Baumann, 177
Baumgärtner, 312
Bauwens, 66
Bawden, 352
Bayat, 263
Bayazit, 125
Bazaga, 405
Bazhukov, 387
Beauchamp, 313
Beaver, 297
Bechet, 129
Beepath, 262
Beese, 48
Behar, 152
Behzad, 116, 326
Beigi, 51, 130
Beigl, 166
Beinborn, 236
Beirami, 131
Bel, 199
Belanec, 292
Belay, 340
Belinkov, 65, 178
Belongie, 299
Belém, 191
Ben-David, 190, 318
Ben-Kish, 246
Bendersky, 230, 332
Benedetto, 292, 324
Benedict, 399
Beniwal, 171, 219, 257
Benkirane, 146
Benson, 170
Bentham, 124
Bentivogli, 123, 167
Berant, 41, 70, 212
Berard, 143
Bercean, 75
Berchansky, 385
Berg-Kirkpatrick, 201
Bergamaschi, 406
Bergen, 57, 141
Bergman, 259
Berlot-Attwell, 314
Bernard, 358
Bernardi, 98, 228
Berrachedi, 221
Berrada, 221
Bertasius, 316
Bertolazzi, 98
Bertran, 198
Berzak, 55
Bespalov, 140
Betala, 78
Bethge, 270
Bhagwatkar, 319
Bhambhoria, 326
BHAN, 76
Bharadwaj, 305
bharadwaj, 344
Bhardwaj, 97, 344
Bhargav, 57
Bhatia, 149, 241, 258, 277
Bhatt, 299
Bhattacharjee, 51, 110, 307
Bhattacharya, 338
Bhattacharyya, 153, 157, 195, 198, 263
Bhosale, 91

- Bhuiya, 248
Bhuiyan, 73, 201
Bi, 333, 355, 363, 388
Bibi, 261
Biderman, 145
Bie, 239
Bielikova, 196, 216, 292
Biemann, 157, 253
Biggs, 120
Bikakis, 285
Bilen, 317
Binbin, 354
Bing, 330, 363, 383
Bingert, 59
Bingliwu, 51
Biran, 193
Birrer, 295
Bisazza, 171, 248
Bisk, 237, 279, 325
Bissyandé, 204
Biswas, 307
Bitterman, 217, 322
Bitton, 100
Biyik, 259
Bjerring-Hansen, 294
Bjerva, 75
Blain, 45
Blanchard, 116
Blanco, 49, 65, 254, 311
Blanco-Cuaresma, 307
Blaschko, 197
Blevins, 147, 148
Blloshmi, 85, 265
Blodgett, 59
Blouir, 133
Bobinac, 158
Bocklet, 177
Boeker, 139
Boenninghoff, 326
Bogdanov, 358
Bogin, 166, 212
Boix-Adserà, 76
Bolandraftar, 90
Boleda, 367
Bollegala, 123, 351
Bonaldi, 72
Bonifacio, 285
Boonnag, 388
Borah, 123
Borchert, 145
Borenstein, 133, 171, 299
Borges, 97
Borgholt, 41
Borgne, 104
Borimann, 91
Borthwick, 89
Borzilov, 397
Bos, 254
Bose, 313
Bosselut, 97, 125, 398
Bossi, 121
Bostrom, 138
Bottarini, 71
Bouayad-Agha, 90
Boudin, 401
Boulle, 67
Bouraoui, 254
Bout, 380
Bouyarmane, 56
Bouzoubaa, 150
Bowden, 228
Bowen, 349
Boyd-Graber, 86, 161, 246, 249, 284, 325
Brack, 134
Bradford, 116
Bradley, 56
Bragg, 77, 258
Brahman, 190, 325
Brandon, 292
Brannon, 49
Bransom, 166
Braun, 91
Breazeal, 47, 49
Brief, 74
Brinner, 269
Broeck, 69
Broek, 252
Broman, 80
Brown, 353, 370
Browne, 124
Bruggeman, 175
Bruni, 72
Bruns, 329
Brunskill, 44
Bruseva, 234
Brutti, 167
Bryksin, 397
Bu, 320, 404
Buchmann, 126
Buda, 113
Budhiraja, 282
Buettner, 314
Bugbee, 307
Bui, 215
Bukharin, 279
Bulat, 374

- Bullough, 91
Bunner, 100
Burdissò, 221, 222, 300
Burg, 278
Burja, 227
Burns, 100
Burtell, 170
Buschmeier, 270
Bussotti, 110
Buttery, 196, 236, 324
Byerly, 124
Byrne, 251
Byun, 137
Bárcia, 171
Bärmann, 166
Böhm, 166
Bkowicz, 303
Cabot, 156
Cabrio, 72, 174
Cachola, 326
Cafarella, 162
Cagliero, 292
Cahill, 171
Cahyawijaya, 142, 145, 148, 212
Cai, 84, 94, 101, 105, 109, 131, 137, 157, 185,
 212, 269, 286, 327, 350–352, 356,
 372, 373, 377, 381, 405
Caillou, 179
Caines, 236
Calderon, 190
Caldwell, 187
Caliskan, 261
Callanan, 305
Callison-Burch, 66, 223, 247, 265
Callot, 95
Cam-Tu, 224
Camacho-Collados, 123, 145
Cambria, 201, 358
Cambrin, 292
Cambronero, 95
Camburu, 76, 269
Campbell, 279
Campese, 161
Campos, 66
Cannon, 214
Canny, 397
CAO, 294, 323, 388
Cao, 44, 51, 65, 85, 103, 156, 159, 163, 171,
 191, 205, 209, 223, 238, 271, 289,
 320, 333, 337, 338, 346, 350, 352,
 361, 388, 391, 405
cao, 161, 369
Cappelli, 324
Caragea, 139, 176, 213, 253
Carbonneau, 141
Cardie, 114, 164
Cardoso, 213
Carlantonio, 399
Carlson, 53
Carmeli, 252
Carpenter, 318
Carpuat, 156, 288
Carrell, 314
Cascante-Bonilla, 105
Caselli, 296
Cashman, 287
Casola, 266
Cassani, 110
Cassell, 225
Cassotti, 213, 406
Castelli, 159
Castilho, 397
Casula, 71
Catanzaro, 73, 214
Cava, 255
Caverlee, 42, 191
Cecchi, 232
Cegin, 292
Ceker, 56
Celi, 217
Cen, 386
Cercel, 288, 311
Cerisara, 298
Ceritli, 279
Ceron, 52, 296
Cettolo, 167
CH, 265
Cha, 273
Chacko, 327
Chadha, 325
Chae, 61, 183, 207, 264
Chafei, 221
Chai, 235, 243, 263, 316, 341, 365–367, 374
Chairuengitjaras, 388
Chaitanya, 313
Chakrabarti, 403
Chakrabarty, 254
Chakraborty, 47, 81, 105, 183, 206, 213, 241,
 294, 295, 307
Chakravorti, 120
Chalaki, 279
Chaleshtori, 195
Chalkidis, 396
Chambers, 65
Chamoli, 218

- Chan, 67, 100, 152, 215, 253, 305
 Chandak, 276
 Chandar, 138
 Chander, 183
 Chandler, 281
 Chandra, 74, 111, 249, 357, 375
 Chandresh, 147
 Chang, 44, 49, 52, 69, 79, 84, 85, 88, 102, 106, 111, 115, 127, 130, 131, 134, 141, 154, 160, 165, 178, 179, 220, 235, 237, 247, 277, 278, 280, 281, 310, 316, 317, 319, 323, 335, 338, 343, 357, 391, 403
 Chanie, 340
 Chao, 228, 270
 Chapados, 202
 Chapanin, 190
 Chatterjee, 53, 213, 241
 Chattpadhyay, 272
 Chaturvedi, 117, 271, 299, 301
 Chaudhary, 191, 229, 296, 325
 Chawla, 322
 Che, 82, 384, 392
 Chechik, 318
 Chehade, 175
 Chekalina, 138
 Chelle, 116
 Chemla, 219
 CHEN, 87, 124, 269, 378
 Chen, 40, 46, 51, 52, 56, 60, 65, 68, 69, 73, 75, 78, 84, 89, 92–94, 101, 106, 107, 109–112, 115, 116, 119–121, 125–128, 130–132, 134, 136, 137, 140, 141, 143, 146, 148, 151, 154, 155, 157, 159–162, 165–168, 171, 176, 179, 180, 182, 186, 189–191, 193, 194, 197, 199, 202, 204, 208, 210, 213, 216, 217, 220, 226, 228, 231, 233, 236, 238, 241–244, 248, 249, 252, 255, 256, 258, 261, 263, 268, 270, 271, 273, 275–278, 280, 282–285, 287, 290, 293, 294, 296, 300, 305–309, 311, 318–323, 325, 327, 328, 330, 331, 333, 335–337, 339, 340, 343–348, 350, 351, 353–355, 359–363, 367–369, 372, 374–377, 379, 380, 383, 385–387, 389–393, 395, 397–401, 405, 406, 408, 409
 chen, 197
 Cheng, 50, 51, 57, 58, 62, 64, 65, 68, 83, 86, 90, 106, 117, 124, 139, 166, 179, 194, 202, 205, 206, 225, 241, 259, 260, 272, 275, 277, 291, 300, 314, 321, 328, 333, 345, 354, 358–361, 365, 376, 379, 386, 388, 404, 406
 cheng, 351, 378
 Chengbaolian, 56
 ChengFu, 51
 Chenglin, 368
 Chenkang, 56
 Cheong, 148
 CHESNEAU, 76
 Cheung, 350
 Chevalier, 94
 Chi, 320
 Chia, 142, 383
 Chiang, 107
 Chieng, 215
 Chigrupaatii, 59
 Chilimbi, 239
 Chilton, 47
 Chim, 64
 Chin, 193
 Chinchure, 149
 Ching, 372
 Chinkamol, 388
 Chinta, 289
 Chirkova, 250
 Chiu, 76, 110
 Chiyah-Garcia, 47
 Chizhov, 275
 Chng, 219
 Cho, 58, 61, 91, 162, 178, 181, 183, 212, 229, 240, 252, 297, 305, 321
 Chodroff, 96
 Choenni, 259
 Choi, 60, 77, 83, 89, 91, 112, 114, 146, 170, 174, 175, 177, 179, 187, 188, 190, 193, 206, 213, 216, 221, 222, 227, 229–231, 255, 267, 276, 280, 301, 303, 319, 324, 325, 328, 355, 369, 395
 choi, 112, 183
 Cholakkal, 327
 Cholakov, 292
 Chollampatt, 147
 Chon, 327
 Chong, 365
 Choo, 149, 216, 324
 Chopra, 280
 Choquette-Choo, 74
 Choshen, 129, 131
 Chou, 152
 Choudhary, 293
 Choudhury, 43, 71, 151

- Chow, 205
Chowdhury, 105
Christ, 324
Christmann, 251
Christophe, 240
Chrysostomou, 256
Chu, 224, 295, 331, 354, 370, 386, 388, 392
Chua, 161, 234, 302, 330, 346
Chuang, 69, 191, 298
Chuangsuwanich, 143, 358
Chugh, 306
Chun, 137
Chung, 49, 54, 88, 103, 127, 240, 264, 314
Church, 172, 282
Churpek, 322
Cideron, 279
Cifuentes, 258
Clark, 46, 166, 200, 253, 370
Clarke, 282
Clifton, 134
CLINCHANT, 250
Clotan, 288
Coady, 170
Coavoux, 407
Cognetta, 117
Cohan, 51, 90, 167, 169, 170, 223, 257, 262,
 266, 287, 345
Cohen, 65, 114, 128, 154, 164, 194, 209, 212,
 213, 233
Cohen-Addad, 236
Cohn, 163, 248
Collier, 58, 189, 268, 340, 401
Colombo, 156
Comar, 63
Cong, 82, 228
Conia, 45, 250
Conroy, 294
Constantin, 358
Constantinescu, 256
Contalbo, 116
Cook, 110
Cooper, 70
Corallo, 50
Corlatescu, 217
Cornejo, 193
Correia, 185
Corro, 260, 409
Cosma, 253
Costa-jussà, 77, 158, 220, 272
Costes, 307
Cotterell, 72, 86, 96, 97, 99, 100, 118, 133,
 256, 271
Cotton, 260
Couturier, 357
Cox, 57
Crabbé, 358
Craciun, 288
Cremer, 276
CREPY, 397
Cristea, 142
Crook, 218
Crouse, 57
Cruz, 142
Cucerzan, 326
Cuervo, 219
Cui, 57, 110, 134, 136, 157, 165, 209, 214,
 228, 230, 237, 240, 243, 289, 322,
 355, 363
Curry, 294
- D, 219
d'Amore, 212
D'Avirro, 219
D'Haro, 106
D'Oosterlinck, 188
Dabrowski, 132
Dabre, 195, 198
Dachsbacher, 166
Dagan, 172, 237
Dahan, 326
Daheim, 225
DAI, 320
Dai, 94, 124, 131, 137, 179, 194, 231, 242,
 243, 251, 278, 340, 344, 348, 355,
 372
Dale, 158
Dall'Asen, 316
Dalton, 53
Dalvi, 128
Daly, 348
Damanhuri, 142
Damavandi, 91, 218
Dambanemuya, 171
Damiano, 287
Dammu, 71, 325
Damnati, 129
Damo, 72
Dandala, 307
Dandan, 340
Dandekar, 186
Danescu-Niculescu-Mizil, 297
Dang, 75, 334
Dann, 279
Dao, 112, 278, 380, 401
Darmon, 146
Darrell, 315

- Das, 53, 57, 60, 70, 78, 130, 168, 199, 218, 224, 230, 301, 322
 Dasaratha, 384
 Dascalu, 217, 253
 Dasgupta, 183
 Dash, 293
 Dasigi, 182, 290
 Dastani, 224, 348
 Datta, 43
 Davani, 152
 Davinroy, 110
 Davis, 305
 Davoudi, 399
 Dawkins, 52
 Daxenberger, 173
 Dazhi, 395
 Deas, 71
 Deb, 147
 Debray, 70
 Debrunner, 281, 307
 DeButts, 153
 decombas, 214
 Dedert, 253
 Deguchi, 114
 Dehaze, 143
 Dehghan, 329
 Dehghani, 300
 Dehnen, 399
 Deilamsalehy, 215
 Deiseroth, 134
 Del, 324
 Delattre, 179
 Dell, 294
 Dell'Orletta, 163
 Delobelle, 59
 Deluca, 90
 Demberg, 263
 Demszky, 303
 DeNeefe, 155
 Deng, 65, 67, 70, 75, 89, 92, 100, 107, 114, 160, 161, 223, 269, 271, 302, 323, 346, 347, 351, 356, 373
 Denis, 313, 389
 Dennis, 72
 Deoghare, 157
 Deoras, 208
 Dernoncourt, 141, 154, 215, 245, 400
 Deshmukh, 375
 Deshpande, 273, 298
 Dethlefs, 295
 Dev, 398
 Devalla, 338
 Devasier, 114
 Devianti, 254
 Devoto, 42
 Dewan, 325
 Dhalwani, 240
 Dhar, 98
 Dharmani, 233
 Dhillon, 304
 DHIMOÏLA, 128
 Dhingra, 86, 211, 266, 283
 Dhuliawala, 139, 189
 Diab, 408
 Diallo, 285
 Diamantis, 64
 Diandaru, 142
 Diao, 84, 125, 136, 160, 220, 238, 245, 328, 378, 405
 Diaz, 78, 152, 204, 259, 299
 Dibia, 330
 Dickens, 285
 Diesner, 43
 Diffenderfer, 137
 Dige, 90
 Dima, 217, 288
 Dimakis, 60
 Dimitrov, 397
 Ding, 102, 105, 128, 136, 139, 159, 172, 176, 209, 211, 237, 238, 244, 249, 272, 310, 325, 327, 328, 343, 349, 356, 379, 380, 385, 393
 Dingliwal, 232
 Dinh, 166, 334
 Dinkar, 261
 Dinu, 142
 Disha, 232
 Divakaran, 244
 Divekar, 87, 264
 Diwan, 222
 Dixit, 82
 Djanibekov, 142
 Dligach, 322
 Do, 111, 142, 375
 Dobbie, 75
 Dobriban, 136
 Doddapaneni, 64
 Dodge, 182, 290
 DoeunKim, 312
 Dognin, 231
 Dolfi, 307
 Domingo, 44
 Dondrup, 317
 Doneva, 214
 DONG, 225

