EMNLP 2023
SENTOSA, SINGAPORE
DECEMBER 6-10

# Contents

**Table of Contents** i

**1 Conference Information** 1
Message from the General Chair . . . 1
Message from the Program Chairs . . . 4
Message from the Local Chair . . . 9
Organizing Committee . . . 10
Senior Program Committee . . . 13
Program Committee . . . 17

**2 Anti-Harassment Policy** 33

**3 Meal Info** 35

**4 Welcome Reception** 37

**5 Social Events** 39

**6 Keynotes** 43

**7 Panel** 49

**8 Birds-of-a-Feather and Affinity Group Meetup** 51

**9 Tutorials: Wednesday, December 6, 2023** 55
Overview . . . 55
Message from the Tutorial Chairs . . . 57
**T1** - NLP+Vis: NLP Meets Visualization . . . 58
**T2** - Security Challenges in Natural Language Processing Models . . . 60
**T3** - Designing, Evaluating, and Learning from Humans Interacting with NLP Models . . . 61

**T4** - LLM-driven Instruction Following: Progresses and Concerns . . . 63
**T5** - Mitigating Societal Harms in Large Language Models . . . 65
**T6** - Creative Natural Language Generation . . . 67

**10 Workshops: December 6 & 7, 2023** **69**
Overview . . . 69
**W1** - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only) . . . 71
**W2** - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023) . . . 74
**W3** - The Eighth Conference on Machine Translation (WMT23) . . . 75
**W4** - GenBench: The first workshop on generalisation (benchmarking) in NLP . . . 79
**W5** - The 4th International Workshop on Computational Approaches to Historical Language Change (LChange'23) . . . 82
**W6** - The 4th New Frontiers in Summarization Workshop (NewSumm) . . . 84
**W7** - The 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS) . . . 85
**W8** - The Pattern-based Approaches to NLP in the Age of Deep Learning Workshop (Pan-DL) . . . 87
**W9** - The Seventh Widening NLP Workshop (WiNLP 2023) . . . 88
**W10** - Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP) . . . 91
**W11** - Natural Language Generation, Evaluation, and Metric (GEM) . . . 92
**W1** - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only) . . . 93
**W2** - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023) . . . 96
**W3** - The Eighth Conference on Machine Translation (WMT23) . . . 97
**W12** - The 10th Workshop on Argument Mining (ArgMining) . . . 101
**W13** - The Big Picture: Crafting a Research Narrative (BigPicture) . . . 102
**W14** - BlackboxNLP 2023: The 6th Workshop on Analysing and Interpreting Neural Networks for NLP . . . 103
**W15** - The Sixth Workshop on Computational Approaches to Linguistic Code Switching . . . 104
**W16** - The Natural Legal Language Processing Workshop 2023 (NLLP) . . . 106
**W17** - The First Arabic Natural Language Processing Conference (ArabicNLP 2023) . . . 109
**W18** - The Third Workshop on Multi-lingual Representation Learning (MRL) . . . 116
**W19** - Novel Ideas in Learning to Learn through Interaction (NILLI) . . . 117
**W20** - The First Bangla Language Processing Workshop (BLP) . . . 118

**11 Main Conference: December 8 - 10, 2023** **121**
Main Conference Program (Overview) . . . 122
Main Conference: Friday, December 8, 2023 . . . 125
Information Extraction 1 . . . 125
Machine Translation . . . 126
Computational Social Science and Cultural Analytics . . . 127
Dialogue and Interactive Systems 1 . . . 128
Demo session 1 . . . 129
Poster session 1 . . . 131
Findings 1 . . . 157
Industry 1 . . . 166
Discourse and Pragmatics . . . 167
Commonsense Reasoning . . . 168
Efficient Methods for NLP 1 . . . 169
Ethics in NLP . . . 170
Phonology, Morphology, and Word Segmentation . . . 171

Information Extraction 2 . . . 172
Demo session 2 . . . 173
Poster session 2 . . . 174
Findings 2 . . . 201
Industry 2 . . . 210
Interpretability, Interactivity, and Analysis of Models for NLP 1 . . . 210
Language Grounding to Vision, Robotics and Beyond . . . 211
Language Modeling and Analysis of Language Models 1 . . . 212
Information Retrieval and Text Mining . . . 214
Linguistic Theories, Cognitive Modeling and Psycholinguistics . . . 215
Dialogue and Interactive Systems 2 . . . 216
Demo session 3 . . . 217
Poster session 3 . . . 218
Findings 3 . . . 245
Industry 3 . . . 254
Main Conference: Saturday, December 9, 2023 . . . 256
Multilinguality and Linguistic Diversity 1 . . . 256
Natural Language Generation 1 . . . 257
NLP Applications 1 . . . 258
Theme Track: Large Language Models and the Future of NLP 1 . . . 259
Efficient Methods for NLP 2 . . . 260
Human-Centered NLP . . . 261
Demo session 4 . . . 262
Poster session 4 . . . 263
Findings 4 . . . 282
Industry 4 . . . 299
Interpretability, Interactivity, and Analysis of Models for NLP 2 . . . 299
Language Modeling and Analysis of Language Models 2 . . . 301
Multilinguality and Linguistic Diversity 2 . . . 302
Natural Language Generation 2 . . . 303
Question Answering . . . 304
Resources and Evaluation 1 . . . 305
Demo session 5 . . . 306
Poster session 5 . . . 308
Findings 5 . . . 329
Industry 5 . . . 344
Main Conference: Sunday, December 10, 2023 . . . 345
Semantics 1 . . . 345
Sentiment/Stylistic Analysis . . . 346
Speech & Multimodality 1 . . . 347
Summarization . . . 348
Machine Learning for NLP . . . 349
Syntax, Parsing and their Applications . . . 350
Demo session 6 . . . 351
Poster session 6 . . . 352
Findings 6 . . . 374
Industry 6 . . . 388
NLP Applications 2 . . . 389
Resources and Evaluation 2 . . . 390
Semantics 2 . . . 391
Speech & Multimodality 2 . . . 392
Theme Track: Large Language Models and the Future of NLP 2 . . . 393
Industry track . . . 394

Demo session 7 . . . 395
Industry 7 . . . 396
Findings 7 . . . 407

**12 Local Guide** **433**
Conference Venue . . . 433
About Singapore . . . 433
Useful Information . . . 436
Visa & Passport . . . 436
Travel to the Conference Venue . . . 437

**13 Venue Map** **439**

**Author Index** **443**

**Sponsorship** **487**

# 1 Conference Information

## Message from the General Chair

I am happy to welcome you to EMNLP-2023 in Singapore! Like EMNLP-2021, EMNLP-2022, and other ACL-related meetings, we decided to host EMNLP-2023 as another hybrid conference having both in-person and virtual presentations and participants. We are not sure how long this style of our meetings will last. However, we have already accustomed to this style of conferences, which has its own advantages, while it causes a heavy burden to those organizing such events.

The past one year has been a terrific and thrilling year since the advent of ChatGPT and other Large Language Models. Any people having access to those models has posed a big impact on people's impression about AI and has started to give them a feeling of fear. People now can do not only natural conversation with AI but also conduct various natural language tasks using our own languages. We now know it is difficult to guarantee that Large Language Models produce honest and harmless outputs. We have found that good prompting, demonstrations and complex ones like the Chain of Thought prompting draw out or enhance the emergent abilities of Large Language Models. However, we still don't know precisely why and how such in-context learning works. This year's EMNLP highlights a theme track, "Large Language Models and the Future of NLP." I hope we can see enthusiastic discussions and innovative ideas will be presented in EMNLP-2023.

One big trial is that the Program Chairs decided to use OpenReview as the cradle of the main conference papers, for making reviews and author responses publicly available. The motivation and effects of this trial will be explained by the PC Chairs.

Another important trial is to rent out the Universal Studio Singapore for our Social Event. I hope everyone will enjoy this event.

EMNLP-2023 is the biggest conference ever in the SIGDAT history. Organizing such a big event is very difficult. As the General Chair, the most important and difficult task is to organize all the committees by a group of enthusiastic and talented people. I was very fortunate to be able to collect great committee members. Without such a wonderful group of colleagues, it almost has been impossible to make this great event happen. I would like to send my sincere thanks to all the members of our organization teams. Here, I only list the chairs by names, but I also like to send gratitude from my heart to all the people involved in EMNLP-2023, including keynote speakers, panelists, workshop organizers, tutorial tutors, senior area

chairs, area chairs, reviewers, volunteers, sponsors, the Underline team, and all of you attending EMNLP-2023 in-person or virtually.

- The program chairs — Houda Bouamor, Juan Pino, and Kalika Bali — who made a number of innovations and handled a huge number of submitted papers. I cannot help but be grateful for their tireless work.
- The Local Chair and the Local Team — Haizhou Li the Chair organized and lead a wonderful group of people. While I cannot name every one of them, weekly meetings with the team members including related Chairs made our communication smooth and worked as a good time-keeper.

For the remaining committee chairs, I only list them by names, as I cannot give all my gratitude only with short messages.

- The Industry Track Chairs — Mingxuan Wang and Imed Zitouni
- The Workshop Chairs — Sujian Li, Alex Marin, and Hao Fei
- The Tutorial Chairs — Qi Zhang and Hassan Sajjad
- The Ethics Chairs — Kemal Oflazer and Francisco Guzman
- The Demonstration Chairs — Yansong Feng and Els Lefever
- The Publication Chairs — Nadi Tomeh, Atsushi Fujita, Aixin Sun, Bin Wang, Rong Tong, and Ryan Cotterell
- Publicity Chairs: Weiwei Sun and Noriki Nishida
- The Student Volunteer Chairs — Mamoru Komachi, Kevin Duh, and Ekaterina Kochmar
- The Diversity/Inclusion Chairs — Jing Li, Luisa Bentivogli, Eva Vanmassenhove, and Shi Zong
- The Reviewer Mentoring Chairs — Roi Reichart and Roy Lee
- The Sponsorship Chairs — Deyi Xiong, Shyam Upadhyay, and the ACL Sponsorship Director, Chris Callison-Burch
- The Website Chairs — Yan Zhang and Benyou Wang
- The Virtual Infrastructure Chairs — Feng Jiang, Chen Zhang, and the Underline Science team, especially Damira Mršić
- Last but least, I want to express special thanks to Jennifer Rachford, the ACL Business Manager for her endless attention and guidance to all aspects of the organization.

I also express my gratitude to our sponsors, without their donation this conference would not be possible:

- Diamond Sponsors: Apple, Colossal-AI, Google Research, GTCOM, King Salman Global Academy for Arabic Language, LivePerson, and SONY
- Platinum Sponsors: Ahrefs, Alibaba Cloud, Amazon, Baidu, ByteDance, Cohere, Megagon Labs, and NEC

- Gold Sponsors: ANT GROUP, Bloomberg, HUAWEI, J.P.Morgan Chase, Salesforce, and SAP
- Silver Sponsors: aiXplain, duolingo, jenni, and Translated
- Bronze Sponsors: Adobe, Babelscape, ModelBest, and nyonic
- Diversity and Inclusion, Champion: RIKEN AIP
- Diversity and Inclusion, Ally: Amazon

Finally, I hope all of you enjoy the main conference, workshops, tutorials and all other events, and make your stay in Singapore an unforgettable experience!

Yuji Matsumoto
RIKEN AIP, Tokyo, Japan
EMNLP 2023 General Chair

# Message from the Program Chairs

Welcome to EMNLP 2023, one of the most-attended conferences in the field of Natural Language Processing, held in "hybrid" mode this year serving both virtual and in-person participants in Singapore, where we have increased the number of Findings papers to be more inclusive and to showcase the diversity and quality of research in our field.

## Submission and Acceptance

EMNLP 2023 received 4,909 full paper submissions, the largest number to date. This number includes 134 ARR papers that were committed to EMNLP (see further discussion of ARR below). 256 papers were desk rejected for various reasons (missing limitation section, anonymity policy, multiple submission policy or formatting violations), leaving us with submissions that were fully reviewed. This is a record-breaking number of submissions, an increase of 719 submissions over last year. Based on the reviewers, area chairs and senior area chair comments, we have tried to keep the EMNLP 2023 main acceptance rates similar to previous events while increasing the number of Findings.

We accepted 1,105 papers to the main conference including 9 nominated by the Computational Linguistics (CL) journal, and 49 by the Transactions of the Association of Computational Linguistics (TACL). Out of these, 234 are oral presentations and 871 poster presentations. We also accepted 1,060 papers for "Findings of EMNLP". Additionally, a total of 77 and 52 submissions were accepted to be presented in the Industry and Demo tracks, respectively.

More statistics on the accepted papers can be found below.

| | Long | Short | Total |
|---|---|---|---|
| Submitted (Including ARR commits) | 3,868 | 1,041 | 4,909 |
| Accepted to the Main conference | 901 | 146 | 1,047 |
| Acceptance Rate (main conference) | 23.3% | 14.0% | 21.3% |
| Accepted to Findings | 857 | 203 | 1,060 |
| Acceptance Rate (Main + Findings) | 45.4% | 33.5% | 42.9% |
| Presented TACL papers | | | 49 |
| Presented CL papers | | | 9 |
| Accepted Industry | | | 77 |
| Accepted Demo | | | 52 |

## Limitations Statement

We continued the practice started in the previous year with a requirement that each submitted paper must include an explicitly named Limitations section, discussing the limitations of the work. This discussion does not count towards the page limit, and was strictly implemented. A large number of desk rejects were due to a missing Limitations section.

## Tracks

To ensure a smooth process, the submissions to EMNLP 2023 were divided into 27 tracks. The tracks mostly followed those of previous EMNLP conferences, reflecting the "standard" divisions in the field. We did however make the following changes: the "Multilinguality" track became "Multilinguality and Linguistic Diversity" to explicitly call out the work on diverse languages; and we added ONE additional track on "Human-centered NLP". Finally, we continued the inclusion of a "Theme Track" and solicited

papers on "Large Language Models and the Future NLP", discussing empirical and theoretical research, as well as position and survey papers on the ways in which the new generation of Large Language Models perform for NLP tasks and applications, and what this means for the future of NLP as a field. Of the 27 tracks, the Resources and Evaluation, NLP Applications, Dialogue and Interactive Systems, Information Extraction, and the Theme tracks were the most popular with over 200 submissions per track.

## Program committee structure & reviewing

Similar to prior NLP conferences, we adopted the hierarchical program committee structure, where for each area we invited between 1 to 5 Senior Area Chairs (SACs), who worked with a team of Area Chairs (ACs), and an army of reviewers. We relied on statistics from prior years to estimate how many SACs, ACs and reviewers will be needed and ended up with 85 SACs and 458 ACs. For the reviewers, we used the reviewer lists from prior conferences, solicited volunteer reviewers. To this end, we used the reviewers from ACL2023. We provided the information to the program committee for making reviewer assignments. This resulted in a reviewer pool of 4,094 reviewers of which at least 3,643 reviewers submitted at least one review. For each submission, we assigned three reviewers and an AC. The initial paper assignment was made using an automatic algorithm that matches the abstracts and submitted PDFs with ACs/reviewers' past publication records, then the assignments were further refined by the SACs/PCs. This matching was done at the global level and not localized per track, which meant that a reviewer could review two (or more) relevant papers from different tracks.

Open Review (see Transparency and Quality of Reviews below) calculated affinity scores making sure that each reviewer does not have a load exceeding 5 papers. We also accommodated requests for a limited load. Besides the overall recommendation, reviewers were asked whether there was any ethical concern. To ensure the review quality, we provided detailed guidelines about what reviewers should and shouldn't do in a review.

We made final decisions according to the rankings and SAC recommendations. Our final decisions were made not just on the review scores, but also took into account the reviews, author responses, discussion, meta-reviews and SAC/AC recommendations.

## Transparency and Quality of Reviews

For a transparent and open reviewing system, this year we have implemented a process by which some of the reviews, author responses and meta reviews will be made publicly available. Our motivation here was to provide increased transparency in the review process, to foster more accountability for reviewers and higher quality reviews as well as enabling peer review research by providing an open collection of papers and reviews. Only reviews, author responses and meta reviews of accepted papers and opt-in rejected papers (where opting in is done by authors) will be made publicly available after the acceptance notifications. The requirement of making reviews open meant that we had to move the conference platform from SoftConf to OpenReview. The OpenReview platform is also currently being used by the ACL Rolling Review and some other related conferences and is well suited to this type of process.