- Dong, 42, 64, 95, 105, 110, 122, 131, 135, 136, 146, 168, 182, 189, 197, 208, 211, 218, 224–226, 278, 309, 312, 318, 355, 365, 370, 378, 386, 388, 405
DONGDONG, 293
Donvito, 324
Doo, 93
Dorn, 71, 295
Doshi, 198
Dotzel, 290
Dou, 79, 92, 106, 134, 176, 198, 226, 239, 321, 323, 344
Doveh, 131
Dover, 148
Downey, 147
Dowpati, 343
Dragut, 175, 213
Dras, 203
Dredze, 44, 123, 132, 234, 260
Drummond, 248
Du, 62, 65, 72, 117, 120, 135, 151, 186, 196, 201, 205, 248, 271, 299, 302, 317, 355, 376
du, 124
Duan, 217, 287, 291, 361, 376, 377, 382, 406
Duarte, 234
Dubey, 279
Dubossarsky, 269
Ducatelle, 90
Duderstadt, 148
Dufaux, 214
Dugan, 223, 247
Duh, 156
Duki, 255
Dumitru, 288
Duong-Tran, 327
Dupoux, 77, 219
Dupuy, 280
Duquenne, 77
Duraiswami, 58
Durmé, 132, 138, 212, 245, 253, 318
Durrani, 128
Durrett, 60, 184, 211, 264
Dusek, 216, 274
DuSell, 99
Dutta, 84, 113, 206, 213, 296, 357
Dwivedi-Yu, 304
Déjean, 250
Dönmez, 194
Döring, 91
Eack, 110
Earls, 67
Eberhardt, 281, 307
Ebert, 145
Ebling, 314
ECH-CHAMMAKHY, 221
Echterhoff, 188, 189
Eddie, 405
Edin, 41
Edman, 132
Eger, 48, 210, 387, 407
Ehghaghi, 399
Eickhoff, 145
Eide, 164
Eisape, 227
Eisenstein, 131
Eisner, 132, 138
Ekbal, 59, 66, 69, 79, 162, 227, 246, 343
El-Kurdi, 307
El-Refai, 88
El-Shangiti, 221
Elangovan, 232
Elaraby, 175
Elazar, 41, 84, 259
Elbayadb, 77
Elesedy, 121
Elhoseiny, 282
Elisha, 74
Ellendorff, 214
Elliott, 246
Elozeiri, 82
Emami, 52, 350
Emde, 75
Eom, 221, 375
Epps, 106
Epure, 298
Erdogan, 223
Erler, 246
Ermis, 143, 270
Ernst, 301
Ertekin, 262
Esfandiarpoor, 103
eshemChoshen, 218
Eshghi, 47, 245, 269
Esperanca, 121
Espinosa-Anke, 123
Estarrona, 174
Estrada, 280
Etchegoyhen, 158, 265
Etter, 318
Ettinger, 128
Eugenio, 204
Eustratiadis, 171
Evans, 120
Everaert, 231, 400

- Evuru, 58
Ewaleifoh, 218
Eyal, 40
Ezquerro, 118
Ezzini, 204

Fabbri, 63, 170, 183
Fabrikant, 312
Fadaee, 40, 143
Faddoul, 278
Fadhilah, 142
Fadlallah, 177
Faghri, 188
Fagnou, 179
Fahim, 201
Fahrezi, 178
Fairstein, 233
Falenska, 194, 272
Falk, 52, 188
Faloutsos, 199, 218
Faltungs, 214
FAN, 346, 348
Fan, 40, 67, 70, 100, 102, 121, 144, 154, 170,
 185, 228, 259, 309, 333, 343, 358,
 360, 362, 375, 385, 386, 389, 392,
 394
Fancher, 307
Fang, 44, 65, 82, 90, 101, 109, 110, 138, 172,
 179, 238, 244, 276, 277, 284, 286,
 287, 301, 320, 331, 340, 346, 366,
 376, 397
Far, 105
Farag, 115
Farahani, 362
Farchi, 389
Farfade, 63
Farinhas, 66, 155
Farmaner, 399
Farn, 409
Farrús, 199
Farup, 164
Faruqui, 87
Farzana, 66, 326
Fashandi, 232, 305
Fathullah, 108
Fayyazsanavi, 202
Fazel-Zarandi, 154
Feder, 40, 193
Fei, 133, 301, 371, 381, 406
Feichtenhofer, 54
Feith, 207
Feizi, 101, 315
Fekri, 285

Feldhus, 162
Feldman, 99, 100
Fellenz, 334
FENG, 317
Feng, 51, 75, 83, 86, 100, 108, 114, 115, 136,
 154, 155, 159, 163, 181, 187, 199,
 206, 208, 211, 225, 234, 249, 252,
 289, 302, 311, 316, 320, 347, 350,
 352, 359–361, 372, 373, 381, 385,
 387, 389, 394, 398
FengTao, 368
Fenogenova, 387
Ferdinan, 201
Fernandez, 109, 279
Fernández, 247, 248
Ferrand, 147
Ferrando, 129, 272
Ferraz, 88
Ferreira, 232, 304, 305
Ferret, 104, 279
Fetahu, 89
Feucht, 127
Ficek, 264
Fidler, 245
Field, 149, 261
Fierro, 54, 98
Figueiredo, 66
Filandrianos, 86, 283
Filatov, 397
Filos, 372
Finch, 227, 255
Fini, 316
Finn, 276
Firdaus, 69
Fishel, 324
Fisher, 83, 187, 218
Flanagan, 62
Flanigan, 48
Fleischer, 385
Fleisig, 121, 259
Flet-Berliac, 276
Fletcher, 399
Flores-Herr, 145
Florian, 61
Flynn, 142
Fogliato, 165
Fok, 165
Fons, 169
Foo, 112
Forde, 145
Forey, 70
Formal, 250
Foroosh, 183

- Foroutan, 125
Foss, 154
Fourney, 330
Frank, 100, 284
Frantar, 368
Fraser, 52, 117, 132, 173, 295
Frassinelli, 99, 274
Frei, 139
Freire, 234
Freitag, 234
Freitas, 67, 72, 197
Fridman, 260
Fried, 111
Friedrich, 62
Frigo, 298
Frisoni, 110
Fritz, 128
Frohmann, 178, 349
Fromm, 145
Frydenlund, 275
Fu, 65, 154, 211, 212, 224, 244, 284, 306, 332, 334, 337, 353, 358, 366, 405
Fucci, 259
Fuchs, 146
Fuge, 340
Fukatsu, 56
Fukumoto, 110
Fulay, 49
Fung, 212, 222, 223
Furniturewala, 258
Furumai, 302
Färber, 67
Gabburo, 161
Gabriel, 150, 325, 399
Gacche, 66
Gadeppally, 181
Gagliardelli, 406
Gai, 86, 330
Gaido, 167
Galan, 231
Gales, 108, 126, 128, 220
Galgani, 172
Galimzianova, 356, 369
Gallifant, 217
Gallipoli, 292
Gallé, 286
Galstyan, 278, 280, 281
Gambardella, 238
Gan, 55, 166, 211
Ganadi, 406
Ganapathiraju, 221, 222
Ganapathiraman, 279
Ganchev, 70
Gandhi, 116, 261
Ganeeva, 195
Ganesan, 281
Ganeshmohan, 240
Gang, 277
Ganguly, 135, 168, 403
Gantt, 184
GAO, 142
Gao, 50, 65, 75, 83, 87, 94, 106, 109, 115, 119, 122, 124–126, 134, 137, 164, 166, 192, 202, 217, 223, 224, 238, 251, 262, 312, 318, 322, 342, 350, 356, 358–360, 363, 369, 373, 379, 389, 392, 396, 398, 404
Gareev, 267
Garg, 80, 100, 176
Garinella, 215, 267, 319, 325
Garland, 110
Garneau, 98
Garrison, 337
Garza, 292
Gashteovski, 273
Gastaldi, 99
Gat, 77
Gatt, 98
Gatto, 67
Gaube, 76
Gaur, 285
Gautam, 42, 59, 119, 124, 346, 402
Ge, 110, 161, 211, 319, 353, 367, 406
Gedeon, 187
Geh, 69
Gehrman, 44, 132
Geierhos, 312
Geiger, 188
Geigle, 243
Geist, 276
Gekhman, 40, 190
Gelmi, 279
Gemechu, 408
Genabith, 313
Gendron, 75
Geng, 82, 114, 171, 257
Gentile, 90
George, 229, 287
Georgescu, 142
Georgiev, 130, 257
Gerasimenko, 397
Gerasimov, 307
Gerdjikov, 73
Gere, 46
Gerlach, 207

- Gerstein, 223
Gertz, 234
Gessler, 60
Gete, 158
Geva, 65, 124–126, 193
Ghaddar, 92
Ghaffari, 247
Ghafouri, 329
Ghanem, 261
Ghanim, 123
Ghannay, 104
Ghassemi, 150, 399
Ghazarian, 280
Ghinassi, 117
Ghodsi, 400
Gholaminejad, 223
Ghosal, 66, 195, 315
Ghose, 277
Ghosh, 54, 58, 77, 88, 135, 184, 195, 280, 281, 307, 394
Giadikiaroglou, 86
Gigant, 214
Gilbert, 44
Gill, 195
Gillani, 149
Gilsenan-McMahon, 286
Gim, 307
Ginn, 144, 336
Gipp, 184, 192, 230
Giryes, 246
Gispert, 251
Gittens, 266
Giulianelli, 98–100
Giuliani, 345
Gizzi, 339
Glass, 61, 69
Glava, 145, 234, 243
Globerson, 193
Glória-Silva, 316
Godbole, 86
Goddard, 399
Goel, 289
Goharian, 183
Golac, 89
Golany, 172
Golazizian, 300
Goldberg, 209, 233, 234
Goldfarb-Tarrant, 86, 143
Goldman, 214, 237
Goldstein, 148
Goldwasser, 48, 152
Gollakota, 221
Gomez, 221, 222
Gomez-Sebastia, 90
Gonen, 142, 148
GONG, 160
Gong, 68, 141, 196, 207, 209, 211, 221, 238, 242, 249, 289, 291, 312, 331, 357, 373, 390
Gongas, 146
GongQue, 405
Gonzalez, 180
Gonzalo, 79
Goodman, 100
Gor, 161
Gorade, 375
Gorbunov, 157
Gordon, 83, 239
Goriely, 236
Goswami, 120, 194
Gottesman, 125, 193
Gou, 87, 186, 205
Gouda, 208
Govindarajan, 297
Gowda, 186
Goyal, 61, 81, 94, 166, 168, 205, 278, 280, 281, 298, 313
Grabowski, 135
Gratch, 322
Graça, 234
Greco, 255
Greene, 111
Greenstein-Messica, 306
Gretter, 167
Grezes, 307
Griffiths, 274
Grimmelmann, 213
Grimmer, 295
Grinsztajn, 276
Grishina, 309
Groschwitz, 254
Gross, 200
Grossman, 366
Grotov, 397
Grundkiewicz, 186
Gröner, 210
Gschwind, 305
Gu, 65, 85, 128, 134, 146, 173, 197, 220, 225, 252, 257, 258, 271, 286, 289, 293, 309, 333, 334, 340, 348, 351, 352, 356, 357, 373, 389
Gualdoni, 367
Guan, 102, 106, 111, 209, 226, 238, 246, 262, 293
Gubelmann, 97
Guerberof-Arenas, 123

- Guerini, 48, 72
Guerra, 116
Guerraoui, 174, 175
Guerreiro, 156, 158
Guerrero, 374
Gui, 42, 109, 127, 226, 249, 275, 311, 321, 328, 333, 374, 389
Guidotti, 60
Guinaudeau, 214
Gul, 104
Gulati, 278
Gulcehre, 321
Gulla, 164
Gulwani, 113, 203, 286
Gumma, 165
Gunasekara, 57
Gunel, 185
Guo, 50, 52, 62, 65, 83, 105, 107, 133, 136, 137, 139, 143, 149, 160, 212, 237, 279, 290–292, 295, 307, 308, 320, 321, 333, 339, 344, 348, 353, 355, 358, 360, 370, 372–374, 380–382, 384–386, 400, 401
guo, 404
Gupta, 65, 79, 89, 120, 124, 128, 146, 162, 166, 195, 196, 203, 205, 217, 270, 278–281, 288, 306, 307, 319, 344, 352, 396, 403
Gurevych, 82, 94, 126, 150, 171, 173, 218, 225, 248, 257, 295, 309, 387, 397
Gurung, 265, 307
Gururangan, 148
Gwak, 216
gweon, 189
Gállego, 220
Gómez-Rodríguez, 60, 118
- Ha, 349
Haake, 191
Habash, 82
Habernal, 59, 262
Habibi, 142
Habiboullah, 221
Hachiuma, 122
Hadar, 55
Haddow, 143, 154
Hadfi, 152
Hadiwijaya, 142
Haehn, 322
Haensch, 188
Haf, 105, 333, 339, 342
Haga, 56
Hagen, 236
- Hahm, 207
Hai, 310
Haider, 201
Hain, 142
Hajishirzi, 182, 213
Hakimov, 265
Hale, 98
Halevy, 398
Hallinan, 83
Halliwell, 353
Hamborg, 198
Hammond, 84
Hammoud, 261
HAN, 170, 287
Han, 41, 43, 53, 54, 56, 65, 70, 73, 75, 89, 90, 95, 96, 112, 114, 115, 122, 128, 139, 156, 159, 175, 179, 192, 201, 209, 221, 222, 226, 229, 245, 252, 267, 276, 284, 289, 291, 292, 296, 297, 302, 305, 308, 310–312, 344, 352–354, 406
han, 316
Hanawal, 200
Handan-Nader, 49
Hangya, 173, 295
Hanif, 180
hanminwang, 223
Hao, 40, 83, 164, 205, 249, 384, 392
Haotian, 385
Harchaoui, 83
Hardalov, 49
Hardy, 110
Harel-Canada, 78
Harrenstien, 403
Harris, 222
Harrison, 228
Harsha, 82, 384
Hartvigsen, 217, 288, 324
Harwath, 222
Hasan, 169
Hasanain, 169, 298
Hasegawa-Johnson, 375
Hashimoto, 276
Hasibi, 201
Hasnat, 169
Hassani, 136
Hassid, 129
Hasson, 235
Hatzel, 253
Hauptmann, 64, 355
Havtorn, 41
Hayati, 120
Haydarov, 282