We also introduced a rebuttal and discussions cycle to the reviewing process, where we allowed the reviewers and the authors to have conversations as a part of it. We also allowed for some additional experimentations during the rebuttal. This has, on the whole, seen a positive engagement from the authors and the reviewers. These discussions were extremely useful for the meta-reviews, and final paper decisions. This, we believe, also helps in ensuring high quality reviews to a certain extent.

A lot of behind-the scenes technical and other work has gone into making this possible. We are grateful to everyone for their patience with some of the glitches on the way. We hope that more and more authors will choose to make their reviews and data available for the community, and more and more reviewers will engage in the rebuttal discussions.

## Ethics committee

We also formed an Ethics Committee (EC) dedicated to ethical issues. Besides, we strongly encouraged the authors to include an ethics statement that did not count towards the page limit. After the technical reviews, but before the author-response and discussion phases, the ethics committee considered 54 papers that were flagged by the technical reviewing committee for ethical concerns. The two EC chairs went over the papers, to determine whether a full EC review would be required. As a result, 23 papers received two dedicated ethics reviews from a committee of 42 reviewers recruited by the EC chairs. We thank the committee for their excellent work.

## ACL Rolling Review

ACL Rolling Review (ARR) is an initiative of the Association for Computational Linguistics, where the reviewing and acceptance of papers to publication venues are done in a two-step process: (1) centralized rolling review and (2) the ability to commit the reviewed papers to be considered for publication by a publication venue. For EMNLP 2023, we continued the process from last year where the authors could either submit papers to EMNLP 2023 directly, or commit ARR reviewed papers by a certain date. We coordinated with the ARR team to extract the submission, review and meta-review. ARR submissions already have their reviews and meta-recommendation. These ARR papers were then ranked by the SACs of the given tracks.

Overall, EMNLP had 134 papers committed from ARR, of these 52 were accepted to the main conference and 45 were accepted to Findings of EMNLP.

## Best paper selection

This year we increased the number of best paper awards as following:

**a)** Best Paper (Long)

**b)** Best Paper (Short)

**c)** Best Paper (Theme)

**d)** Best Paper (Industry)

**e)** Best Paper (Demo)

**f)** Outstanding Paper in each of the tracks including the Theme, Industry and Demo tracks

Based on the nominations from SACs and ACs, we have identified 107 candidates for consideration for the best papers and outstanding papers award. These papers are assessed by the Best Paper Award Committee. The award winners will be announced and present their works at the closing ceremony.

## Presentation Mode

We attempted the decision for oral vs poster presentations not to be made based on the quality/merit of the papers, but rather on the authors' interest in the presentation mode, and our understanding of what would be the best format for presentation of each individual paper.

## Keynote talks

Another highlight of our program is the three exciting keynote talks, presented by Prof. Jong Park, KAIST, on "Human-Centric Natural Language Processing"; Prof. Emily M. Provost, University of Michigan, on "From Speech to Emotion to Mood: Mental Health Modeling in Real-World Environments" & Prof. Christopher Manning from Stanford University, on "Academic NLP research in the Age of LLMs: Nothing but blue skies!"

## Gratitude

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

- Our General Chair, Yuji Matsumoto, who led the whole organizing team, and helped with many of the decision processes;
- 85 SACs who helped us throughout the entire review process, from assigning papers, checking review quality, making final recommendation, suggesting presentation formats to recommending best paper candidates;
- 458 ACs who checked the initial submissions, led paper discussions, wrote meta reviews, ensured review quality and suggested best paper candidates;
- 3,643 reviewers who reviewed the papers and actively participated in paper discussions; special thanks to those who stepped in at the last minute to serve as emergency reviewers;
- 42 Ethics Committee members, chaired by Kemal Oflazer and Francisco Guzman, for their hard work in providing ethical reviews and meta-reviews for all papers with serious ethical issues, and ensure that all the conditionally accepted papers have addressed the ethical issues appropriately;
- The 26 members of the Best Paper Award Committee for selecting the best papers;
- Publication Chairs Nadi Tomeh, Atsushi Fujita, Aixin Sun, Bin Wang, Rong Tong, and Ryan Cotterell for completing the final proceedings within a short period;
- ACL Anthology Director Matt Post and his team, for his help in the production of the conference proceedings and maintenance of the ACL Anthology;
- TACL editor-in-chief Asli Celikyilmaz & Roi Reichart, TACL Editorial Assistant Cindy Robinson, and CL Editor-in-Chief Hwee Tou Ng for coordinating TACL and CL presentations with us;
- The ARR team for their continued effort in running ARR, and for coordination with us;
- The Open Review team, especially Harold Rubio and Celeste Matinez for multiple rounds of technical help in setting up EMNLP 2023 on the OR platform;
- Website chairs Yan Zhang and Benyou Wang for their continued effort in prompt updates to the website;
- Publicity chairs Weiwei Sun and Noriki Nishada for publicizing the conference and handling communications on social media.
- Damira Mršic and the whole Underline team, for helping to manage the logistics of both the virtual and online conference.
- Jenn Rachford for her patience, and professional invaluable help in organizing the logistics of the conference.

- Haizhou Li and the rest of the Local Organizing Committee, for various discussions on organizing EMNLP, and making the local arrangements.
- 14,105 authors who submitted their work to EMNLP 2023.

We hope that you will enjoy this year's program and hybrid conference!

Houda Bouamor, Carnegie Mellon University in Qatar
Juan Pino, Meta
Kalika Bali, Microsoft Research Labs India
EMNLP 2023 Program Co-Chairs

# Message from the Local Chair

In early 2022, Singapore announced a major relaxation of most pandemic countermeasures. At the same time, it was selected to host EMNLP 2023. It has been an exciting journey preparing for a major post-pandemic conference. The local community is now ready to welcome all of you.

Singapore embraces the coexistence of Malay, Mandarin, Tamil, and English as the national languages. It showcases how a multilingual society can thrive and prosper in harmony. The vibrant community of natural language processing in Singapore has hosted ACL 2009 and INTERSPEECH 2014 international conferences. We are thrilled once again to host EMNLP 2024.

The Chinese and Oriental Languages Information Processing Society (COLIPS) of Singapore was founded in 1988 to advance the research of computer processing of Chinese and other Asian languages. At its 35th anniversary, COLIPS is proud to be the local supporting organization of the conference.

Along the way, the local organization has been helped greatly by the General Chair Yuji Matsumoto and others from SIGDAT and ACL executive team, to whom we are extremely thankful. I would like to take this opportunity to thank COLIPS council and the local volunteers for their efforts and dedications, in particular, Yan Zhang and Benyou Wang for managing the website, Feng Jiang and Chen Zhang for virtual conference infrastructure, Bin Wang for local publication, Rong Tong for volunteer services, Celine Cheong for visa services, and Yan Wu, Minghui Dong, Lei Wang as the Singapore liaisons.

Finally, I really hope that you all enjoy the conference and your stay in Singapore.

Haizhou Li
Local Chair, EMNLP 2023

# Organizing Committee

**General Chair**

Yuji Matsumoto , RIKEN Center for Advanced Intelligence Project

**Program Chairs**

Houda Bouamor , Carnegie Mellon University in Qatar
Juan Pino , Meta
Kalika Bali , Microsoft Research Labs India

**Local Chair**

Haizhou Li , The Chinese University of Hong Kong, Shenzhen & National University of Singapore

**Tutorial Chairs**

Qi Zhang , Fudan University
Hassan Sajjad , Dalhousie University

**Workshop Chairs**

Sujian Li , Peking University
Alex Marin , Microsoft
Hao Fei , National University of Singapore

**Demonstration Chairs**

Yansong Feng , Peking University
Els Lefever , LT3, Ghent University

**Industry Track Chairs**

Mingxuan Wang , ByteDance AI Lab
Imed Zitouni , Google

**Ethics Chairs**

Kemal Oflazer , CMU
Francisco Guzman , Meta

**Publication Chairs**

Nadi Tomeh , Université Sorbonne Paris Nord
Atsushi Fujita , NICT
Aixin Sun , Nanyang Technological University
Bin Wang , Institute for Infocomm Research, A*STAR, Singapore
Rong Tong , Singapore Institute of Technology
Ryan Cotterell , ETH Zürich

**Publicity and Social Media Chairs**

Weiwei Sun , University of Cambridge
Noriki Nishida , RIKEN AIP

**Website Chairs**

Yan Zhang , National University of Singapore
Benyou Wang , The Chinese University of Hong Kong, Shenzhen

**Student Volunteer Chairs**

Mamoru Komachi , Hitotsubashi University
Kevin Duh , Johns Hopkins University
Ekaterina Kochmar , MBZUAI

**Virtual Infrastructure Chairs**

Feng Jiang , The Chinese University of Hong Kong, Shenzhen
Chen Zhang , National University of Singapore

**Diversity and Inclusion Chairs**

Jing Li , The Hong Kong Polytechnic University
Luisa Bentivogli , Fondazione Bruno Kessler
Eva Vanmassenhove , Tilburg University
Shi Zong , University of Waterloo

**Reviewer Mentoring Chairs**

Roi Reichart , Technion
Roy Lee , Singapore University of Technology and Design

**Sponsorship Chairs**

Deyi Xiong , Tianjin University
Shyam Upadhyay , Google

**ACL Event Director**

Jennifer Rachford

**ACL On-Site Team**

Brandy Dorsey
Megan Maloy
Sally Stevenson

**AV Team - Lee Hartman & Sons**

Jon Dorsey
Trevor Laffoon

**Virtual Hybrid Team - Underline**

Nateo Antonic
Borna Bevanda
Dorian Fildor
Rafael Grabovica
Jernej Masnec
Damira Mrsic
Alexandru Pricop
Lucijan Prpic
Luka Simic

**Local Organizing Committee**

Yan Zhang , National University of Singapore
Benyou Wang , The Chinese University of Hong Kong, Shenzhen
Feng Jiang , The Chinese University of Hong Kong, Shenzhen
Chen Zhang , National University of Singapore
Bin Wang , Institute for Infocomm Research, A*STAR, Singapore
Rong Tong , Singapore Institute of Technology
Celine Cheong , National University of Singapore
Yan Wu , Institute for Infocomm Research, A*STAR, Singapore
Minghui Dong , Institute for Infocomm Research, A*STAR, Singapore
Lei Wang , Huawei International Pte. Ltd.

**Supporting Organization**

Chinese and Oriental Languages Information Processing Society (COLIPS)

# Senior Program Committee

**Commonsense Reasoning**

Antoine Bosselut , École Polytechnique Fédérale de Lausanne
Soujanya Poriya , Singapore University of Technology and Design
Dan Roth , University of Pennsylvania

**Computational Social Science and Cultural Analytics**

Ashique KhudaBukhsh , Rochester Institute of Technology
Animesh Mukherjee , Indian Institute of Technology, Kharagpur
Brendan O'Connor , University of Massachusetts Amherst

**Dialogue and Interactive Systems**

Malihe Alikhani , University of Pittsburgh
Zhou Yu , Columbia University

**Discourse and Pragmatics**

Christian Hardmeier , IT University of Copenhagen
Yufang Hou , IBM Research
Shafiq Joty , Salesforce AI Research
Sebastian Schuster , Saarland University
Manfred Stede , Potsdam University

**Efficient Methods for NLP**

Sara Hooker , Cohere For AI

**Ethics and NLP**

Marta R. Costa-jussà , Meta AI
Preslav Nakov , MBZUAI

**Human-Centered NLP**

Cecilia O. Alm , Rochester Institute of Technology
Jeffrey P. Bigham , Carnegie Mellon University
Alison Smith , Dataminr
Diyi Yang , Stanford University

**Information Extraction**

Hao Fei , National University of Singapore
Bhaskar Mitra , Microsoft Research
Sudip Naksar , Jadavpur University
Scott Yih , Meta AI

**Information Retrieval and Text Mining**

David Mimno , Cornell University
Gabriella Pasi , University of Milano-Bicocca
Qifan Wang , Meta AI

**Interpretability, Interactivity and Analysis of Models for NLP**

Yonatan Belinkov , Technion
Nadir Durrani , Qatar Computing Research Institute
Sebastian Gehrmann , Bloomberg

**Language Grounding to Vision, Robotics and Beyond**

Mohit Bansal , University of North Carolina, Chapel Hill
Yonatan Bisk , Carnegie Mellon University

**Language Modeling and Analysis of Language Models**

Sandipan Dandipat , Microsoft
Colin Raffel , University of North Carolina, Chapel Hill
Partha Pratim Talukdar , Google

**Linguistic Theories, Cognitive Modeling and Psycholinguistics**

Damien Blasi , Harvard University
Ritesh Kumar , Bhimrao Ambedkar University

**Machine Learning for NLP**

Danish Pruthi , Indian Institute of Science (IISc), Bangalore
Bhiksha Raj , Carnegie Mellon University
William Wang , University of California, Santa Barbara

**Machine Translation**

Alexander Fraser , LMU Munich
Philippe Langlais , University of Montreal
Holger Schwenk , Meta AI
François Yvon , University Paris Sud

**Multilinguality and Linguistic Diversity**

Antonios Anastopoulos , George Mason University
A. Seza Doğruöz , Ghent University
Shruti Rijhwani , Google
Sunayana Sitaram , Microsoft Research India

**Natural Language Generation**

Muhammad Abdul-Mageed , University of British Columbia
Naoaki Okazaki , Tokyo Institute of Technology

Nanyung (Violet) Peng , University of California, Los Angele

**NLP Applications**

Nitin Madnani , ETS AI Labs
Nedjma Ousidhoum , Cardiff University and University of Cambridge
Min Yang , Shenzhen Institute of Advanced Technology (SIAT)

**Phonology, Morphology and Word Segmentation**

Claudia Borg , University of Malta
Ryan Cotterell , ETH Zurich
Brian Roark , Google

**Question Answering**

Roee Aharoni , Google
Hannaneh Hajishirzi , University of Washington
Alessandro Moschitti , Amazon Alexa
Min Joon Seo , Korea Advanced Institute of Science and Technology

**Resources and Evaluation**

Benoit Sagot , INRIA
Claudia Soria , CNR-ILC
Wajdi Zaghouani , Hamad Bin Khalifa University

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

Marianna Apidianaki , University of Pennsylvania
Cristina España i Bonet , Deutsches Forschungszentrum für Künstliche Intelligenz
Sobha Lalitha Devi , AU-KBC Research Centre
Paul Rayson , Lancaster University
Steven Schockaert , Cardiff University

**Sentiment Analysis, Stylistic Analysis, and Argument Mining**

David Jurgens , University of Michigan
Saif Mohammad , National Research Council Canada
Mohammad Taher Pilehvar , Tehran Institute for Advanced Studies

**Speech and Multimodality**

Ahmed Ali , Qatar Computing Research Institute (QCRI)
Murat Saraçlar , Bogazici University
Rita Singh , Carnegie Mellon University
Shinji Watanabe , Carnegie Mellon University

**Summarization**

Annie Louis , Google
Joshua Maynez , Google

Horacio Saggion , Universitat Pompeu Fabra
Xiaodan Zhu , Queen's University

**Syntax, Parsing and their Applications**

Liang Huang , Oregon State University
Joakim Nivre , Uppsala University

**Theme Track: Large Language Models and the Future of NLP**

Monojit Choudhury , Microsoft Research Lab India
Rada Mihalcea , University of Michigan
Vinodkumar Prabhakaran , Google

# Program Committee

**Commonsense Reasoning**

Niranjan Balasubramanian, Ronan Le Bras, Nouha Dziri, Xiang Li, Xuezhe Ma, Navonil Majumder, Debjit Paul, Simon Razniewski, Keisuke Sakaguchi, Niket Tandon

**Computational Social Science and Cultural Analytics**

Somak Aditya, Abhijnan Chakraborty, Anjalie Elena Field, Andrew Halterman, Abram Handler, Kristen Johnson, Katherine Keith, Srijan Kumar, Sumeet Kumar, Roy Ka-Wei Lee, Usman Naseem, Alice Oh, Eugenia Rho, Koustuv Saha, H. Andrew Schwartz, Qinlan Shen, Rob Voigt, Justine Zhang

**Dialogue and Interactive Systems**

Paul Crook, Devamanyu Hazarika, Simon Keizer, Bill Yuchen Lin, Odhisattwa Majumder, Alexandros Papangelis, Xiaojun Quan, Samira Shaikh, Lei Shu, Anthony Sicilia, Fanghua Ye, Yi Zhang