- Hays, 78
 Hazarika, 227
 Haznitrama, 169
 Hazra, 357, 365
 He, 42, 53, 58, 62, 95, 112–115, 119, 125–127,
 130, 132, 135, 137, 139, 144, 151,
 153, 167, 172, 176, 177, 189, 192,
 194, 201, 203, 206, 208, 211, 212,
 216, 235, 240, 243, 249, 252, 277,
 278, 280, 286, 292, 295, 305, 308,
 311, 317, 320, 321, 328, 331, 333,
 337, 339, 347, 359, 365, 367, 375,
 383, 384, 387, 388, 392, 393, 401,
 404, 405
- Heck, 235
 Hedderich, 188
 Hee, 149, 294
 Heer, 286
 Hegde, 394
 Heidari, 91
 Heidemann, 253
 Heil, 189
 Heineman, 79
 Heinzelring, 194
 Helcl, 117
 Heldring, 294
 Hell, 97
 Helm, 148
 Hemanthage, 317
 Hemmati, 395
 Hendria, 142
 Hendricks, 54
 Hengle, 396
 Hennen, 312
 Hennequin, 298
 Hensman, 68
 Heo, 175
 Hermawan, 142
 Hernandez, 118, 337
 Herold, 181
 herring, 326
 Hershcovich, 174, 246, 294, 359
 Herzig, 40, 315
 Hessel, 54, 114, 325
 Heumann, 267
 Heyer, 368
 Hiatt, 191
 Hill, 212
 Himmi, 156
 Hinck, 295
 Hinkle, 353
 Hirasawa, 146
 Hirota, 122, 258
- Hirschberg, 85, 110, 220, 227, 304, 312
 Hng, 344
 Ho, 49, 207, 322, 323
 Hobson, 150
 Hockenmaier, 240
 Hoefer, 368
 Hoeken, 172, 173
 Hoeve, 241
 Hofmann, 267
 Holliday, 72
 Homann, 296
 Honavar, 134
 Honda, 80
 HONG, 112
 Hong, 49, 65, 127, 195, 198, 230, 267, 270,
 290, 305, 327, 341, 361, 376, 377,
 385
 Hongchao, 293
 Hoogs, 110
 Hooker, 40, 75, 143, 293
 Hooper, 223
 Hooshmand, 84
 Hoque, 73, 169, 172, 266, 312
 Horie, 63
 Horvitz, 265
 Horváth, 157
 Hosking, 257
 Hossain, 193, 266
 Hosseini, 312, 326
 Hou, 58, 67, 135, 163, 210, 217, 218, 279, 307,
 309, 312, 325, 327, 347, 351, 355,
 364, 385, 393
 Houmansadr, 84
 Hovy, 259, 294
 Howe, 218
 Hoyle, 86
 Hrckova, 399
 Hsiao, 190
 Hsieh, 52, 69, 180, 241, 244, 260, 267
 Hsu, 100, 196, 224, 229, 241
 Hsueh, 260, 277
 HU, 216
 Hu, 43, 59, 64, 65, 87, 89, 91, 101, 102, 108,
 115, 120, 124, 131, 135, 136, 143,
 149, 158, 159, 163, 164, 173, 174,
 178, 179, 183, 188, 189, 191, 207,
 211, 212, 220, 226, 238, 242, 246,
 251, 258, 278, 280, 292, 298, 305,
 308, 311, 316, 317, 325, 334, 336,
 342–344, 348, 350, 353, 359, 360,
 363, 367, 371, 373, 375–377, 379,
 382, 388, 396, 397, 405
- HUA, 159

- Hua, 130, 134, 154, 237, 260, 349
HUANG, 124
Huang, 44, 46, 51, 53, 54, 56, 59, 60, 65, 68, 73, 75–78, 83, 86, 88, 92, 94, 100, 106, 115, 127, 130, 134, 136, 138, 139, 144, 150, 152, 154, 155, 159, 163, 165, 167, 170, 173, 179, 182, 184, 190, 203, 204, 207, 210, 211, 214, 224, 226, 233, 238, 240–242, 244, 246, 247, 249, 251, 260, 262, 266, 271, 272, 275, 277, 280–283, 289–292, 300, 308, 309, 311, 312, 317, 319–322, 328, 333, 335–338, 340–342, 346, 347, 349–351, 353, 355, 357–359, 363, 364, 372, 374, 376, 381, 383, 385, 389, 391–394, 397, 402, 406
- huang, 65
Huber, 174, 357
Huda, 290
Hudi, 142
Huerta-Enochian, 239
hufeng, 374, 391
Hui, 85, 110
Hulden, 336
Hull, 297
Hunter, 117, 285
Huo, 154, 282, 336, 360
Huot, 46, 70
Hupkes, 72
Hurtado, 278
Hus, 156
Hussenot, 279
Huynh, 109, 121
Hwang, 90, 137, 141, 144, 149, 162, 185, 199, 201, 250, 269, 280, 301
hwang, 77, 96, 113, 178, 207
Hyeon, 303
Hänni, 128
Häatty, 91
- Iakovenko, 142
Iana, 234
Idris, 345
Ie, 135
Igamberdiev, 262
Ignatov, 344
III, 44, 77, 161, 174, 258
Ikbal, 57
Ikeda, 302
Ilhan, 292
Ilie-Ablachim, 217
Ilievski, 55, 273
- Ilin, 245
Iluz, 259
Ilyas, 231
Imoto, 369
Imperial, 70, 142, 255
Indu, 231
Indurthi, 138, 147
Ineichen, 214
Ingle, 229
Ingvaldsen, 164
Inkpen, 235
Inoue, 174, 175
Inui, 174, 175, 194, 364
ION, 311
Iordache, 142
Ippolito, 345
Iqbal, 82, 257
- Irawan, 142
Irsoy, 132
Irving, 102
Isaac, 259
Ishigaki, 264
Ishii, 212
Ishmam, 201, 337
Iskander, 164
Islam, 169, 172, 266, 316
Iso, 166
Ittachaiwong, 388
Ivanova, 174
Ive, 64
Iwasawa, 238
Iyyer, 84, 168, 170, 267
Izsak, 385
- Jaakkola, 69
Jacovi, 237
Jada, 340
Jafari, 201
Jagatap, 57
Jagerman, 332
Jaggi, 259
Jahan, 73
Jaidka, 258
Jaimes, 171
Jain, 54, 88, 91, 140, 160, 229, 245, 261, 264, 343, 356
JAISWAL, 179, 180
Jaitly, 293
James, 297, 334
Jampani, 304
Janardhanan, 62
Jandial, 258

- Jang, 91, 95, 103, 149, 189, 199, 233, 301, 303, 305, 310, 327, 355
Jansen, 253, 269
Janssen, 212
Jaques, 397
Jaravine, 139
Jarrar, 221
Jatowt, 385
Jauhar, 326
Java, 258
Jaya, 142
Jayanthi, 250
Jedema, 212
Jen, 184
Jennings, 73
Jeon, 91, 112, 189, 222, 267, 314
Jeong, 80, 162, 229, 250, 382
Jetter, 191
Jeung, 178
Jha, 223, 290
Jhamtani, 132, 138, 212, 285
Jhunjhunwala, 73
Ji, 56, 58, 73, 84, 93, 102, 105, 109, 125, 136, 144, 178, 192, 222, 223, 233, 235, 264, 299, 317, 323, 333, 335, 337, 376, 378, 380
ji, 154
Jia, 121, 127, 137, 160, 161, 231, 254, 270, 291, 323, 349, 374, 401
Jian, 98, 374
JIANG, 228
Jiang, 51, 53, 55, 56, 60, 68, 83, 85, 88, 102, 108, 109, 121, 126, 134, 136, 137, 154, 158, 161, 164, 167, 181, 187, 192, 203, 204, 206, 224, 228, 229, 233, 248, 253, 261, 270, 271, 277, 279, 282, 286, 292, 293, 314, 317, 325, 327, 341, 361, 362, 364, 365, 367, 373, 375, 379, 381, 384, 393, 404
jiang, 368
Jiao, 112, 148, 159, 163, 203, 207
Jiayang, 65, 100, 165, 253, 315
Jiayu, 206
Jie, 201
JIN, 293
Jin, 52, 70, 73, 95, 106, 124, 130, 131, 137, 151, 154, 159, 174, 181, 182, 189, 191, 198, 209, 214, 215, 227, 242–244, 258, 260, 275, 280, 301, 305, 308, 310, 313, 355, 374, 376, 385, 391, 396, 401
JING, 244
Jing, 248, 374
Jinnai, 140
Jinxiaojia, 347
Jo, 91, 183
Johansson, 362
Johnson, 89, 330
Johri, 310
Jones, 57, 110
Joo, 221, 227
Joshi, 47, 57, 63, 71, 78, 125, 178, 192, 285
Joty, 51, 63, 73, 159, 169, 170, 266, 287, 330
Jou, 54
JU, 154
Ju, 129, 333
Jumelet, 76
Junczys-Dowmunt, 186
Juneja, 206
Jung, 50, 71, 77, 117, 260, 297, 299, 301, 328, 355, 382, 400
Junker, 270
Junlin, 224
Jurgens, 273, 298, 299, 324
Jwa, 170

K, 232
Kabbara, 49
Kabir, 249
Kabra, 45
Kadaoui, 170, 221
Kadlík, 201
Kadurin, 195
Kaelin, 66
Kahn, 54
Kai, 307
Kailkhura, 137
Kairouz, 74
Kaiser, 301
Kale, 334
Kalinsky, 233
Kallala, 44
Kamaloo, 285
Kambadur, 132
Kambhatla, 152
Kamigaito, 45, 114, 156, 157, 276, 291
Kammakomati, 263
Kamoi, 41
Kampman, 142
Kamruzzaman, 121
Kan, 86, 95, 166, 217, 290, 318
Kanagaraj, 307
Kanagarajan, 56
Kandpal, 74
Kane, 346

- Kaneko, 266
Kang, 53, 91, 95, 112, 117, 120, 144, 152, 173,
196, 207, 223, 234, 250, 270, 324,
382, 407
kang, 339
Kangaslahti, 178
Kanithi, 240
Kannen, 278
Kanojia, 45, 91, 157
Kanoulas, 158
Kantarcioğlu, 53
Kantu, 217
Kao, 154, 267, 381
Kapadnis, 225
Kapanipathi, 57
Kapur, 225
Karagöz, 118
Karami, 51
Karanam, 57
Kargupta, 302
Karidi, 366
Karim, 232
Karimi, 340
Karlinsky, 131
Karls, 326
Karlsson, 142
Karmaker, 254
Karnin, 164
Karpinska, 168
Karpov, 344
Karpukhin, 399
Karray, 114
Kartik, 217
Karver, 123, 260
Karypis, 199, 326
Kasai, 170
Kasat, 205
Kashima, 162
Kasianenko, 329
Kasneci, 298
Kataria, 195
Katsigiannis, 234
Katsimpras, 93
Katz, 65, 234
Katz-Samuels, 239
Kaufman, 123, 260
Kaur, 169
Kautsar, 142
Kavathekar, 218
Kawabata, 193
Kawaguchi, 166, 290
Kawarada, 264
Kayser, 75
Kazienko, 201
Ke, 355, 357
KediChen, 216
Keh, 142
Keller, 210, 389
Kelly, 191
Kementchedjhieva, 54
Kersting, 134
Kesarwani, 307
Keutzer, 180, 211, 223
Kew, 146
KHADEMI, 397
Khadivi, 181
Khan, 64, 73, 240, 246, 327
Khandelwal, 146, 217
Khanehzar, 329
Khanuja, 64
Khapra, 64
Kharlamov, 328
Khasanova, 306
Khashabi, 124, 193
Khasin, 306
Khatib, 176, 296
Khattab, 80, 303
Khattak, 90
Khelli, 142
Khetan, 199
Khetani, 66
Khondaker, 206
Khot, 166
Khrabrov, 195
KhudaBukhsh, 296
Khule, 408
Kidambi, 279
Kiegeland, 97
Kiela, 105
Kil, 137
Kilicoglu, 175
KIM, 149, 199, 204, 309
Kim, 47, 54, 58, 60, 69, 70, 76–79, 84, 86, 90,
91, 94–96, 102–104, 106, 107,
111, 113, 119–121, 125, 130, 131,
133, 141, 144, 146, 149, 162, 164,
165, 167, 170, 175, 177–180, 182,
183, 189, 199, 201, 204, 205, 207,
208, 221–223, 226, 227, 229,
231–233, 238, 250, 252, 260, 263,
264, 266, 269, 280, 283, 289, 292,
301, 303, 304, 307, 309, 310, 315,
317–319, 323–325, 327, 328, 334,
345, 375, 378, 382
Kimyeeun, 280
Kinchagawat, 388

- King, 48, 115, 163, 251
Kipnis, 372
Kirchner, 110
Kiritchenko, 52
Kirstein, 184, 230
Kirtania, 203
Kiseleva, 282
Klakow, 59, 119, 124, 154
Klein, 121, 204
Klie, 397
Klinger, 287, 297
Kloft, 334
Kloots, 256
Knaus, 139
Knill, 220
Knuple, 272
Ko, 61, 91, 143, 167, 191, 231, 239, 242, 277,
 303
Kobayashi, 247, 256
Kobren, 249
Koeppel, 94
Koh, 182, 187, 213, 260, 278
Koide-Majima, 99
Koike, 266
Kojima, 238
Koller, 118, 119, 254
Kolossa, 326
Komachi, 146, 256
Komatsu, 153
Koncha, 387
Konen, 176
Kong, 89, 95, 107, 196, 203, 229, 239, 340,
 353, 376
Konidaris, 244
Konovalov, 344, 356
Konstas, 52, 104, 261, 269
Koo, 47, 56, 105, 173, 232, 310, 312, 375
Koopman, 93
Korakakis, 140
Korhonen, 58, 95, 189, 271, 284, 340, 409
Korotkova, 275
Korzanova, 344
Kosecka, 202
Koshiyama, 262
Koshkin, 371
Kostikova, 48
Kosugi, 200
Koto, 142, 329
Kou, 364, 404
Koumoundouros, 71
Koutra, 262
Kovacec, 283
Kovashka, 314
Kowshik, 87
Kramer, 139
Kratel, 230
Krayko, 356
Kreiss, 188
Kretchmar, 252
Kreuter, 188
Kreutzer, 40, 75, 143
Krishna, 61, 69, 84, 87, 100, 180
KRISHNAN, 203
Krishnan, 89
krishnasamy, 267
Krishnaswamy, 116, 247
Krivobok, 397
Kriz, 318
Kropko, 324
Krumdick, 141
Krumnack, 253
Kryscinski, 170
Ku, 375
Kuai, 354
Kuan, 107
Kuang, 286, 335, 354, 382
Kubo, 99
Kuchaiev, 264
Kuchmiichuk, 184
Kudo, 364
Kudugunta, 398
Kulikov, 397
Kulkarni, 78, 254, 282, 396
Kulshrestha, 322
Kumar, 53, 57, 58, 63, 66, 77, 91, 140, 153,
 168, 216, 221, 222, 259, 263, 264,
 322, 394
Kumarage, 110
Kumaraguru, 218
Kumaravel, 57
Kumaresan, 149
Kumari, 343
Kummerfeld, 211
Kundu, 200, 205
Kuo, 102, 268
Kurakin, 239
Kurata, 63
Kurabayashi, 364
Kurtic, 168
Kushilevitz, 233
Kushnareva, 380
Kuznetsov, 380, 387, 397
Kwak, 89, 207, 264
Kwakpovwe, 99
Kwan, 282, 310
Kweon, 58