**Discourse and Pragmatics**

Feng Jiang, Murathan Kurfalı, Ekaterina Lapshinova-Koltunski, Junyi Jessy Li, Yang Janet Liu, Sharid Loáiciga, Vincent Ng, Tatjana Scheffler, Juntao Yu, Amir Zeldes

**Efficient Methods for NLP**

Cody Blakeney, Patrick H. Chen, Tim Dettmers, Beyza Ermis, Utku Evci, Yuhang Li, Bradley McDanel, Vishvak Murahari, Jonas Pfeiffer, Edoardo Ponti, Mohammad Rostami, Andreas Rücklé, Tal Schuster, Roy Schwartz, Dustin Wright, Mengzhou Xia, Ahmet Üstün

**Ethics in NLP**

Kathleen Fraser, Lea Frermann, Dirk Hovy, Anne Lauscher, Margot Mieskes, Debora Nozza, Chan Young Park, Sherif Saad, Zeerak Talat, Jieyu Zhao

**Human-Centered NLP**

Su Lin Blodgett, Kianté Brantley, Marine Carpuat, Hal Daumé III, Kenny Davila, Shi Feng, Kristina Gligoric, Saad Hassan, Michael A. Madaio, Khanh Xuan Nguyen, Weiyan Shi, Sherry Wu, Tongshuang Wu

**Information Extraction**

Jun Araki, Yixin Cao, Muhao Chen, Hao Cheng, Christos Christodoulopoulos, Asif Ekbal, Debasis Ganguly, Xu Han, Ruihong Huang, Fei Li, Hongyu Lin, Kang Liu, Tingwen Liu, Zhiyuan Liu,

Wei Lu, Thien Huu Nguyen, Qiang Ning, Ankur Padia, Jay Pujara, Alan Ritter, Mrinmaya Sachan, Yu Su, Wenpeng Yin, Yue Zhang

**Information Retrieval and Text Mining**

Maria Antoniak, Dallas Card, Fabio Crestani, Wenqi Fan, Yi Fang, Fuli Feng, Lifu Huang, Yiqun Liu, Jian-Yun Nie, Denis Peskoff, Lynda Lechani Tamine, Jinggang Wang, Zenglin Xu

**Interpretability, Interactivity and Analysis of Models for NLP**

Leila Arras, Pepa Atanasova, Jasmijn Bastings, Wei Cheng, Fahim Dalvi, Zhenyun Deng, Greg Durrett, Jacob Eisenstein, Antske Fokkens, Mor Geva, John Hewitt, Dieuwke Hupkes, Yangfeng Ji, Divyansh Kaushik, Piyawat Lertvittayakumjorn, Zaiqiao Meng, Pasquale Minervini, Isar Nejadgholi, Leonardo Ranaldi, Abhilasha Ravichander, Roi Reichart, Ashish Sabharwal, Xingyi Song, Mariya K Toneva, Martin Tutek, Elena Voita, Sarah Wiegreffe

**Language Grounding to Vision, Robotics and Beyond**

Valts Blukis, Daniel Fried, Zhe Gan, Aniruddha Kembhavi, Jie Lei, Paul Pu Liang, Fangyu Liu, Roma Patel, Jivko Sinapov, Alane Suhr, Hao Tan, Xin Eric Wang, Zirui Wang

**Language Modeling and Analysis of Language Models**

Jan Buys, Kevin Clark, Kumar A Dubey, Orhan Firat, Manish Gupta, Hany Hassan, Harsh Jhamtani, Melvin Johnson, Mandar Joshi, Urvashi Khandelwal, Lingpeng Kong, Ni Lao, Moontae Lee, Bing Liu, Peter J Liu, Qiang Liu, Naomi Saphra, Emma Strubell, Huan Sun, Lijun Wu, Chunting Zhou

**Linguistic Theories, Cognitive Modeling, and Psycholinguistics**

Carolyn Jane Anderson, Yunfei Long, Tiago Pimentel, Géraldine Walther, Philipp Wicke

**Machine Learning for NLP**

Heike Adel, Mikhail Burtsev, Giuseppe Castellucci, Kai-Wei Chang, Sourish Chaudhuri, Wenhu Chen, Arman Cohan, Danilo Croce, Julian Martin Eisenschlos, Francis Ferraro, Matthias Gallé, Philip John Gorinski, Tatsunori Hashimoto, Junxian He, Ricardo Henao, Jack Hessel, Estevam Hruschka, Dongyeop Kang, Pei Ke, Yoon Kim, Jay Yoon Lee, Irene Li, Lei Li, Zemin Liu, Ashutosh Modi, Thanh Tam Nguyen, Giannis Nikolentzos, Raphael Olivier, Ankur P Parikh, Pasquale Restaino, Freda H. Shi, Jun Suzuki, Swabha Swayamdipta, Hao Tang, Pat Verga, Taro Watanabe, Junyu Xuan, Kaisheng Yao, Se-Young Yun, Ningyu Zhang

**Machine Translation**

Loic Barrault, Rachel Bawden, Laurent Besacier, Alexandra Birch, Boxing Chen, Colin Cherry,

Angela Fan, Marcello Federico, Mark Fishel, George Foster, Markus Freitag, Thanh-Le Ha, Barry Haddow, Gholamreza Haffari, Lifeng Han, Shujian Huang, Jindřich Libovický, Qun Liu, Xuebo Liu, Lili Mou, Masaaki Nagata, Stephan Peitz, Maja Popovic, Anoop Sarkar, Rico Sennrich, Felix Stahlberg, Zhaopeng Tu, David Vilar, Joern Wuebker, Tong Xiao

**Multilinguality and Linguistic Diversity**

David Ifeoluwa Adelani, Priyanka Agrawal, Mikel Artetxe, Yoshinari Fujinuma, Dan Garrette, Constantine Lignos, Manuel Mager, Benjamin Muller, Xinyi Wang, Guillaume Wisniewki, Zheng-Xin Yon, Marcos Zampieri

**Natural Language Generation**

Ife Adebara, Alham Fikri Aji, Nora Al-Twairesh, Hanan Aldarmaki, Prithviraj Ammanabrolu, Elizabeth Clark, Mohit Iyyer, Ganesh Jawahar, Hidetaka Kamigaito, Masahiro Kaneko, Young Jin Kim, Laks V. S. Lakshmanan, Xiaodong Liu, Hassan Sajjad, Younes Samih, João Sedoc, Chiyu Zhang

**NLP Applications**

Kolawole Adebayo, Fernando Alva-Manchego, Meriem Beloucif, Daniel M. Bikel, Edward Choi, Manuel Rafael Ciosici, Colin Clement, Tirthankar Ghosal, Zhijiang Guo, Andrea Horbach, Daiki Kimura, Ekaterina Kochmar, Yanchi Liu, Natalie Parde, Lis Pereira, Daniel Preotiuc-Pietro, Marek Rei, Alla Rozovskaya, Michael Sejr Schlichtkrull, Kevin Small, Giancarlo Sperlì, Anaïs Tack, James Thorne, Subhashini Venugopalan, V.G.Vinod Vydiswaran, Xuan Wang, Qianqian Xie, Ruifeng Xu, Victoria Yaneva, Torsten Zesch, Taolin Zhang, Xin Zhao, Arkaitz Zubiaga

**Phonology, Morphology, and Word Segmentation**

Yugo Murawaki, Emily Tucker Prud'hommeaux, Kairit Sirts, Ekaterina Vylomova

**Question Answering**

Akari Asai, Iz Beltagy, Pradeep Dasigi, Markus Dreyer, Simone Filice, Siddhant Garg, Jonathan Herzig, Robin Jia, Daniel Khashabi, Tom Kwiatkowski, Ivano Lauriola, Mengwen Liu, Sewon Min, Masud Moshtaghi, Salvatore Romeo, Olga Uryupina, Thuy Vu

**Resources and Evaluation**

Firoj Alam, Federico Boschetti, Emily Chen, Vera Demberg, Rotem Dror, Iria de-Dios-Flores, Angelo Mario del Grosso, Yannick Estève, Nazli Goharian, Valeria Quochi, Sophie Rosset, Yves Scherrer, Sajad Sotudeh, Daan Van Esch

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

Alan Akbik, Valerio Basile, Eduardo Blanco, Elena Cabrio, Emmanuele Chersoni, Allyson Ettinger, Goran Glavaš, Andrey Kutuzov, Alessandro Lenci, Nafise Sadat Moosavi, Roser Morante, Constantin Orasan, Mohammad Taher Pilehvar, Yuval Pinter, Lidia Pivovarova, Alessandro Raganato, Ivan Vulić, Lonneke van der Plas, Gijs Wijnholds, Wei Xu, Sho Yokoi

**Sentiment Analysis, Stylistic Analysis, and Argument Mining**

Nikolaos Aletras, Ehsaneddin Asgari, Isabelle Augenstein, Nathaniel Blanchard, Elena Kochkina, Lun-Wei Ku, Ivan Vladimir Meza Ruiz, Joonsuk Park, Ehsan Shareghi, Shabnam Tafreshi, Eva Maria Vecchi, Serena Villata, Hao Wang, Shuai Wang, Yadollah Yaghoobzadeh

**Speech and Multimodality**

Shammur Absar Chowdhury, David Harwath, Takaaki Horiat, Wei-Ning Hsu, Ryo Ishii, Herman Kamper, Suyoun Kim, Ondrej Klejch, Gakuto Kurata, Sheng Li, Erfan Loweimi, Soumi Maiti, Florian Metze, Shruti Palaskar, Leibny Paola Garcia Perera, Suwon Shon, Katsuhito Sudoh, Matthew Wiesner, Koichiro Yoshino

**Summarization**

Reinald Kim Amplayo, Florian Boudin, Yue Dong, Yang Gao, Tanya Goyal, Hiroaki Hayashi, Sathish Reddy Indurthi, Jey Han Lau, Manling Li, Marina Litvak, Fei Liu, Aiala Rosá

**Syntax, Parsing and their Applications**

Agnieszka Falenska, Zhenghua Li, Haitao Mi, Kenji Sagae, Kewei Tu, David Vilares, Rob van der Goot, Meishan Zhang

**Theme Track: Large Language Models and the Future of NLP**

Maxamad Axmed, Gasper Begus, Laura Biester, Lidong Bing, Rishi Bommasani, Jordan Lee Boyd-Graber, Vishrav Chaudhary, Sunipa Dev, Aparna Garimella, Pawan Goyal, Michael Hahn, Oana Ignat, Zhijing Jin, Preethi Jyothi, Mitesh Khapra, Shayne Longpre, Vukosi Marivate, Shashi Narayan, Amrita Saha, Charles Welch, Steven R Wilson, Winston Wu, Rui Zhang

**Primary Reviewers**

Amirhossein Abaskohi, Harika Abburi, Abdelrahman Abdallah, Asad Abdi, Omri Abend, Gavin Abercrombie, Alafate Abulimiti, Griffin Thomas Adams, Tosin Adewumi, Jiban Adhikary, Yossi Adi, Ankur Agarwal, Dhruv Agarwal, Milind Agarwal, Shivam Agarwal, Saaket Agashe, Arshiya Aggarwal, Karan Aggarwal, Kriti Aggarwal, Pranjal Aggarwal, Zeljko Agic, Ameeta Agrawal, Ravi Agrawal, Sweta Agrawal, Iftakhar Ahmad, Wasi Uddin Ahmad, Sina Ahmadi, Shafiuddin Rehan Ahmed, Natalie Ahn, Sumyeong Ahn, Vicent Ahuir, Kabir Ahuja, Lin Ai, Xi Ai, Ankit Aich, Laura Aina, Akiko Aizawa, Aswathy Ajith, Reina Akama, Pritom Saha Akash, Mohammad Akbari, Nader Akoury, Burak Aksar, Taha Aksu, Mousumi Akter, Mst Shapna Akter, Arjun

Reddy Akula, Hend Al-Khalifa, Hussein Al-Olimat, Badr AlKhamissi, Özge Alacam, Mehwish Alam, Belen Alastruey, Alon Albalak, Abdullah Albanyan, Chris Alberti, Vasiliy Alekseev, Georgios Alexandridis, Robin Jonathan Algayres, Asaad Alghamdi, Israa Alghanmi, Bashar Alhafni, Abdulaziz Alhamadani, Hamed Alhoori, Hassan Alhuzali, Emily Allaway, Eugenia Alleva, Tiago Almeida, Kenneth Alperin, Sawsan Alqahtani, Abdullah Alrajeh, Milad Alshomary, Maha Jarallah Althobaiti, Duygu Altinok, Rami Aly, Chiara Alzetta, Shmuel Amar, Bharat Ram Ambati, Maxime Amblard, Enrique Amigo, Saadullah Amin, Afra Amini, Maaz Amjad, Guozhen An, Haozhe An, Jisun An, Shengnan An, Ashish Anand, Nikhil Anand, Raviteja Anantha, Aparna Ananthasubramaniam, Rafael Anchiêta, Nicholas Andrews, Adrita Anika, Tatiana Anikina, Alan Ansell, Abrar Anwar, Xiang Ao, Emilia Apostolova, Alessio Palmero Aprosio, Erik Arakelyan, Matheus Araujo, Arturo Argueta, William Scot Armstrong, Hiba Arnaout, Akhil Arora, Arnav Arora, Ravneet Singh Arora, Siddhant Arora, Philip Arthur, Anjana Arunkumar, Mohammad Arvan, Viraat Aryabumi, Saurav Keshari Aryal, Elliott Ash, Zhenisbek Assylbekov, Md. Atabuzzaman, Duygu Ataman, John Atkinson-Abutridy, Yash Kumar Atri, Giuseppe Attanasio, Giuseppe Attardi, Aitziber Atutxa, Katherine Atwell, Lauriane Aufrant, Hayastan Avetisyan, Eleftherios Avramidis, Parul Awasthy, Busayo Awobade, Ayodele Awokoya, Gorka Azkune, Salah Aït-Mokhtar, Matthias ASSenmacher

Senthil Kumar B, Petr Babkin, Nguyen Bach, Sarkhan Badirli, Seongsu Bae, Ashutosh Baheti, Vikas Bahirwani, Seyed Ali Bahrainian, Bing Bai, Chongyang Bai, Fan Bai, Jiaxin Bai, Jinbin Bai, Jincheng Bai, Jun Bai, Long Bai, Xuefeng Bai, Yu Bai, Yushi Bai, Divya Jyoti Bajpai, JinYeong Bak, Amir Bakarov, Pedram Bakhtiarifard, Vidhisha Balachandran, Mithun Balakrishna, Oana Balalau, Ananth Balashankar, Gunjan Balde, Ioana Baldini, Mohammadreza Banaei, Juan M Banda, Dibyanayan Bandyopadhyay, Amar Banerjee, Atmadeep Banerjee, Namo Bang, Parikshit Bansal, Guangsheng Bao, Yu Bao, Yuwei Bao, Jorge Baptista, Kfir Bar, Claire Barale, Mohamad Hardyman Barawi, Adrien Barbaresi, Elham Barezi, Antonio Valerio Miceli Barone, Marco Baroni, Alberto Barrón-Cedeño, Marion Bartl, Sabine Bartsch, Sabyasachee Baruah, Pierpaolo Basile, Kinjal Basu, Somnath Basu Roy Chowdhury, Partha Basuchowdhuri, Ian Beaver, Björn Bebensee, Melika Behjati, Shabnam Behzad, Meriem Beloucif, Alejandro Benito-Santos, Himanshu Beniwal, Imene Bensalem, Gábor Berend, Leon Bergen, Maria Berger, Nathaniel Berger, Toms Bergmanis, Dario Bertero, Amanda Bertsch, Rasika Vinayak Bhalerao, Rohan V Bhambhoria, Rishabh Bhardwaj, Aditya Bhargava, Nirav Pravinbhai Bhatt, Abari Bhattacharya, Sumanta Bhattacharyya, Plaban Kumar Bhowmick, Rajarshi Bhowmik, Mukul Bhutani, Nikita Bhutani, Guanqun Bi, Qi Bi, Sirui Bi, Giovanni Maria Biancofiore, Irina Bigoulaeva, Arne Binder, Arianna Bisazza, Debmalya Biswas, Johannes Bjerva, Philippe Blache, Louis Blankemeier, Terra Blevins, Jelke Bloem, Michael Bloodgood, Victoria Bobicev, Ben Bogin, Joanne Boisson, Valeriia Bolotova, Alessandro Bondielli, Nadav Borenstein, Logan Born, Mihaela Bornea, Emanuela Boros, Cristina Bosco, Digbalay Bose, Robert Bossy, Kaj Bostrom, Houda Bouamor, Nadjet Bouayad-Agha, Zied Bouraoui, Andrey Bout, Maharaj Brahma, Faeze Brahman, Pavel Braslavski, Daniel Braun, Adrian M. P. Braşoveanu, Jacob Bremerman, Jonathan Brennan, Chris Brew, Thomas Brovelli, Hannah Brown, Henrico Bertini Brum, Yuqi Bu, Kyle Buettner, Nghi D. Q. Bui, Trung Bui, Laurie Burchell, Victor Bursztyn, Aslan Wong Butjamlong, Miriam Butt, Joan Byamugisha, Bill Byrne, Olga Bystrova, Jaeseok Byun, Necva Bölücü