- Kwiatkowski, 160
Kwok, 172
KWON, 217
Kwon, 91, 133, 207, 226, 240, 264, 287, 296, 322, 400
Kyle, 402
Käser, 97
Köksal, 95, 271, 284, 313
Kühnberger, 253
- Laban, 63, 183, 211
Labat, 287
Labbi, 305
Labrum, 110
Lacasse, 240
Ladhak, 47
Ladia, 219
Lahabi, 90
Lai, 120, 141, 173, 252, 327, 351, 383
Lakshmanan, 206
Lal, 65
LAM, 203
Lam, 46, 84, 187, 251, 269, 302, 303
Lamba, 171, 223
Lan, 73, 109, 175, 192, 208, 209, 283, 342, 359, 377
Lanchantin, 304
Land, 80
Lange, 62, 240
Langis, 173
Langlais, 129
Lango, 176, 274
Lapastora, 76
Lapata, 46, 70, 210, 251, 257, 265
Lapesa, 296
Laputin, 356
Larionov, 407
Larson, 78
Laskar, 73, 169, 172, 306
Lastras, 57
Latapie, 225
Latecki, 213
Latifah, 280
Latouche, 141
Lau, 112, 278
Laurito, 128
Lauscher, 59, 259, 295
Lauvrak, 164
Lavania, 151
Lavie, 302
Lavrak, 74
Layek, 357, 365
Le, 41, 113, 114, 186, 286, 287, 310, 390, 401
- Leach, 78
Lease, 152
Lederman, 255
Lee, 45, 47, 53, 54, 60, 61, 77, 82, 89–91, 95, 96, 100, 101, 103, 104, 107, 111, 120, 133, 136, 137, 141, 142, 144, 149, 164–166, 175–178, 180–182, 185, 192, 196, 199, 204, 207, 208, 216, 222, 223, 228–231, 233, 236, 241, 242, 247, 250, 252, 260, 266, 267, 269, 273, 276, 277, 283, 289, 290, 294, 297, 303–307, 314, 317–319, 323, 325, 327, 328, 335, 339, 344, 353, 355, 358, 369, 382, 409
- Leemann, 298
Legaspi, 302
Lehmann, 145
Lei, 147, 150, 167, 244, 262, 300, 346, 355, 386
- Leippold, 379
Leiter, 387
Lemon, 245, 317
Leng, 365
Leong, 142
Leordeanu, 217
Lepri, 48
leqingchen, 388
Lerman, 151, 295
Leroy, 212
Lerzer, 166
Lesci, 196
Lesot, 76
Leurent, 279
Levi, 153
Levin, 144
Levine, 116, 152
Levy, 54, 123, 234, 260
Lhoneux, 66, 75
LI, 61, 96, 107, 275, 297, 321, 346, 348, 375
Li, 42, 45, 50, 51, 53–58, 60, 65, 67, 68, 70, 76, 79, 82–88, 90, 92, 93, 100, 101, 105–108, 110–115, 120–122, 124–126, 131–137, 141, 142, 144, 148, 154, 155, 157, 159–164, 166, 168, 172, 175, 177, 179–181, 183–188, 190, 192, 199, 200, 202–205, 208, 209, 213, 216–218, 220, 222, 227, 228, 230–235, 238, 239, 241, 242, 244, 246, 248, 251, 254, 258, 260, 267–272, 275, 276, 278–282, 284, 285, 289–291, 294, 297, 299, 300, 302, 304, 309, 311,

- 312, 315, 317, 318, 320–323, 325,
327, 330, 332, 333, 335–338,
340–346, 349–351, 353–355,
357–359, 361, 363–368, 370–372,
374–377, 379–383, 385, 386,
388–395, 398, 401, 403–408
- li, 344, 391, 397
Liakata, 64, 71
LIAN, 122, 245
Lian, 124, 155, 351, 378
Liang, 82, 103, 111, 112, 137, 139, 161, 171,
182, 187, 207, 244, 246, 249, 251,
263, 284, 289, 301, 309, 322, 339,
346, 347, 356, 366, 382, 386, 393
- LIAO, 277
Liao, 68, 76, 112, 141, 155, 171, 181, 190,
197, 226, 245, 246, 280, 301, 311,
338, 347, 378, 391, 404
- liao, 337
Liaoshengyi, 51
Libov, 233
Libovický, 117
Liermann, 204
Liew, 344
LIM, 148
Lim, 47, 119, 176, 229, 230, 232, 310, 314,
327, 342, 344, 347, 353
- lim, 379
Lima, 78, 307
LiMingDa, 341
Limišiewicz, 148
Limkonchotiwat, 142, 143, 358, 388
Lin, 47, 68, 76, 77, 82, 83, 87, 88, 91, 93, 95,
96, 105, 107, 111, 112, 115, 116,
124, 126, 130, 133, 136, 138, 154,
155, 163–165, 170, 179, 186, 196,
199, 201, 203, 206, 218, 228, 229,
231, 233, 241, 243, 245, 247, 252,
254, 277–279, 282, 285–287, 297,
299, 306, 309, 311, 315–319, 322,
324, 331, 334, 337, 341, 344, 345,
347, 352, 355, 357, 361, 363, 375,
377, 379, 385, 387, 394, 398, 407
- Lindemann, 118
Lindenbaum, 402
Lindquist, 99
Ling, 134, 315
Linzen, 55
Lio, 405
Lioma, 273
Lipani, 198
Lipka, 223
Lipton, 80
- Liscio, 188
Lisevych, 354
Litman, 175
Litschko, 189, 274
Little, 307
LIU, 65, 172, 216, 228
Liu, 43, 44, 50, 51, 54, 58, 60, 62–65, 67–69,
71, 73, 76, 82, 83, 85–89, 91, 92,
94, 100, 105–113, 115, 116,
119–121, 123, 125, 126, 129–131,
134–137, 141, 143, 146–148, 151,
154, 155, 157, 159, 162–167, 170,
172, 178–183, 186, 187, 189, 191,
194, 196–199, 202, 203, 205,
207–210, 212, 214, 216, 219, 220,
225–227, 229, 231–233, 235, 237,
240, 241, 244, 246, 249, 251–253,
255, 257, 258, 261–264, 266,
268–271, 274, 276, 277, 281–283,
285, 287–293, 296, 299, 300, 305,
308, 311, 316, 317, 319–321, 323,
325, 327, 329–332, 337, 339–342,
344–349, 351–362, 365–368, 370,
372–388, 391–395, 398, 402,
405–407, 409
- liu, 51, 110, 202, 327, 344, 377
Liujianfeng, 113
Liusie, 108, 126
Livescu, 220
liyunfei, 275
Liétard, 389
Lloret, 139
Lo, 165, 168, 215, 266
Loakman, 96, 155
Lockhart, 307
Logeswaran, 273, 303, 304
Lomeli, 304
Lomshakov, 354
Long, 51, 167, 240, 259, 290, 348, 355, 364
Longpre, 164
Lopez, 57, 407
Lopo, 142
Lotan, 151
Lothritz, 204
Lou, 65, 84, 123, 207, 238, 260, 364, 367
Louie, 44
Louridas, 372
Louzoun, 209
Lovenia, 142
Low, 112, 278
LU, 115, 225
Lu, 46, 58, 62, 76, 77, 82, 83, 107, 124,
130–132, 134, 162, 168, 171, 177,

- 190, 193, 210, 240, 247, 249, 262,
269, 272, 282, 311, 316, 323, 325,
333, 336, 338, 339, 341, 344, 345,
358, 359, 366, 368, 373, 379, 380,
383, 388, 390, 392, 401
- Luan, 112, 113, 356
Lucas, 216, 322
Lucchetti, 324
Lucero, 66
Lucy, 215
Ludaecher, 43
Luger, 296
Luhtaru, 324
Lukasiewicz, 76
Lukasik, 140
Lukito, 296
Lum, 259
Lundberg, 91
Luo, 49, 51, 61, 82, 92, 105, 108, 134, 154,
167, 172, 173, 185, 203, 209, 233,
244, 250, 257, 258, 277, 300, 308,
318, 326, 329, 331, 344, 345, 348,
356, 362, 363, 377, 383, 384, 392,
402
- luo, 83
Lutz, 259
Luu, 87, 102, 162, 167, 290, 349, 383, 408
Lv, 135, 193, 237, 238, 249, 357, 366, 386, 392
Lymperaiou, 86, 283
Lyu, 66, 114, 150, 161, 174, 197, 203, 286,
336, 373, 375, 399
- Ma, 43, 51, 58, 68, 73, 92–94, 103, 107, 109,
114, 129, 131, 134, 137, 151, 155,
161, 173, 187, 188, 197, 205, 208,
209, 220, 233, 246, 274, 277, 278,
281, 284–286, 305, 307, 320, 321,
343, 345, 351, 363, 364, 380, 385,
386, 394, 398, 400, 402
- Maaløe, 41
Mabrey, 116
MacAvaney, 53
MacDonald, 331
Macina, 225
Mackie, 53
Macko, 216
Madaan, 261
Madabushi, 70, 255
Madan, 232
Madhusudan, 52
Madhusudhan, 265
Madikeri, 221, 300
Madotto, 91, 218
- Maduabuchi, 339
Magai, 380
Magalhaes, 304, 316
Magdy, 221, 287
Magnusson, 290
Magooda, 175
Mahadevan, 378
mahajan, 254
Maharaj, 82, 153
Mahendra, 142, 329
Maheshwari, 267
Maheshwary, 265, 357
Mahfouz, 305
Mahmood, 245
Mahmoud, 82
Mahmud, 337
Mahongxia, 367
Mahowald, 55, 128, 255, 297
Mai, 408
Maimaiti, 375
Maimon, 237
Maistro, 41, 273, 359
Maiti, 220
Maity, 266, 294
Maiya, 128
Majumder, 57
Mak, 370
Makhlof, 354
Makinae, 45, 157
MAKOUAR, 221
Malagutti, 99
Malaviya, 70, 212
Malfa, 163
Malgaroli, 399
Malik, 57, 87, 341
Malin, 53, 57
Mallick, 164
Malmasi, 89
Malon, 254
Malviya, 234
Malwat, 286
Mamo, 172
Mamta, 59
Man, 154, 336
Manakul, 143, 358
Manatkar, 352
Mancenido, 194
Manchanda, 326
Mandal, 281
Mandelkern, 55, 72
Manerba, 60
Mangla, 60
Manikantan, 116

- Manmatha, 378
Manning, 49
Manocha, 58, 77, 245, 246, 315
Manohar, 371
Mansurov, 82, 142
Manzoor, 130, 298
Mao, 87, 92, 104, 122, 136, 225, 283, 286,
 289, 293, 323, 330, 342, 344, 353,
 360, 364, 367, 370, 390, 391
Marasovic, 195
Marchal, 259
Marchant, 272
Marchisio, 75, 143, 293
Marchitan, 142
Marciniak, 295
Marco, 79, 117
Marcotte, 202
Marculescu, 337
Marjanovic, 273
Markham, 99
Markov, 301, 368
Markus, 153
Marques, 234
Marraffini, 260
Marrese-Taylor, 247
Marro, 163
Martin, 184, 318
Martinez, 196, 236, 374
Martino, 169
Martins, 66, 155, 158
Marttinen, 245
Martín, 379
Martínez, 265
Marxer, 219
Masala, 217
Maskey, 307
Maslenkova, 240
Masoudian, 349
Masry, 172
Mass, 252
Massiceti, 101
Masud, 295
Masullo, 296
Matassoni, 167
Mateo-Girona, 79
Mathur, 78, 217, 223, 245, 246
Matos, 370
Matsuda, 363
Matsunaga, 226
Matsuo, 238
Matsuyama, 99
Matthes, 120
Matton, 286
Maturi, 123
Maurer, 296
Mavlyutov, 77
Mavromatis, 199
May, 60, 153, 212
Mazurek, 44
Mazzacara, 228
Mazzei, 266
McAuley, 174, 189, 223
McCabe, 76
McCallum, 86
McCoy, 274
McCurdy, 97
McDonald, 52
McDuff, 337
McGhee, 220
Mcgranaghan, 307
McKeown, 47, 71, 265
McKinzie, 55
Mcmanah, 230
McNichols, 208
Mcquade, 399
Meagher, 113
Mediratta, 114
Medya, 294
Meer, 188
Mehrab, 337
Mehrabi, 278, 280, 281
Mehrabian, 307
Mehrafarin, 269
Mehta, 88, 152, 279, 281, 282, 307
Mei, 230, 388
Meichanetzidis, 370
Meiri, 55
Meisenbacher, 120
Meister, 98
MEIYING, 170
Mejia, 71
Mekala, 201, 304
MEKKI, 221
Melamed, 76
Mellor, 259
Melo, 387
Melvin, 64
Men, 191
Menapace, 101
Mendes, 78, 185, 186
Mendonça, 302
Meng, 61, 79, 85, 106, 155, 190, 219, 255,
 270, 276, 278, 308, 310, 327, 331,
 358, 367, 401, 405
Menghini, 88, 103
Menon, 104, 130, 140, 299

- Menzat, 75
Mercer, 314
Merdjanovska, 214
Merlin, 98
Merrill, 41, 286, 288
Merugu, 356
Mesgar, 91
Metcalf, 241
Meulemans, 68
Meyer, 200
Meyers, 399
Mezentsev, 138
Mgbahurike, 222
Mi, 112
Miao, 46, 50, 354, 359
Miaschi, 163
Michel, 298
MICHELIN, 283
Michi, 279
Michieli, 261
Micklem, 405
Mickus, 144
Miehling, 348
Mihalcea, 123, 145, 174, 189
Mijovic, 326
Mikaelyan, 68
Mikhailov, 82, 387
Mikhalev, 138
Miksovic, 305
Milan, 77
Milani, 110
Milchevski, 91
Mileti, 256, 272
Miller, 285, 322
Mills, 252
Min, 65, 159, 213, 230, 237, 254, 279, 324,
 345
Minakov, 77
Minervini, 42, 269, 321
Ming, 175, 373
Mingeong, 332
Minhas, 45
Minixhofer, 178
Minkov, 151
Miranda, 142
Mire, 49, 71
Miresghallah, 213
Miret, 210
Mironova, 397
Mirylenka, 305
Mishaeli, 74
Mishra, 47, 104, 162, 184, 227, 253
Misra, 55, 128
Mita, 80, 256
Mitchell, 276
Mitra, 68, 71, 78, 299
Mittal, 375
Mitts, 154
Miyao, 128, 247, 254
Mo, 43, 92, 309, 382
Modarres, 361
Modarressi, 313
Modas, 258
Modi, 151, 285
Mody, 184
Modzelewski, 169, 339
Moeini, 168
Moens, 150, 197
Mohamed, 221, 282
Mohammad, 99, 168
Mohammed, 170
Mohan, 90, 130
Mohankumar, 232
Mohanty, 49
Mohapatra, 225
Molfese, 250
Monath, 86
Mondal, 135, 284, 325
Mondorf, 211
Monfort, 165
Moniz, 142
Montalan, 142
Montanelli, 406
Montariol, 398
Monteiro, 202, 334
Montgomery, 100
Monz, 45, 116, 147, 171, 181
Moon, 91, 152, 178, 207, 218, 223, 327
Mooney, 65
Moore, 64, 218, 298
Morabito, 52
Moraffah, 202
Morariu, 245
Moreira, 217, 283
Morency, 47, 103, 246
Morgado, 75
Morimura, 140
Moro, 110, 216
Morris, 40, 44
Morrison, 182, 313
Mortensen, 117
Moryossef, 186, 314
Mosbach, 124
Moschitti, 161, 212
Moskvoretskii, 157
Moslem, 142

- Motlicek, 221, 222, 300
Mou, 184, 214
Mouchtaris, 177
Mousavi, 216, 231
Movva, 187
Mozafari, 200
Mu, 361, 378, 396
Mudhiganti, 231
Mueller, 147
Muennighoff, 142
Muhamed, 408
Mujahid, 82, 171
Mukherjee, 88, 151, 261, 357, 403
Mukhopadhyay, 319
Mukkavilli, 307
Mukku, 56
Mullappilly, 327
Muller, 77
Mullick, 313
Mullov, 166
Mun, 215
Munakata, 153
Munawar, 57
Munch, 217
Munoz, 140
Muppalla, 78, 310
Murali, 259
Muraoka, 63, 307
Muresan, 254
Murphy, 110
Murray, 144
Murugesan, 91, 139
Murukannaiah, 188
Musa, 64
Mustafa, 254
Muti, 296
Myers, 322
Möller, 162
Müller, 314
- Na, 267, 277, 290, 375
Nabih, 167
Nag, 134, 403
Nagarajan, 91
Nagata, 46
Nagireddy, 231, 348
Naglik, 176
Nagoudi, 221
Nagrani, 316
Naguib, 283
Nahid, 328
Nahrstedt, 102
Naidu, 325
- Naik, 165, 188
Nair, 46, 100, 339
Naito, 174, 175
Nakada, 153
Nakagi, 99
Nakamura, 78, 282, 371
Nakashima, 122, 258
Nakov, 82, 130, 171, 245, 257, 294, 298
Nam, 152, 229
Name, 223
Nan, 90, 170, 251, 269, 345
Nandi, 44, 294, 354
Nandy, 168, 313
Naous, 74
Narasimhan, 140
Narayan, 408
Narayana, 304
Narayananam, 352
Narsupalli, 375
Naseem, 380
Naszadi, 116
Natarajan, 222, 267, 325
Nath, 116
Navigli, 156, 250
Nayak, 54, 319
Nayeem, 172, 187
Nayeri, 328
Nechaev, 199
Nediyanchath, 307
Neelam, 57
Negri, 123, 167
Nejadgholi, 52
Nenkova, 400
Neo, 194
Neubig, 64, 101, 144, 164, 250, 256
Neville, 264
Newell, 117
Newman, 165
Ng, 102, 112, 113, 161, 211, 302, 322, 370,
 406
Nghiêm, 174, 258, 284
Ngo, 154, 310, 401
Ngomo, 156
Nguyen, 41, 53, 75, 77, 88, 89, 102, 109, 121,
 154, 195, 211, 222, 252, 274, 277,
 278, 297, 310, 322, 334, 337, 401,
 403
Nguyen-Son, 380
Ngweta, 266
Ni, 51, 170, 235, 325, 333, 375, 379
Nickel, 326
Nicolai, 117
Nie, 43, 92, 236, 244, 289, 321, 347, 354

- Niehues, 166, 220
Nieto, 58
Nigam, 134, 239
Nigatu, 43
Nikandrou, 245
Nikishina, 157
Niklaus, 174
Nikolaev, 52
Nikolenko, 354, 380
Nikoulina, 250
Nilizadeh, 185
Ning, 190, 243, 257, 311, 331
ning, 82
Nirunwiroy, 298
Nisar, 257
Nishida, 239
Nishimoto, 99
Nishimura, 153, 302
Nithyanand, 64
Niu, 57, 63, 85, 112, 187, 224, 241, 278, 309, 332, 399
Niwa, 166
Nixon, 142
Noble, 213
Noel, 202
Noh, 50, 279
Noriega-Atala, 313
Noronha, 353
Norouzian, 90
Nourbakhsh, 285, 313
Nouri, 128
Novak, 75
Novovi, 91
Nowak, 99, 118
Nozza, 59
Nutanong, 143, 358
Nwogu, 146
Nygaard, 353
Névéol, 283
O'Brien, 174
O'Connor, 64, 118
O'Neill, 151
Oba, 56, 363
Oberst, 80
Ocampo, 72
Ogren, 249
Ogueji, 265
Ogundepo, 285
Oguz, 313
Oh, 50, 58, 90, 96, 169, 278, 301, 303, 323, 382
Ohmer, 72
Ohrimenko, 272
Ok, 290
Oka, 80
Okazaki, 117, 266
Okotore, 78
Okumura, 200
Olaru, 217
Oliehoek, 116
Oliveira, 234
Olson, 113, 171
Omar, 104
Omraní, 300
Ong, 207, 264, 315
Onizuka, 296
Onoe, 100
Oosterhuis, 332
Opedal, 96, 100
Opitz, 253
Oprea, 74, 121
Opsahl-Ong, 80
Oraby, 88
Orasan, 45, 91
Ordonez, 105
Oren, 129
Oresko, 399
Orlando, 250
Ornelas, 109
Ortega, 278
Oseki, 56
Oseledets, 138
Ossowski, 303
Ostermann, 162, 313
Otterbacher, 339
Ou, 330
Ouali, 374
Ouchi, 56
Ouyang, 98, 183, 184, 186, 292, 320, 328, 387
OuyangRuyi, 109
Ovadia, 74
Oved, 190
Overney, 49
Oveson, 326
Oyama, 268
Oyamada, 309
Ozay, 121, 261, 279
Ozdaglar, 150
Ozkan, 279
Paassen, 48
Pack, 84
Padhi, 231
Padro, 90
Padó, 52, 296
-

- Paganelli, 116
 Pai, 312
 Pakazad, 250
 Pal, 158, 202
 Palande, 169
 Palioras, 93
 Pallarés, 90
 Palmer, 144, 336
 Palo, 177
 Palta, 288
 Pal tenghi, 357
 Pan, 84, 105, 108, 110, 115, 122, 130, 135,
 136, 159–161, 167, 171, 194, 205,
 213, 245, 251, 273, 286, 297, 345,
 349, 352, 355, 366, 367, 370, 388,
 393, 395, 396, 408
 Panchenko, 138, 407
 Panda, 57
 Pande, 199
 Pandey, 184, 297
 Pandita, 296
 Pandya, 195
 Pang, 43, 249, 330, 361
 Panov, 114
 Panswan, 142
 Pantazopoulos, 245
 Pantha, 307
 Paoli, 294
 Papadimitriou, 305
 Papakyriakopoulos, 258
 Papangelis, 277
 Papi, 123, 167
 Papiez, 76
 Papotti, 50, 110
 Papoudakis, 210
 Pappas, 49
 Parameswaran, 346
 Parasol, 172
 Parby, 294
 Parde, 66
 Parekh, 52, 104, 115, 310, 313
 Pari, 279
 Parikh, 147
 Paris, 211
 Parisien, 135, 226
 Park, 47, 49, 53, 70, 89–91, 103, 107, 119,
 139, 148, 149, 162, 170, 177, 178,
 181, 187, 191, 204, 205, 208, 216,
 227, 229–232, 247, 252, 280, 289,
 297, 299, 304, 305, 307, 310, 312,
 318, 323, 327, 345, 369
 Parmar, 73, 78, 215, 282
 Parmonangan, 142
 Parthasarathi, 138
 Paruchuri, 337
 Parvez, 73, 169, 266
 Pasti, 118
 Patankar, 240, 396
 Patel, 78, 88, 183, 219, 265, 282
 Patil, 76, 165, 180
 Patki, 231, 400
 Patti, 287
 Pattnaik, 229, 265
 Patwa, 168
 Patwary, 73, 214
 Paul, 207, 357
 Paula, 283
 Paulheim, 234
 Pavlick, 145
 Pavlopoulos, 372
 Pavlov, 397
 Pavlova, 354
 Payani, 285
 Payoungkhamdee, 143
 Pechenizkiy, 284, 286
 Pecher, 196, 292
 Pedram, 200
 Pegueroles, 259
 Pei, 85, 88, 211, 273, 404
 PeiguangLi, 131
 Pejovic, 339
 Pelayo, 90
 Peller-Konrad, 166
 Pelles, 146
 Peloquin, 221
 Penamakuri, 104
 Pendzel, 151
 Peng, 44, 49, 60, 67, 75, 78, 79, 85, 88, 100,
 106–108, 115, 134, 135, 142, 153,
 164, 170, 182, 187, 189, 197, 198,
 220, 224, 246, 247, 280, 282, 296,
 300, 309, 310, 312, 333, 341, 351,
 355, 360, 366, 400
 Pengpun, 388
 Penn, 119
 Penzo, 48
 Peper, 175
 Perera, 219
 Perez, 265
 Perez-Lebel, 140
 Perez-Ortiz, 262
 Peris, 261, 278
 Periti, 340, 406
 Perlitz, 218
 Pernes, 185
 Perrault, 137

- Perrella, 156
Persad, 280
Pesaranghader, 62, 306
Peteleaza, 288
Petricek, 239
Petry, 71
Pezzelle, 247
Pfeffer, 298
Pfister, 93, 236
Pfitzmann, 307
Pfrommer, 42
Pham, 218, 267, 401, 403
Phan, 109
Phimsiri, 388
Phogat, 384
Phukan, 79, 246
Pi, 102, 122, 136, 160, 232, 245
Picot, 156
Pieri, 327
Pierson, 187
Pietquin, 276
Pikuliak, 399
Pilehvar, 361, 409
Pillai, 171, 371
Pillutla, 74
Pimentel, 98, 240, 256, 267
Pimparkhede, 263
Pino, 77
Pinter, 61
Piontkovskaya, 380
Piorkowski, 348
Piotrowski, 303
Piper, 71, 150
Pizzati, 261
Plank, 188, 189, 211, 274
Plaza-del-Arco, 294
Ploeger, 75
Podivilov, 354
Podolak, 303
Poelman, 75
Poh, 215
Poli, 219
Pombal, 158
Pond, 217
Ponomareva, 239
Poole-Dayan, 49
Poon, 211, 244
Poovendran, 85
Pop, 311
Popa, 180, 235
Popescu, 217
Popov, 344
Popovic, 67
Popuri, 77
Porat, 190
Poria, 344, 365, 383
Porter, 231
Post, 186
Potdar, 45, 231, 233
Pothong, 174, 175
Potnis, 149
Potter, 120, 371
Potthast, 236
Potts, 80, 127, 188, 231, 400
Potyka, 328
Pound, 231
Pour, 160
Pouransari, 188
Pouw, 256
Prabhakar, 274
Prabhakaran, 152
Prabhumoye, 73, 214
Pradeep, 231, 345
Prajapati, 78, 257
Pramanik, 43
Pratama, 345
Prato, 138
Press, 212
Preum, 67
Priebe, 148
Prindle, 258
Proietti, 156
Prud'hommeaux, 147
Pruthi, 237, 273, 304
Pryzant, 397
Pu, 208, 292, 349
Pucci, 144
Puerto, 248
Pujari, 48
Puranam, 384
Puranik, 57
Purason, 324
Puri, 399
Purohit, 338
Purtell, 80
Purushothama, 118
Purver, 117
Purwarianti, 142, 148
Putnikovic, 91
Putra, 142
Putri, 169
Pyarelal, 313
Pütz, 48
Qazi, 180
-

- Qi, 65, 67, 101, 103, 135, 140, 159, 170, 171, 189, 217, 224, 248, 270, 287, 299, 312, 332, 354, 356, 398, 401
Qian, 45, 111, 204, 208, 338, 343, 355, 384, 392, 397, 405
Qiao, 53, 170, 271, 278, 287, 352, 366, 386
Qidwai, 319
Qin, 82, 86, 92, 101, 151, 154, 157, 159, 187, 192, 274, 309, 311, 325, 332, 337, 345, 347, 352, 365, 366, 370, 382, 384–386, 392, 394
Qing, 238, 353, 382
Qiu, 69, 73, 83, 87, 129, 133, 139, 163, 165, 175, 218, 235, 253, 262, 272, 290, 330, 333, 345, 348, 349, 354, 356, 364, 376, 383, 384, 387, 392, 398, 405, 409
Qorib, 113, 142
Qu, 156, 157, 276, 307, 314, 320, 333, 339, 354, 361, 365, 380
QUAN, 67, 72
Quan, 134, 288, 363
quezibing, 216
Quick, 64

Rabatin, 111, 357
Rabinovich, 389
Radev, 51, 170, 287
Radhakrishna, 203
Radharapu, 63
Raffel, 155
Rafiei, 187, 328
Ragan-Kelley, 292
Ragazzi, 110
Raghu, 57, 184, 224, 301
Raghuraman, 307
Raha, 240
Rahaman, 70
Rahamim, 178
Rahimi, 94
Rahman, 73, 172, 266
Rahmati, 118
Raileanu, 304
Raina, 108, 126, 128, 220
RAJ, 168
Raj, 261
Rajabzadeh, 400
Rajadesigan, 152
Rajagopal, 120
Rajan, 272
Rajmohan, 83, 357
Rajtmajer, 120
Ram, 49

Ramachandran, 307
Ramakrishna, 49, 278, 384
Ramakrishnan, 394
Ramamoorthy, 64
Ramamurthy, 231
Ramanathan, 208
Ramasamy, 346
Ramasubramanian, 85, 307
Rame, 279
Ramesh, 239, 261
Ramus, 307
Ramponi, 71
Ramu, 194, 215, 319
Ran, 137, 276, 308
Ranaldi, 144, 197
Ranasinghe, 45, 296
Ranathunga, 65
Rando, 125
Rangappa, 222
Ranger, 379
Rangwala, 199
Rani, 218
Ranjit, 71
Rankel, 288
Rao, 54, 342, 351, 362, 378, 379, 390
Rashkin, 46
Rassim, 233
Rasteh, 90
Rastogi, 237
Rathore, 147
Rau, 250
Ravanelli, 270
Ravfogel, 53
Ravi, 149
Rawal, 49, 118, 325
Rawat, 86
Ray, 61, 70, 105, 205, 208, 264
Raychev, 147
Raza, 284, 378
Razniewski, 62
Rebedea, 135, 217, 226
Recknor, 318
Reddy, 54, 61, 93, 141, 171, 222, 233
Reed, 408
Reeson, 211
Reforgiato, 216
Rei, 155, 158, 269
Reich, 97
Reichart, 40, 148, 190, 318
Reichman, 235
Reisert, 174, 175
Reiter, 216
ReiSS, 166

- Rekabsaz, 349
Ren, 43, 51, 90, 92, 114, 159, 161, 163, 190,
198, 206, 211, 229, 269, 302, 323,
362, 364, 368, 386
Renduchintala, 241
Renze, 331
Rep, 255
Rezagholizadeh, 92, 138, 285, 400
Rezapour, 150
Ribeiro, 212, 265
Riboni, 216
riccardi, 116, 216
Ricci, 316
Rice, 71, 144
Richards, 348
Richardson, 166, 308
Riddell, 51, 170, 287
Riedhammer, 177, 311
Rieser, 259
Rijke, 158, 161, 171, 360
Rimchala, 170
Rippeth, 186
Rish, 319
Risher, 63
Ritter, 78, 186, 234
Rizk, 57
Rizzoli, 116
Ro, 103, 106, 179, 221
Robbani, 174, 175
Roberts, 218
Robey, 86
Robinson, 398
Roccabruna, 116
Rodemann, 267
Rodriguez, 259
Roemmle, 239
Rogers, 298
Rohanian, 134
Romary, 225
Romero, 302
Rondeau, 350
Rong, 308, 361, 392
Roosta, 325
Ropers, 77
Rosati, 277
Rose, 313
Rosenberg, 132, 250
Rosenthal, 61
Rosman, 43
Rossi, 215, 400
Roth, 164, 195, 198, 217, 272, 274, 302, 308,
319
Rothkopf, 254
Roukos, 57
Roukus, 61
Roy, 40, 49, 135, 235, 240, 265, 293, 314
Roy-Chowdhury, 105
Rozner, 402
RRV, 78
Ru, 345, 348
Ruan, 353, 365, 387
Ruas, 184, 192, 230
Rubashevskii, 171
Rubin, 41
Rubinstein, 272
Rudenko, 138
Ruder, 40, 142, 143, 293
Rudinger, 119, 288, 295
Rudzicz, 277, 314
Ruggeri, 296
Ruitter, 303
Runwal, 137
Ruotsalo, 41
Ruppenhofer, 395
Ruseti, 253
Rush, 164, 290
Rustagi, 121
Ruths, 150
Ryan, 74, 80
Ryanda, 142
Ryu, 111, 189, 290
Rühle, 357
S, 77, 221, 222, 232
Saad-Falcon, 400
Saadany, 91
Saadia, 221
Sabharwal, 166
Sabour, 355
Sachan, 139, 174, 189, 225, 236, 271
Sachdeva, 229
Sadagopan, 280
Sadeq, 223
Sadhu, 306
Sae-jia, 388
Saenger, 295
Sagot, 77, 352
Saha, 184, 225, 281, 285, 293, 301
Sahabandu, 85
Sahai, 78, 289
Sahoo, 357
Sahu, 66, 229, 244, 346
Saikh, 246
Sailor, 170
Saito, 239
Sajedinia, 48