Kishan K C, Sky CH-Wang, Pere-Lluís Huguet Cabot, Michele Cafagna, Samuel Cahyawijaya, Deng Cai, Erica Cai, Jason Cai, Mingzhu Cai, Weixin Cai, Xiangrui Cai, Yuxiang Cai, Zefan Cai, Ruken Cakici, Agostina Calabrese, Eduardo Calò, Nguyen Cam-Tu, Jose Camacho-Collados, Giovanni Campagna, Stefano Campese, Leonardo Campillos-Llanos, Daniel F Campos, Jon Ander Campos, Burcu Can, Boxi Cao, Jiarun Cao, Jie Cao, Lei Cao, Pengfei Cao, Qian Cao, Qingqing Cao, Rui Cao, Shuyang Cao, Yihan Cao, Yixuan Cao, Yu Cao, Yuan Cao, Spencer Caplan, Rémi Cardon, Lucien Carroll, Samuel Carton, Danilo Carvalho, Silvia Casola, Giovanni Cassani, Pierluigi Cassotti, Arie Cattan, Andrew Cattle, Paulo Cavalin, Francesco Cazzaro, Roberto Centeno, Dumitru-Clementin Cercel, Suchet Chachra, Dong-Kyu Chae, Haixia Chai, Heyan Chai,

Joyce Chai, Junyi Chai, Yekun Chai, Tuhin Chakrabarty, Tanmoy Chakraborty, Ilias Chalkidis, Kate Challis, Chi-Min Chan, Chunkit Chan, Hou Pong Chan, Zhangming Chan, Chandrahas, Buru Chang, Haw-Shiuan Chang, Serina Chang, Shuaichen Chang, Tyler A. Chang, Yingshan Chang, Rajen Chatterjee, Akshay Chaturvedi, Iti Chaturvedi, Aditi Chaudhary, Hardik Hansrajbhai Chauhan, Kushal Chawla, Ciprian Chelba, Emmanuel Chemla, Beiduo Chen, Catherine Chen, Chacha Chen, Chen Chen, Deli Chen, Derek Chen, Francine Chen, Fuxiang Chen, Guanhua Chen, Guanhua Chen, Guanliang Chen, Guanyi Chen, Hang Chen, Hanjie Chen, Hao Chen, Haotian Chen, Hui Chen, Huiyuan Chen, Hung-Ting Chen, Jiaao Chen, Jiangjie Chen, Jiaoyan Chen, Jiawei Chen, Jifan Chen, Jingqiang Chen, Jintai Chen, John Chen, Joya Chen, Junyang Chen, Keqin Chen, Kezhen Chen, Lei Chen, Liang Chen, Lihu Chen, Lin Chen, Lingwei Chen, Luoxin Chen, Maximillian Chen, Mei-Hua Chen, Meng Chen, Mingda Chen, Nan Chen, Nuo Chen, Nuo Chen, Pei Chen, Pinzhen Chen, Qian Chen, Qianglong Chen, Qiyuan Chen, Rui Chen, Shijie Chen, Shizhe Chen, Shuang Chen, Sihao Chen, Sishuo Chen, Tao Chen, Tao Chen, Tongfei Chen, Wei-Lin Chen, Wei-Rui Chen, Xiang Chen, Xiaojun Chen, Xiaoyin Chen, Xilun Chen, Xingran Chen, Xinhong Chen, Xiuying Chen, Yang Chen, Yangbin Chen, Yangyi Chen, Yanping Chen, Yi Chen, Yidong Chen, Ying Chen, Yirong Chen, Yiyi Chen, Yu Chen, Yubo Chen, Yue Chen, Yue Chen, Yulong Chen, Yun Chen, Yunmo Chen, Zeming Chen, Zeyuan Chen, Zhenghan Chen, Zhengyu Chen, Zhihong Chen, Zhiyu Chen, Zhiyu Chen, Zhongwu Chen, Zhuang Chen, Zichen Chen, Emily Cheng, Fei Cheng, Liying Cheng, Lu Cheng, Myra Cheng, Pengxiang Cheng, Qinyuan Cheng, Sijie Cheng, Tiffany Cheng, Weiwei Cheng, Xin Cheng, Xuxin Cheng, Yun Cheng, Zhi-Qi Cheng, Zifeng Cheng, Jianfeng Chi, Zewen Chi, Cheng-Han Chiang, David Chiang, Luis Chiruzzo, Hyun Chang Cho, Hyundong Justin Cho, Hyunsoo Cho, Jaemin Cho, Sangwoo Cho, Young Min Cho, Byung-Ju Choi, Jihun Choi, Minje Choi, Seungtaek Choi, Yejin Choi, Yun-Seok Choi, Kostadin Cholakov, Jaegul Choo, Harshita Chopra, Shubham Chopra, Sagnik Ray Choudhury, Arijit Ghosh Chowdhury, Jishnu Ray Chowdhury, Koel Dutta Chowdhury, Md Faisal Mahbub Chowdhury, Md Towhidul Absar Chowdhury, Lukas Christ, Chenhui Chu, Yun-Wei Chu, Zhendong Chu, Hsiu-Min Chuang, Yung-Sung Chuang, Jin-Woo Chung, Yi-Ling Chung, Mark Cieliebak, Alexandra Ciobotaru, Jorge Civera, Christian Clark, Christopher Clark, Christopher Clarke, Vincent Claveau, Ann Clifton, Anne Cocos, Davide Colla, Pedro Colon-Hernandez, Andrei Catalin Coman, Anna Corazza, Francesco Corcoglioniti, Camille Couturier, Benoit Crabbé, Mathias Creutz, Liam Cripwell, James Cross, Maxwell Crouse, Ganqu Cui, Leyang Cui, Peng Cui, Shaobo Cui, Shiyao Cui, Wenyao Cui, Yiming Cui, Rossana Cunha

Gautier Dagan, Deborah A. Dahl, Haixing Dai, Qin Dai, Wenliang Dai, Xiang Dai, Yi Dai, Yinpei Dai, Yong Dai, Daniel Dakota, Rumen Dangovski, Marina Danilevsky, Verna Dankers, Aswarth Abhilash Dara, Amitava Das, Mithun Das, Sarkar Snigdha Sarathi Das, Souvik Das, Sarthak Dash, Sam Davidson, Forrest Davis, Joseph Douglas Davison, Hillary Dawkins, Steve DeNeefe, Jay DeYoung, Alok Debnath, Zahra Delbari, Jean-Benoit Delbrouck, Marc Delcroix, Pieter Delobelle, Louise Deléger, Daryna Dementieva, Çağatay Demiralp, Dorottya Demszky, Mingkai Deng, Shijian Deng, Shumin Deng, Yang Deng, Yuntian Deng, Michael Denkowski, Jay Desai, Chris Develder, Joseph Dexter, Hira Dhamyal, Zonglin Di, Maria Pia Di Buono, Luca Di Liello, Marion Di Marco, Giorgio Maria Di Nunzio, Aissatou Diallo, Shizhe Diao, Bosheng Ding, Caiwen Ding, Chenchen Ding, Hantian Ding, Kaize Ding, Keyang Ding, Ruiqing Ding, Shuoyang Ding, Wenjian Ding, Wentao Ding, Yangruibo Ding, Yuning Ding, Zeyuan Ding, Ajay Divakaran, Anuj Diwan, Tanay Dixit, Nemanja Djuric, Phong Nguyen-Thuan Do, Quyet V. Do, Simon Dobnik, Sumanth Doddapaneni, Jonathan Dodge, Miguel Domingo, Bin Dong, Chenhe Dong, Haoyu Dong, Liu Wei Dong, MeiXing Dong, Ming Dong, Qianqian Dong, Qingxiu Dong, Ruihai Dong, Tiansi Dong, Xiangjue Dong, Xin Dong, Yuanzhe Dong, Vishnu Sashank Dorbala, Zi-Yi Dou, Jad Doughman, Timothy Dozat, Eduard Dragut, Andrew Drozdov, Jiangshu Du, Jinhua Du, Li Du, Li Du, Mengnan Du, Pan Du, Tianyu Du, Wanyu Du, Wei Du, Wenchao Du, Ye Du, Yifan Du, Yulun Du, Yupei Du, Brian DuSell, Dheeru Dua, Chaoqun Duan, Jiaxin Duan, Jinhao Duan, Junwen Duan, Sufeng Duan, Kevin Duh, Jonathan Dunn, Tejas Duseja, Ritam Dutt, Sourav Dutta, Sujan Dutta, Tomasz Dwojak, Gaël de Chalendar, Maddalen Lopez de Lacalle, David Martins de Matos, Andrea Gregor