- Sajjad, 128, 277
Sakaguchi, 364
Sakai, 45, 92, 114, 157, 387
Sakamoto, 140
Sakhovskiy, 195
Sakshi, 58, 77
Salakhutdinov, 103
Saley, 224, 301
Salim, 70
Salinas, 193
Salisbury, 186
Salman, 240
Salto, 71
Samdarshi, 254
Sameti, 118
Samir, 149
Sanchetti, 119
Sanchez, 169
Sanders, 245, 253, 307, 318
Sandhan, 61
Sang, 231
Sanjabi, 101
Sankararaman, 57
Sankowski, 303
Sanner, 62, 160
Sanochkin, 369
Sant, 314
Santamaría, 301
Santos, 79
Santoso, 142
Sanyal, 196
Sap, 49, 71, 227
Saparov, 125, 192
Saphra, 120, 178, 314
Saralajew, 273
Sarahthy, 303
Sarfati, 67
Sargsyan, 216
Sarikaya, 125
Sarkar, 169
Sarti, 248
Sassano, 353
Satapara, 143
Satapathy, 201
Sataudom, 388
Sathe, 261
Sathvik, 343
Sati, 229
Satpathy, 110
Sattigeri, 231, 348
Satzoda, 378
Saunders, 155
Savchenko, 195
Savin, 354
Savkin, 344
Savoldi, 123, 259
Savov, 169
Sawhney, 245
Saxena, 153, 194, 293
Saxon, 215
Scardapane, 42
Scarlatos, 109
Scells, 236
Schedl, 178, 349
Scheible, 139
Schein, 271
Schick, 95
Schiller, 64, 173
Schimanski, 379
Schirmer, 298
Schlangen, 265
Schlechtweg, 213, 406
Schlegel, 248
Schlichtkrull, 329
Schlötterer, 195
Schmid, 132
Schmidt, 61, 145, 278
Schnabel, 264
Schneider, 97, 214, 320, 326
Schockaert, 254
Schoelkopf, 170
Schoene, 102
Schottmann, 146
Schrader, 62
Schramowski, 134
Schröder, 368
Schuetze, 95, 146, 199, 240, 271, 284, 313
Schuler, 96
Schulte, 198
Schwab, 407
Schwartz, 129, 131
Schwettmann, 105
Schölkopf, 174, 189, 214
Scivetti, 116
Scotton, 305
Sedova, 274
Seeberger, 177, 311
Segonne, 144
Seifert, 195
Seifi, 174
Sek, 305
Seleznyov, 407
Selfridge, 56
Semedo, 304, 316
Semnani, 187, 251, 302
Senel, 284

- Sennrich, 146, 314
Seo, 162, 164, 175, 185, 191, 230, 233, 238,
 266, 315, 324
Seol, 277
Seonwoo, 301
Serai, 147
Serapio-García, 397
Sert, 181
Seshadri, 165, 369
Sesia, 136
Seth, 58, 77
Sethi, 343
Seto, 241
Sewunetie, 340
Seyssel, 219
Sha, 68, 85, 224
Shah, 65, 97, 296, 298, 305, 308, 313, 325
Shahaf, 306
Shahariar, 318
Shahi, 84
Shahid, 258
Shaib, 84
Shaier, 249
shaiik, 394
Shalumov, 190
Shan, 237, 397
Shang, 113, 168, 197, 198, 208, 225, 233, 282,
 286, 304, 315, 338
Shankar, 55
Shao, 187, 304, 323, 338, 358, 400, 404
Shapiro, 164, 318
Shareghi, 285, 339, 342
Sharif, 67
Sharma, 57, 136, 174, 193, 212, 215, 231, 253,
 294, 306, 314
Sharpnack, 63
Shashidhar, 289
Shatnawi, 221
Shayanfar, 326
Shayegani, 105
She, 351
Shea, 44, 227
Sheafer, 153
Shehata, 221
Shehu, 133
Shekkizhar, 278
Shelby, 63
Shelmanov, 82
Shen, 49, 51, 58, 76, 78, 83–85, 110, 121, 123,
 130, 134, 148, 154, 180, 188, 193,
 196, 199, 204, 211, 216, 224, 226,
 243, 250, 274, 287, 299, 305, 323,
 327, 332, 336, 339, 343, 351,
 361–365, 381, 392, 405, 406
shen, 292
Sheng, 79, 126, 211
Shengelia, 403
Shenhav, 153
Shenoy, 105
Sherborne, 286, 290
Sheth, 128, 218, 257
Shetty, 285
SHI, 225
Shi, 51, 52, 64, 84, 93, 111, 112, 136, 151,
 159, 161, 203, 207, 208, 220, 223,
 240, 241, 244, 279, 302, 315, 317,
 322, 327, 333, 351, 354–357, 363,
 365, 375, 376, 380, 381, 387, 392,
 393, 406
Shieh, 166, 237
Shifat, 201
Shilton, 44
Shim, 130, 179, 317, 375
Shimabucoro, 40
Shimodaira, 268, 274, 313
Shimomoto, 247
Shin, 91, 132, 138, 164, 177, 189, 247, 266,
 283, 297
Shiran, 306
Shiri, 105
Shiwakoti, 296
Shiwei, 353
Shmatikov, 40
Shoeybi, 73, 214
Shopov, 73
Shou, 160
Shrimal, 307
Shrivastava, 217, 246
Shtok, 131
Shu, 85, 86, 225, 249, 278, 290, 308, 328
Shubi, 55
Shui, 326
Shukla, 285
SHUM, 378
Shwartz, 149
Si, 335, 337, 345, 383
Siangliulue, 165
Siarohin, 101
Sibue, 285, 305
Sick, 214
Sidat, 350
Siddique, 123
Siderius, 150
Sidorov, 356
Sieker, 270

- Sikka, 244
Sil, 61, 93
Silcock, 294
Sileo, 390
Silfverberg, 117
Silva, 158, 311
Sim, 278
Simko, 216, 292, 399
Simoulin, 181
Sinapov, 339
Sindhgatta, 281, 307
Sindhujan, 45
Singh, 63, 71, 77, 113, 146, 151, 169, 171,
 188, 191, 219, 229, 235, 257, 265,
 268, 273, 286, 295, 305, 311
singh, 69
Singha, 203
Singhal, 88
Singhi, 177
Singla, 135, 147
Sinha, 89, 93, 184, 230
Siriwardhana, 399
Sitaram, 151, 159, 165, 261, 320
Siu, 223, 375, 400
Sivek, 63
Skachkova, 313
Skorokhodov, 101
Skorupska, 339
Skotti, 353
Slavutsky, 269
Slobodkin, 237
Small, 125, 265
Smith, 41, 91, 109, 121, 133, 142, 148, 182,
 408
Smolensky, 97
Smyth, 191
Smdu, 311
Snajder, 255
Snæbjarnarson, 271
Soares, 160, 203
Soatto, 378
Sodhani, 138
Soedarmadji, 152
Soga, 208
Soh, 310
Sohn, 117, 304
Sohoney, 357
Sojoudi, 42
Solawetz, 400
Soldaini, 215
Solihin, 123
Solorio, 43
Sommerauer, 55
Son, 208, 289
Sone, 364
SONG, 380
Song, 49, 57, 62, 64, 67, 70, 92, 95, 96, 100,
 105, 107, 120, 121, 126, 138, 139,
 147, 155, 157, 170, 172, 175, 183,
 185, 197, 204, 207, 209, 210, 244,
 249, 253, 273, 284, 297, 305, 315,
 338, 344, 345, 351, 355, 360, 361,
 381, 384, 386, 390, 396
Soni, 375
Sonkar, 241, 287, 325
Soremekun, 272
Sorensen, 187
Soria, 57
Soricut, 100
Soroa, 285
Sorodoc, 251
Sosnowski, 339
Sotnikova, 398
Sotudeh, 183
Soulos, 97
Sourati, 55, 273
SOUROVE, 201
Sousa-Silva, 213
Souza, 99
Soylu, 80
Spangher, 44, 60, 153
Spilsbury, 245
Sra, 268
Srba, 196, 216, 292
Sreedhar, 57, 135, 226
Sreekar, 280
Srihari, 225
Srijith, 143
Srikumar, 124
Srinath, 65, 120
Srinet, 91
Srinivasa, 225, 233
Srinivasan, 64, 70, 88, 194, 199, 222
Srivastava, 47, 129, 130, 299
Staab, 328
Staar, 307
Stacey, 269
Staerman, 156
Stallone, 57
Stamatiou, 98
Stammbach, 321
Stamou, 86, 283
Stanczak, 54, 60, 72
Stanković, 110
Stanojević, 98
Stanovsky, 234, 259

- Steedman, 395
Steenkiste, 54
Stein, 176
Steinert-Threlkeld, 147
Stella, 398
Stenetrop, 76, 146, 396
Stepputtis, 279
Sternbentz, 84
Sternier, 178
Stevenson, 296
Stewart, 295
Steyvers, 191
Stiefelhagen, 166
Stigt, 158
Stoehr, 271
Stolcke, 57
Stone, 140
Storks, 316
Stranisci, 287
Stranjanac, 91
Stripelis, 305, 308
Strohmaier, 259
Strong, 209
Stroud, 296
Strub, 276
Strubell, 279, 290
Strötgen, 240
Studdiford, 298
Sturman, 63
SU, 327
Su, 46, 65, 116, 119, 164, 185, 186, 216, 225,
 227, 241, 274, 277, 281, 314, 315,
 320, 335, 358, 374, 377, 391, 395,
 400, 409
Suarez-Tangil, 329
Subakan, 270
Subbiah, 47
Subramanian, 88, 142, 279, 315
Subramanyam, 152
Subramonian, 119
Such, 329
Suchanek, 140
Sudoh, 371
Sugawara, 56, 193
Sugiyama, 315
Suglia, 47, 104, 245
Suh, 152, 189
Suhr, 104, 142
Sui, 131, 163, 201, 291, 384, 386, 392
Suk, 164
Suleiman, 78
Sullivan, 279
Sultan, 93, 306
SUN, 113
Sun, 43, 51, 58, 63, 67, 82, 109, 119, 121, 127,
 129, 131, 133, 136, 138, 139, 142,
 144, 168, 170, 173, 177, 179, 180,
 182, 185, 186, 205, 207, 209, 215,
 217, 223, 226, 237, 238, 244, 264,
 266–269, 274, 283, 311, 314, 321,
 331, 337–339, 341, 342, 344, 345,
 350–352, 355, 356, 360, 362,
 364–366, 368, 378, 382, 389, 397,
 400, 403, 407, 409
Sundaram, 346
Sung, 87, 249, 402
Sunshine, 337
Supholkhan, 388
Supryadi, 355
Surdeanu, 288, 313
Sureshan, 82
Suri, 245
Surikuchi, 247
Surve, 281
Susanto, 142, 345
Susladkar, 375
Sutawika, 145
Suvarna, 115
Suzuki, 63, 110, 146
Svete, 86, 118, 133
Sviridova, 174
Swain, 93, 219
Swaminathan, 165
Swanson, 141
Swayamdipta, 71, 88, 278
Sycara, 279
Syed, 239, 330
Synnaeve, 77
Sypherd, 287
Szabó, 170
Szarvas, 301
Szekely, 212
Sánchez, 146
Søgaard, 54, 98, 142, 160, 219, 246
T, 394
Ta, 82
Tabrizi, 223
Tadiparthi, 279
Tadmor, 326
Tafjord, 253
Tagarelli, 255
Taha, 199
Tahaei, 400
Tahan, 209
Tahmasebi, 213, 340, 406

- Takagi, 99
 Takamura, 247, 264
 Takase, 274
 Takeuchi, 162
 Taktasheva, 387
 Talafha, 221
 Talat, 59, 119
 Talwalkar, 256
 Tam, 107, 306
 Tamilselvam, 263
 Tan, 45, 46, 51, 54, 58, 73, 86, 106, 108, 113,
 114, 128, 180, 195, 200, 202–204,
 209, 215, 233, 240, 274, 288, 320,
 325, 327, 331, 339, 353, 354, 360,
 364, 386
 Tandon, 97
 Taneja, 289
 Tang, 76, 83, 97, 104, 126, 133–135, 137, 157,
 167, 191, 204, 211, 218, 225, 231,
 246, 250, 257, 260, 275, 296, 299,
 300, 307, 311, 317, 323, 330, 331,
 337, 344, 347, 349, 354, 355, 365,
 370, 378, 383, 385, 386, 401, 404
 tang, 365
 Tangherlini, 294
 Taniguchi, 364
 Tanner, 61
 Tannier, 283
 Tanzer, 403
 TAO, 337
 Tao, 89, 95, 109, 130, 191, 351, 357, 364, 375,
 393
 tao, 336, 367
 Tapaswi, 116
 Tar, 87
 Tasawong, 358
 Taslakian, 202
 Tassiulas, 181
 Tata, 185
 Tatariya, 66
 Tater, 99
 Taubenfeld, 148
 Tavanaei, 56
 Taveekitworachai, 133
 Tavor, 389
 Tayir, 375
 Taylor, 164, 192
 Tejaswi, 146
 Tekin, 292
 Teng, 340
 Teodorescu, 99
 Terian, 217
 Terzis, 239
 Testoni, 228
 Tetreault, 171
 Thai, 131, 168
 Thakur, 134, 285
 Thaler, 262
 Thanh, 310
 Thawonmas, 133
 Thellmann, 145
 Theobald, 241
 Thirukovalluru, 86, 266
 Thomas, 212, 283, 337
 Thomason, 127
 Thomczyk, 139
 Thompson, 117, 301
 Thomson, 138
 Thorbecke, 221, 222
 Thorne, 127, 290
 Thota, 185
 Thrush, 105
 Tian, 60, 83, 95, 102, 131, 135, 154, 197, 203,
 204, 213, 220, 235, 251, 336, 353,
 354, 357, 360, 380, 386
 Tiankanon, 388
 Tiedemann, 144
 Timofte, 243
 Tippmann, 139
 Titov, 118
 Tiwari, 116, 181
 Tiyyala, 193
 Tjhi, 142
 Tjuatja, 144, 256
 Tn, 306
 Todwal, 88
 Toledo, 252
 Tolera, 242
 Tolmach, 164
 Tomar, 82
 Tomkins, 324
 Tommasone, 286
 Tonelli, 48, 71
 Toneva, 98
 Tong, 205, 206, 214, 249, 261, 262, 365, 385,
 395
 Tonglet, 150
 Toniato, 305
 Tonja, 43, 340
 Topi, 264
 Toroghi, 62, 160
 Torr, 192, 261, 282
 Torresani, 244
 Toshev, 55
 Toshniwal, 116
 tourad, 221

- Toutanova, 147
Tovar, 90
Towle, 347
Toyin, 170, 291
Tran, 217, 252, 309, 310, 401
Trancoso, 302
Trapeznikov, 110
Tredup, 56
Tresp, 246, 252, 328
Treviso, 158
Trhlík, 396
Triedman, 251
Tripathi, 229
Tripodi, 113
Tripto, 216
Trivedi, 138, 307
Troiano, 287
Trott, 57
Tsai, 260, 280, 306
Tsakalidis, 71
Tsang, 155
Tsarfaty, 237
Tseng, 190
Tseriotou, 71
Tsipidi, 99
Tsuruoka, 46
Tsvetkov, 114, 142, 149, 159, 187, 299, 372
Tsvigun, 82
Tu, 93, 112, 141, 151, 193, 225, 276, 343, 374,
 378, 386, 392
Tuan, 343
Tulchinskii, 380
Tulyakov, 101
Tupitsa, 157
Tur, 289, 302
Turcan, 71
Ture, 76
Turner, 123
Tutek, 248
Tutubalina, 195
Tuytelaars, 320
Tuzel, 188
Tyagi, 58, 78
Tzimiropoulos, 374
Uban, 142
Uchendu, 216
Uchiyama, 238
Udagawa, 63, 307
Udomcharoenchaikit, 143, 358
Uemura, 339
Umair, 303
Uniyal, 286
Unuvar, 57
Upadhyay, 353
Usbeck, 335
Ushio, 145
Utescher, 270
Uthus, 403
Utiyama, 118
Uzan, 61
Uzzi, 294
V, 338
Vacareanu, 313
Vagenas, 307
Vahidinia, 307
Vainikko, 324
Vakil, 267
Vakulenko, 251
Valdoriez, 198
Valentino, 72
Valipour, 400
Vallebueno, 49
Valter, 397
Valvoda, 86
Vandsburger, 172
Varma, 97
Varoquaux, 140
Varshney, 231, 282
Vasantha, 89
Vashisht, 184
Vashishth, 195
Vasilev, 397
Vassilvitskii, 239
Vaz, 158
Vazquez, 202
Vechev, 147
Veerakanjana, 388
Veerendranath, 111
Velasco, 142
Velicu, 217
Veloso, 169, 285
Veluri, 221
Velutharambath, 297
Vema, 285
Vemulapalli, 69, 188
Venkatesha, 116
Venkateswaran, 57
Venkit, 65, 120
Venturi, 163
Vepa, 229
Verberne, 201
Verbruggen, 113, 286
Verma, 64, 278, 280, 305
Vickers, 172

- Vieira, 99
 Vijiini, 299
 Vilares, 118
 Villata, 72, 174
VILLATORO-TELLO, 221, 222
 Villavicencio, 145
 Villegas, 66
 Vincent, 313
 Vishnubhotla, 99
 Viskov, 407
 Viswanathan, 250
 Vitsakis, 52, 104, 261
 Vittaut, 76
 Vlachos, 107, 115, 140, 209
 Vladimir, 397
 Vo, 185
 Voigt, 151
 Voisin, 286
 Voita, 129
 Volk, 318
 Volkova, 299
 Vondrick, 104
 Vosoughi, 98, 149, 238, 284, 328
 Voss, 204, 222
 Vougiouklis, 84, 251, 395
 Voznyuk, 344
 Vu, 87, 192, 194, 262, 333, 342
 Vuli, 58, 145, 178, 340, 409
 Vutla, 229
 Vyas, 84
 Vydiswaran, 293
 Vyetrenko, 169
- Wachsmuth, 176
 Wadhwa, 125, 184, 203
 Waghjale, 111
 Wagner, 177, 193, 269, 311
 Wahle, 184, 192
 Waibel, 166
 Wakhare, 76
 Walde, 99, 213, 256
 Waldis, 173, 218, 295
 Walker, 228
 Wallace, 84, 125, 127, 203
 WAN, 203, 387
 Wan, 44, 78, 87, 102, 109, 148, 182, 292, 329,
 340, 355, 363, 409
 WANG, 103, 160, 251, 388, 401
 Wang, 42, 43, 46, 47, 50–52, 54, 57–59, 62,
 63, 65, 67–70, 73, 76, 79, 82–84,
 89, 92–94, 100–102, 105–114,
 117–119, 121, 122, 124–127,
 129–133, 135–137, 140, 142, 146,
 151, 153, 155, 157, 159, 161, 162,
 164, 165, 167, 168, 171–175,
 182–184, 186–193, 195–200, 202,
 203, 205–213, 215, 218, 220,
 222–224, 226, 228, 231–234, 236,
 237, 240, 241, 244, 246, 249, 250,
 254, 255, 257, 259, 262–264, 267,
 271, 273, 275, 277–279, 281–283,
 289–293, 296, 298, 299, 302, 303,
 305, 307–310, 312–314, 316, 318,
 320, 322, 323, 325, 327, 330–332,
 334, 337, 340–356, 358, 361,
 364–368, 370–393, 396, 398–402,
 404–406
- wang, 250
 Wardle-Solano, 170
 Warstadt, 99
 Wasserblat, 385
 Wassie, 340
 Wastadt, 256
 Watanabe, 45, 56, 62, 114, 156, 157, 220, 276,
 291
 Watts, 165
 Weber, 145
 Weerasooriya, 296
 Wegmann, 252
 Wehner, 277
 Wei, 46, 65, 92, 101, 111, 155, 164, 167, 212,
 217, 299, 300, 321, 333, 336, 350,
 351, 360, 368, 370, 373, 381, 386,
 391, 392, 402
 wei, 206, 259, 376, 404, 405
 Weidinger, 259
 Wein, 253
 Weinbach, 134
 Weir, 212, 245, 253
 Weiss, 125, 322
 Weissenbacher, 118
 Weissweiler, 271
 Weixin, 379
 Weld, 165
 Welleck, 164
 Weller, 253
 Wen, 86, 111, 115, 126, 159, 193, 218, 241,
 289, 300, 315, 326, 332, 344, 347,
 349, 355, 392
 wen, 309, 346
 Wendler, 351
 Wendt, 185
 Weng, 122, 209, 356
 Wense, 118
 Wersing, 270
 Wertheimer, 307

- West, 113, 207, 271
Weston, 278, 304
White, 184, 297
Whitefoot, 64
Whitehouse, 142
Wichers, 85, 308
Widmer, 321
Wiedemann, 48
Wiegand, 395
Wiegreffe, 288
Wiemerslage, 118
Wierzbicki, 169, 339
Wieting, 84
Wijanarko, 345
Wijaya, 345
Wilcox, 96, 97, 99
Wilczyska, 169
Willie, 212
Wille, 171
Williams, 72, 110, 219, 256, 277
Williamson, 77
Willke, 321
Wilson, 120, 238
Winata, 132, 142, 145, 148
Winkler, 248
Winston, 64
Wirawan, 303
Wisznia, 260
Witbrock, 75
Wojciechowski, 274
Wolf, 65, 402
Won, 208
wonbyung, 269
Wong, 43, 103, 108, 170, 206, 251, 270, 282, 388
Woo, 247
Wood, 193
Woodhead, 109
Woodruff, 408
Wooldridge, 163
Wright, 299
WU, 367, 383
Wu, 43, 45, 46, 48, 54, 56–58, 60, 61, 63, 65, 68, 76, 82, 85, 87, 91, 92, 101, 102, 105–107, 111–113, 116, 119, 125, 127, 131–134, 136, 140, 143–145, 147, 148, 155, 158–160, 167, 176, 180, 183, 188, 192, 196, 197, 204, 205, 207, 209, 222, 225, 226, 236, 238, 242, 244, 246, 247, 250, 256, 258, 262, 264, 269, 276, 278, 281–283, 286, 293, 296, 301, 302, 306, 307, 317, 320–323, 327, 328, 331, 333, 338, 343–346, 350, 352, 355, 358, 360–362, 365, 366, 373, 374, 381, 382, 386, 389, 392, 394, 398, 404, 408
wu, 257, 330
Wuehrl, 297
Wuraola, 295
Xhelili, 146
Xi, 208, 243, 275, 385
Xia, 55, 65, 94, 95, 131, 132, 138, 157, 283, 321, 328, 341, 357, 381, 388
Xian, 126, 246
Xiang, 42, 127, 224, 258, 308, 368
Xiao, 52, 87, 91, 92, 109, 111, 134, 149, 179, 192, 258, 277, 281, 291, 296, 297, 309, 310, 329, 340, 348, 353, 355, 360, 361, 365, 367, 373, 377, 389, 390, 393, 402, 404, 406
XIE, 122
Xie, 54, 82, 122, 136, 148, 164, 174, 185, 193, 205, 211, 216, 217, 223, 237, 271, 276, 283, 318, 319, 342, 347, 355, 360, 364, 367, 376, 390, 399
Xin, 107, 344
Xing, 65, 82, 126, 135, 155, 197, 242, 292, 333, 404
Xiong, 51, 94, 95, 107, 136, 153, 170, 183, 231, 244, 276, 285, 287, 311, 317, 325, 328, 341, 342, 345, 355
Xiayaoxiao, 355
Xu, 47, 50, 51, 54, 65, 67, 74, 76, 78, 79, 83–85, 87, 93, 101, 105, 106, 111, 113, 118, 119, 121, 124–126, 131, 134, 138, 140, 142, 143, 157, 159, 165, 166, 168, 170, 178, 179, 182, 185, 186, 191, 192, 203, 205, 207, 210, 211, 218, 226, 228, 230, 232, 233, 236, 239, 242, 244, 249, 251, 255, 261, 262, 271, 275, 276, 278, 289, 300, 305, 307, 308, 312, 317, 318, 320–322, 329–331, 336, 338, 341, 344–349, 351, 352, 355, 356, 360, 364, 367, 368, 370, 375–378, 381, 383–385, 388, 392, 394, 395, 397–399, 401, 404–406, 409
xu, 252
Xuan, 83
xuan, 338
XUE, 388
Xue, 151, 260, 293, 320, 343, 372, 392, 408
Xue', 337
Xun, 373

- Xuweiliu, 56
- Yadav, 171, 219, 265, 288, 299
- Yadavalli, 165
- Yadwadkar, 178
- Yakovlev, 380
- Yamada, 126
- Yamagiwa, 268, 274
- Yamaguchi, 99, 145, 175
- Yamana, 387
- Yamasaki, 369
- Yamashita, 216
- Yamazaki, 302
- Yamshchikov, 275
- YAN, 122
- Yan, 42, 51, 53, 85, 105, 106, 127, 133, 134, 151, 157, 161, 193, 225, 238, 242, 243, 250, 255, 280, 307, 311, 317, 319, 321, 324, 332, 333, 336, 338, 356, 359, 376
- Yanaka, 363
- Yancan, 166
- Yancey, 63
- YANG, 46, 84
- Yang, 43, 44, 51, 53, 59, 62, 65, 70, 83, 84, 89, 92, 102, 105–108, 112, 116, 117, 122, 127, 128, 130, 134, 137, 151, 158, 163–166, 173, 177, 179, 181, 182, 191, 193, 196, 203, 205, 207–209, 213, 215, 217, 222, 229, 233, 236, 238, 242, 243, 251, 254, 258, 260, 271, 274, 279, 282, 285, 289, 291, 298–300, 310, 319, 320, 322, 327, 331, 332, 336, 339, 341, 343, 347, 348, 352, 355, 358, 359, 367, 369, 373, 375, 376, 378, 384–387, 392, 394, 396–398, 401, 407
- Yanuka, 246
- YAO, 398
- Yao, 83, 119, 136, 158, 168, 183, 184, 202, 217–219, 237, 251, 271, 305, 308, 321, 323, 388, 391, 404, 405
- Yap, 373
- Yarden, 129
- Yarowsky, 144
- Yasin, 57
- Yates, 53, 147, 158
- Yatskar, 70
- Yau, 90, 344
- Yavuz, 51, 170, 276, 287
- Ye, 83, 173, 193, 226, 235, 243, 263, 275, 276, 289, 320, 331, 332, 336, 342, 344, 359, 362, 370, 376, 377, 379, 389, 394, 397, 406
- Yedetore, 96
- Yeginbergen, 174
- Yeh, 44, 54, 78, 91
- Yehudai, 252, 259
- Yen, 290
- Yenigalla, 56, 307
- Yeo, 61, 185, 207, 221, 264
- YEONJU, 106
- Yerragorla, 338
- Yeung, 128, 240
- Yi, 123, 178, 306, 309
- Yih, 54
- Yihua, 404
- Yildiz, 78
- Yim, 67, 100, 315
- Yimam, 340
- Yin, 51, 52, 55, 65, 77, 85, 87, 121, 133, 134, 139, 142, 161, 163, 172, 179, 180, 186, 187, 197, 203, 205, 235, 239, 244, 274, 279, 284, 290, 291, 335, 342, 347, 350, 351, 353, 356, 374, 376, 392, 406
- yin, 177
- Ying, 103, 170, 181, 287
- Yinghui, 354
- Yinjie, 315
- Yishen, 216
- Yiwen, 345
- Yona, 40, 126
- Yong, 88, 142, 228, 357
- YongxueWu, 359
- Yoo, 107, 252, 375
- Yoon, 93, 130, 182, 215, 233, 236, 242, 250, 305, 355, 400
- Yoran, 212
- Yoshinaga, 363
- Yoshioka, 397
- You, 43, 96, 115, 214, 315, 335, 377
- Youn, 153
- youn, 303
- Young, 230
- Youngren, 116
- Youngrok, 280
- Yousefpour, 54
- Youssef, 195
- YU, 66, 190, 221
- Yu, 42, 44, 51, 53, 54, 65, 68, 77, 83, 85, 88, 94, 96, 103, 105, 111, 112, 115, 124–126, 128, 132, 136, 140, 141, 143, 149, 161, 164, 167, 170, 172, 177, 183, 192, 201, 202, 204, 207,

- 209, 215, 225, 227, 244, 258, 264, 265, 268, 270, 273, 279, 281, 286, 287, 292, 300, 304, 310, 314, 316, 322, 325, 327, 331, 332, 334, 335, 341, 342, 344, 349, 351, 355, 359, 367, 368, 374, 381, 386, 389, 393, 395, 398, 401, 404
yu, 184, 206, 217, 251
Yuan, 68, 83, 101, 108, 112, 134, 136, 140, 159, 160, 165, 166, 182, 191, 203, 243, 246, 291, 293, 299, 317, 341, 349, 362, 366, 371, 374, 386, 388, 397
YUCHENG, 268
Yue, 83, 106, 197, 321, 336, 370, 392, 400
Yun, 127, 177, 178, 185, 239, 242, 280, 327, 382
Yunusov, 350
Yurochkin, 266
Yüksel, 284
Yldz, 270

zadeh, 79
Zafar, 162
Zaharia, 80
Zaheer, 86
Zahera, 156
Zahid, 403
Zahraei, 407
Zaiane, 184, 391
Zaib, 79
Zaman, 129
Zamaraeva, 60
Zambre, 306
Zampieri, 296
ZAN, 404
Zang, 297
Zantedeschi, 202
Zaranis, 158
Zare, 319
ZarrieSS, 172, 173, 210, 269, 270
Zavelca, 217
Zayed, 221
Zebaze, 352
Zee, 219
ZekunYao, 394
Zeldes, 60, 116, 326
Zemel, 104, 278, 280, 317
ZENG, 83, 344
Zeng, 45, 50, 65, 83, 84, 121, 124, 126, 133–135, 139, 151, 159, 169, 203, 204, 209, 210, 237, 264, 282, 290, 333, 344, 353, 360, 397, 407
Zeren, 93
Zerva, 66, 283
Zettlemoyer, 54, 148, 213, 278
Zettsu, 380
Zevallos, 199
Zha, 49, 386
ZHAI, 129
Zhai, 63, 102, 246, 343, 374
Zhan, 196, 270, 360
ZHANG, 86, 124, 268, 330, 385
Zhang, 40, 41, 43, 46, 50, 51, 53, 55–58, 61, 63, 65, 67–69, 72, 73, 75–77, 79, 80, 82–87, 89, 90, 94, 95, 98, 100–103, 105–109, 112, 114, 119, 121, 122, 125, 129, 131, 132, 134–139, 142, 144, 145, 147, 148, 154, 155, 157, 159–164, 167, 168, 170, 171, 177, 179, 180, 182, 185, 187, 188, 192, 193, 195, 197–201, 204–208, 210, 211, 213, 214, 219, 220, 222, 224, 226, 229–231, 233, 235–238, 241, 242, 244–246, 249, 251–253, 257, 258, 260, 262, 264, 268, 271, 275, 279, 282, 284–287, 289–291, 293, 302, 305, 307–310, 312–314, 316–319, 321–323, 328, 330, 332, 333, 335–339, 341, 343–351, 353, 355–362, 364, 365, 368, 371–381, 383–393, 396–401, 404–406
zhang, 270, 342, 344, 350, 377
Zhangjijun, 347
zhangwenlong, 101
Zhao, 42, 43, 45, 50–52, 65, 68, 83, 84, 86, 90, 92, 94, 95, 99, 101–103, 107, 108, 111, 114, 115, 119, 122, 128, 129, 131, 136–138, 147, 151, 157, 159, 161, 162, 164–167, 170, 177, 179, 184, 190, 191, 196, 202, 205, 210, 214, 215, 223, 224, 226, 238, 241, 242, 244, 247, 249, 251, 256–258, 268, 271, 277, 279, 281, 286, 287, 306–308, 317, 318, 329, 332–334, 336, 339, 340, 342–345, 347–350, 353, 354, 357, 360, 362, 363, 367, 374, 377, 382, 384–386, 388, 391, 393, 396, 402
zhao, 129, 210, 220, 289, 315, 345, 369
Zharmagambetov, 154
Zharov, 397
Zhen, 177
Zheng, 47, 67, 102, 103, 107, 111, 123, 131, 148, 173, 175, 184, 186, 192, 208,

- 216, 226, 231, 244, 254, 257, 260,
262, 273, 275, 296, 311, 320, 329,
345, 347, 349, 352, 357, 367, 370,
377, 380, 381, 388, 391, 392, 395,
404
- zheng, 293
- Zhi, 110, 363
- Zhong, 57, 163, 191, 197, 292, 315, 335, 350,
352, 365, 368, 373, 382, 385
- zhong, 323
- ZHOU, 182, 335
- Zhou, 42, 43, 51, 58, 62, 78, 95, 99, 111–113,
115, 119, 122, 123, 131, 133, 134,
136, 138, 139, 144, 146, 147, 150,
151, 161, 163, 165, 170, 176, 179,
180, 182, 184, 185, 197, 198, 201,
204, 211, 216, 218, 223, 224, 227,
235, 237, 238, 241, 244, 246, 251,
258, 262, 270, 275, 276, 278, 287,
290, 291, 310, 314, 315, 317, 321,
333, 340, 341, 344, 347, 349, 352,
355, 356, 358, 370, 376, 377, 383,
384, 388, 395, 396, 398, 405, 406,
409
- ZHU, 234
- Zhu, 43, 52, 57, 59, 69, 75, 85, 90, 92, 114,
116, 134, 137, 138, 147, 148, 154,
174, 182, 202, 206–208, 213, 218,
222, 223, 226, 236, 238, 243, 248,
254, 261, 262, 275, 281, 286, 300,
305, 308, 313, 316, 321, 322, 326,
328, 330, 332, 346–348, 351,
353–355, 361, 362, 364–367, 370,
374, 377, 383, 384, 387, 389, 392,
393, 397, 400, 406
- Zhuang, 58, 83, 93, 139, 168, 207, 253, 275,
300, 307, 322, 332, 338, 346, 347,
- 377, 393
- Zhuo, 287, 354
- Ziabari, 300
- ZiangWu, 182
- Ziems, 286
- ZihanWang, 137
- Zihe, 381
- Zimmerman, 56
- Ziqi, 268
- ziqi, 344
- Ziser, 128
- ZiyanLiu, 233
- Zmigrod, 285
- Zoicas, 142
- Zoizner, 151
- Zong, 76, 251, 338
- Zongsheng, 312
- Zou, 50, 65, 90, 101, 112, 201, 204, 231, 249,
291, 300, 346, 347, 350, 391, 393
- Zouhar, 117
- Zu, 249
- Zubiaga, 129, 285
- Zubova, 170
- Zuccon, 93
- Zuchen, 112
- Zuidema, 256
- Zuo, 207, 370, 376
- Zur, 188
- Cavuolu, 181
- Öncel, 270
- Üstün, 75, 293
- ukasik, 303
- ahinuç, 309
- en, 181
- tefánik, 201

AI at Bloomberg

Bloomberg is building the world's most trusted information network for financial professionals. Our 9,000+ engineers, developers, and data scientists are dedicated to advancing and building new solutions and systems for the Bloomberg Terminal and other products in order to solve complex, real-world problems.

We have 15+ years of experience building and using AI solutions to help process and organize the ever-increasing volume of structured and unstructured financial information. In fact, we were one of the first on the street to use AI. Our team of 350+ AI researchers and engineers deploy different AI technologies, including NLP, ML, computer vision, and generative models, as well as neural networks and time-series analysis. This is how we provide our customers around the globe with the reliable, high-quality data, news, and information they need to make well-informed decisions and be more productive.



For more information:



#MAKEITHAPPENHERE

Bloomberg

Engineering

**Make it
happen here.**

++++
++++

Google DeepMind

What if solving one
problem could
unlock solutions to
thousands more?



We are hiring

deepmind.google/careers



Join us in the pursuit of
what's possible with AI.

Open positions



www.metacareers.com

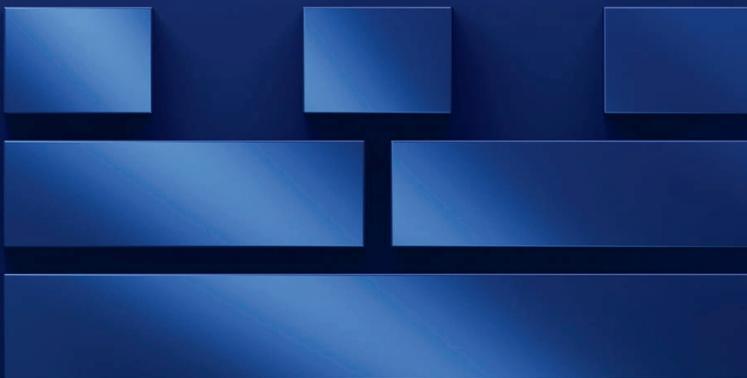


CITADEL



CITADEL | Securities

Build, test, and refine
your ideas at the speed
of the markets.



Learn more about roles and events

citadel.com | citadelsecurities.com



Megagon Labs

**Connecting AI, Data, and People
for Better Opportunities**

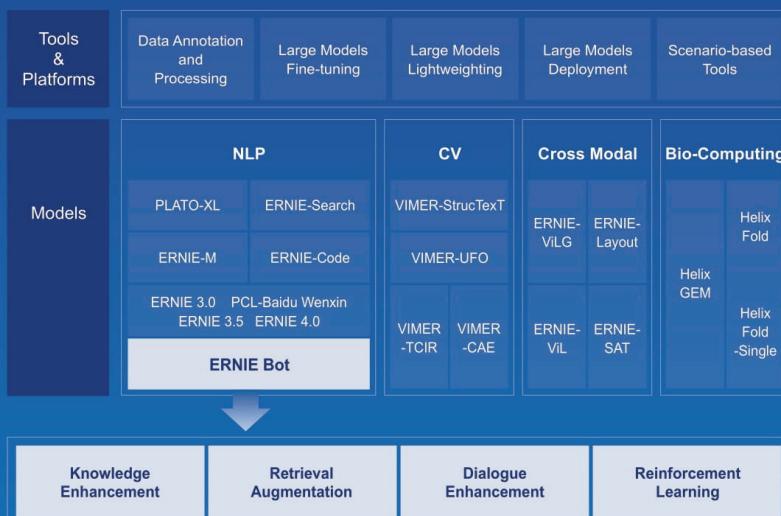




BAIDU WENXIN

A series of knowledge-enhanced large models, from general-purpose to industry-specific, independently developed by Baidu, speed up AI innovations and empower the industry to upgrade.

ERNIE Knowledge - Enhanced Large Models





cohere.com

About Cohere

Cohere is the leading AI platform for enterprise. We build world-class large language models (LLMs) that allow computers to search, understand meaning, and converse in text. Our models are uniquely suited to the needs of business, providing ease of use and strong security and privacy controls across multiple deployment options.

❖ Embeddings Models

Cohere Embed is an embeddings model which translates text into numerical vectors that models can understand. We provide industry-leading English and multilingual models (100+ languages) for use cases including:

- Semantic search
- Text classification
- Search engine for RAG
- Legacy search improvement

❖ Generative Models

Cohere Command is a text generation model, available in two different sizes, that is highly customizable for business use cases, including:

- Text generation
- Text summarization
- RAG
- Chat

↗ Cohere For AI

Cohere For AI (C4AI) is Cohere's research lab that seeks to solve complex machine learning problems. We support fundamental research and are focused on creating more points of entry into machine learning research.

- ❖ **Aya:** our global initiative to push the boundaries of multilingual AI with state-of-the-art models and datasets, widely available for research use.
- ❖ **Our Research:** we work at the frontier of AI progress to solve cutting-edge scientific problems while changing where, how, and by whom research is done. We believe that technology is powerful, and empowering different perspectives ensures responsible innovation.
- ❖ **Open Science:** we collaborate openly with independent researchers all over the world to conduct top-tier ML research. Our open science research community is a space where researchers connect through virtual events, collaborate on research, and support each other on their ML research journeys.
- ❖ **Research Grants:** we provide academic partners, developers, researchers, and other members of our community with subsidized access to the Cohere API.

Learn more, and join us in exploring the unknown, together: cohere.com/research.

To learn more about Cohere, contact us at: cohere.com/contact-sales.



DOUBAO TEAM

About ByteDance Doubao Team

Founded in 2023, the ByteDance Doubao (Seed) Team, is dedicated to pioneering advanced AI foundation models. Our goal is to lead in cutting-edge research and drive technological and societal advancements.

With a strong commitment to AI, our research areas span deep learning, reinforcement learning, Language, Vision, Audio, AI Infra and AI Safety. Our team has labs and research positions across China, Singapore, and the US.

ByteDance **Top Seed** Talent Program

The Top Seed Talent Program is an exclusive initiative by the ByteDance Doubao Team to attract exceptional talent from campuses worldwide. We seek top minds who aspire to "change the world with technology."

How to Apply

Scan the QR code and check our positions



If you have any questions,
please contact us at topseed@bytedance.com

Learn More About Doubao Team



team.doubao.com



Come build the future with us

At Amazon, we fundamentally believe that innovation is essential to being the most customer-centric company in the world. It's the company's ability to have an impact at scale that allows us to attract some of the brightest minds in artificial intelligence, and related fields.

Connect with us at:
emnlp-2024@amazon.com

Academics at Amazon



Careers



Internship Opportunities



ORACLE

Real careers in artificial intelligence

Oracle is proud to sponsor the 2024 Conference on Empirical Methods in Natural Language Processing. And we're just as proud to drive the future of AI forward. Every day, our people strive to build the solutions that empower customers to harness this revolution.

Now, it's your turn. Take your career to the forefront of innovation and put your knowledge to work with the cloud leader for global business.

Discover the next step in your career at oracle.com/joinoci

Create the **future** with us





Toloka powers Post-Training for Generative models

Data stage

Pre-training 1



Fine-tuning 2

Domains
All modalities
Reasoning

Alignment 3

Custom data collection
Static & interactive

Evaluation 4

Human evaluation
Benchmarks
Red Teaming
Responsible AI / Safety



All data types (text, image, video, audio)



Diverse languages



Specialized domains (coding, law, finance, more)



toloka.ai

Transforming the World of Consumer and Small Business Finance with Generative AI



Intuit has accelerated AI innovation at scale for many years to deliver personalized experiences to ~100 million consumer and small business customers with TurboTax, Credit Karma, QuickBooks and Mailchimp. Today, the company is supercharging its financial technology platform with generative AI (GenAI) to deliver game-changing experiences through collaborative partnerships between business units, data scientists, engineers, and AI researchers.

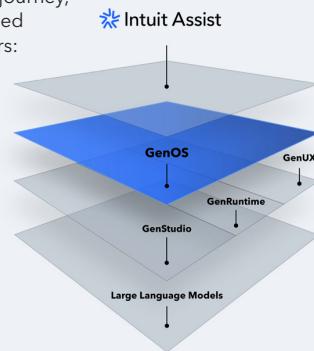
Intuit's investment in a robust AI infrastructure to democratize AI has made it possible for technologists across the company to build AI capabilities into our products at scale. To fuel rapid innovation with generative AI (GenAI), we built GenOS, a proprietary GenAI operating system that's empowering Intuit technologists to design, build, and deploy breakthrough experiences. For example, Intuit Assist is embedded across our platform and products, using powerful and relevant contextual data sets spanning small business, consumer finance, and tax to help our customers make smart financial decisions.

Because we understand where the customer is on their financial journey, we can anticipate or predict what's coming to provide personalized recommendations or insights. The results are clear in the numbers:

- **60B machine learning predictions per day**
- **4M models running in production per day**
- **1,000+ AI, machine learning, and data science US patent assets**
- **20+ papers in genAI in top conferences**

To improve customer experience and instill higher confidence in our GenAI based offerings, the AI Research team at Intuit is actively researching these topics:

- Trustworthiness and Robustness in GenAI
- Controllability and Optimization in GenAI
- Knowledge and Reasoning of LLMs
- Large Language Model Performance
- Multimodal Generative AI
- AI Powered Systems
- Conversational Agents/Dialog Systems



Learn more about
AI & Research at Intuit



Meet the Team at EMNLP 2024!

Join us at the Intuit booth on November 12-16 during exhibit hours and learn more about the 7 papers accepted at the conference.

We are hiring across all roles! jobs.intuit.com

millennium



About Us

Founded in 1989, Millennium is a global, diversified, alternative investment firm, with \$68+ billion AUM, seeking to deliver consistent, high-quality results for our investors. We have over 330 investment teams investing in strategies across industry sectors, asset classes and geographies.



Our technologists: Our 1,400+ technologists have the flexibility and resources to build solutions that work to keep us ahead in an ever-changing world. We leverage AI and machine learning, cloud technology, information security and quantitative modeling to deliver a highly-sophisticated technology platform.

- Solve complex challenges at scale
- Drive business outcomes
- Power innovation
- Hands-on experience, real-world impact

Explore Opportunities

instagram.com/lifeatmillennium
linkedin.com/company/millennium-partners
www.mfp.com/people/technology

TURING

Trusted by the world's leading AI companies

Turing enhances LLM problem solving, reasoning, and multimodality. We power the world's most complex models for performance, accuracy, and reliability.



Code optimization
for efficiency



Tailored
model training



Seamless integration
for smooth scaling

www.turing.com



Ant Group

Ant Group traces its roots back to Alipay, which was established in 2004 to create trust between online sellers and buyers. Over the years, Ant Group has grown to become one of the world's leading open Internet platforms.

Through technological innovation, we support our partners in providing inclusive, convenient digital life and digital financial services to consumers and SMEs. In addition, we have been introducing new technologies and products to support the digital transformation of industries and facilitate collaboration. Working together with global partners, we enable merchants and consumers to make and receive payments and remit around the world.

Digital Payment Digital Connectivity Digital Finance

Digital Technologies

Globalization

<https://www.antgroup.com/>
AntResearch@antgroup.com

translated.

\$100,000

to fund language technology innovators who share the goal of making it easier for everyone to understand and be understood by all others.

Find out more.



JPMorganChase

MACHINE LEARNING CENTER OF EXCELLENCE

Meet the MLCOE, leaders in machine learning and innovation. Visit us to learn how our diverse team is delivering AI/ML solutions to transform finance!

We're looking for curious problem-solvers with a passion for developing innovative ML solutions.

jpmorgan.com/mlcoe

See you at
Booth 12!



Adobe Careers.
Let's create experiences
that matter.

Adobe Research supports EMNLP 2024

November 12th - 16, 2024

We are hiring!

Our Research Areas

AI & Machine Learning

Content Intelligence

AR, VR & 360 Photography

Data Intelligence

Computer Vision, Imaging & Video

Audio

Natural Language Processing

Intelligent Agents & Assistants

Graphics (2D & 3D)

Document Intelligence

Human-Computer Interaction

Systems & Languages

With a team of world-class research scientists, engineers, artists, and designers, Adobe Research combines cutting-edge academic discovery with industry impact. Our researchers shape early-stage ideas into innovative technologies. We collaborate with interns and faculty from universities across the globe. Learn more: research.adobe.com

Adobe





**GET
AI
READY!**

Be ready to lead in
AI Strategy and Implementation.

 To learn more,
visit: ai.fiu.edu/

FIU | Engineering & Computing
FLORIDA INTERNATIONAL UNIVERSITY



Jane Street

A research-driven trading firm
that's technical to the core.

We build models, strategies, and
systems that price and trade a variety
of financial instruments, and analyze
large datasets using a variety of
Machine Learning techniques, exploring
the latest theory and pushing beyond
existing performance limits.

[Learn more](#) 



Figma

Figma is a design platform
for teams who build
products together. Born on
the Web, Figma helps the
entire product team create,
test, and ship better
designs, faster.

Interested in
learning more about
AI at Figma?


[Meet Figma AI](#)


[Building Figma AI](#)

EMNLP 2024

SPONSORS



DIAMOND SPONSORS

Google DeepMind

Meta

Bloomberg

Engineering



CITADEL | CITADEL Securities

PLATINUM SPONSORS



Megagon Labs

Baidu 百度

cohere

INTUIT

Toloka

ORACLE

amazon | science

ByteDance

SCAI

GOLD SPONSORS

millennium

TURING

SILVER SPONSORS

ANT GROUP

BRONZE SPONSORS

JPMorganChase

translated.

Adobe

FIU FLORIDA INTERNATIONAL UNIVERSITY

Jane Street

Figma