de Varda, Éric Villemonte de la Clergerie

Sebastian Ebert, Aleksandra Edwards, Pavel Efimov, Koji Eguchi, Yo Ehara, Maud Ehrmann, Annerose Eichel, Vladimir Eidelman, Roald Eiselen, Yassir El Mesbahi, Samhaa R. El-Beltagy, Aparna Elangovan, Yanai Elazar, Heba Elfardy, Michael Elhadad, Micha Elsner, D. B. Emerson, Joseph Enguehard, Sugyeong Eo, Ori Ernst, Carlos Escolano, Luis Espinosa-Anke, Dominique Estival, Kilian Evang, Saad Ezzini

Alexander Fabbri, Marzieh Fadaee, Hossein Rajaby Faghihi, Fahim Faisal, Ge Fan, Jungwei Fan, Run-Ze Fan, Yao-Chung Fan, Yue Fan, Biaoyan Fang, Jinyuan Fang, Liri Fang, Qingkai Fang, Tianqing Fang, Wanlong Fang, Wei Fang, Xiang Fang, Yimai Fang, Yuwei Fang, Hossein Fani, Ibrahim Abu Farha, Effat Farhana, Nawshad Farruque, Amirhossein Farzam, Amany Fashwan, Jean-Philippe Fauconnier, Adam Faulkner, Amir Feder, Zichu Fei, Aarash Feizi, Nils Feldhus, Anna Feldman, Sergey Feldman, Virginia K. Felkner, Jia Hui Feng, Jianzhou Feng, Jiazhan Feng, Jingrong Feng, Shangbin Feng, Shutong Feng, Steven Y. Feng, Weixi Feng, Xiachong Feng, Yanlin Feng, Yi Feng, Yu Feng, Yujie Feng, Yunhe Feng, Zhangyin Feng, Paulo Fernandes, Nigel Fernandez, Raquel Fernández, Elisa Ferracane, Javier Ferrando, Rafael Ferreira, Manuel Vilares Ferro, Elisabetta Fersini, Besnik Fetahu, Alejandro Figueroa, Marcelo Finger, Matthew Finlayson, Mauajama Firdaus, Tim Fischer, Margaret M. Fleck, Michael Flor, Marco Fonseca, Jennifer Foster, Nathan Fradet, Marc Franco-Salvador, Thomas François, Flavius Frasincar, Dayne Freitag, Francesca Frontini, Deqing Fu, Haomin Fu, Jie Fu, Lisheng Fu, Peng Fu, Quchen Fu, Tingchen Fu, Tsu-Jui Fu, Xingyu Fu, Fumiyo Fukumoto, Richard Futrell, Michael Färber

Matteo Gabburo, Kata Gabor, Marco Gaido, Amit Gajbhiye, Lukas Galke, Leilei Gan, Yanglei Gan, Yujian Gan, Sudeep Gandhe, Ashwinkumar Ganesan, Ananya Ganesh, Govind Krishnan Gangadhar, Achyutarama R Ganti, William Gantt, Chang Gao, Chongyang Gao, Ge Gao, Hongyang Gao, Jiahui Gao, Jialin Gao, Jinhua Gao, Jun Gao, Jun Gao, Lingyu Gao, Mingqi Gao, Pengzhi Gao, Songyang Gao, Tianyu Gao, Xin Gao, Yifan Gao, Yingbo Gao, Ze-Feng Gao, Cristina Garbacea, Diego Garcia-Olano, Krishna Garg, Muskan Garg, Nicolas Garneau, Federico Gaspari, Judith Gaspers, Susan Gauch, Tanja Gaustad, Vagrant Gautam, Mengshi Ge, Ranxiang Ge, Suyu Ge, Tao Ge, Xiou Ge, Yixiao Ge, Michaela Geierhos, Christian Geishauser, Gael Gendron, Ariel Gera, Mozhdeh Gheini, Deepanway Ghosal, Kripabandhu Ghosh, Madhusudan Ghosh, Reshmi Ghosh, Rikhiya Ghosh, Sayontan Ghosh, Shrestha Ghosh, Sohom Ghosh, Soumitra Ghosh, Sourav Ghosh, Sreyan Ghosh, Lee Gillam, John Michael Giorgi, Salvatore Giorgi, Voula Giouli, Serge Gladkoff, Catalina Goanta, Ameya Godbole, Nathan Godey, Vaibhava Goel, Anna Goldie, Sujatha Das Gollapalli, Olga Golovneva, Jose Manuel Gomez-Perez, Jiaying Gong, Linyuan Gong, Shansan Gong, Zhuocheng Gong, Marcos André Gonçalves, Michael Eric Goodale, Sourabh Vasant Gothe, Akhilesh Deepak Gotmare, Isao Goto, Antoine Gourru, Venkata Subrahmanyan Govindarajan, Thamme Gowda, Kartik Goyal, Navita Goyal, Palash Goyal, Poonam Goyal, Nathan Green, Yulia Grishina, Adam Grycner, Stig-Arne Grönroos, Alex Gu, Jia-Chen Gu, Jiasheng Gu, Jing Gu, Xiaodong Gu, Xiaotao Gu, Yu Gu, Yue Gu, Yuxian Gu, Yuxuan Gu, Che Guan, Jian Guan, Saiping Guan, Xinyu Guan, Yi Guan, Yong Guan, Zihan Guan, Nuno M Guerreiro, Anchun Gui, Adrien Guille, Varun Gumma, Kalpa Gunaratna, James Gung, Tunga Gungor, Han Guo, Honglei Guo, Jiaqi Guo, Jiayan Guo, Jinyang Guo, Meiqi Guo, Quan Guo, Ruocheng Guo, Ruohao Guo, Shaoru Guo, Xin Guo, Xinnan Guo, Yanzhu Guo, Yinpeng Guo, Yue Guo, Yuhang Guo, Zhen Guo, Zixian Guo, Akshat Gupta, Amulya Gupta, Ankit Gupta, Ankita Gupta, Ashim Gupta, Ashish Gupta, Jai Gupta, Mithun Das Gupta, Prakhar Gupta, Raghav Gupta, Shashank Gupta, Suchin Gururangan, Bernal Jimenez Gutierrez, Ximena Gutierrez-Vasques, Jeremy Gwinnup, Jana Götze, Ramiro H. Gálvez, Carlos Gómez-Rodríguez

Le An Ha, Nizar Habash, Salim Hafid, Christopher Hahn, Joonghyuk Hahn, Zhen Hai, Samar Haider, Hossein Hajipour, Huda Hakami, Sherzod Hakimov, Harald Hammarström, Thierry Hamon, Chengcheng Han, Chi Han, Hojae Han, Jiuzhou Han, Kelvin Han, Sang-eun Han, Ting

Han, Ting Han, Wei Han, William Han, Xiaohui Han, Xiaotian Han, Xudong Han, Xue Han, Yo-Sub Han, Zhaowei Han, Zhen Han, Ziwen Han, Kunal Handa, Chung-Wei Hang, Viktor Hangya, Hongkun Hao, Tianyong Hao, Yongchang Hao, Momchil Hardalov, Fabrice Y Harel-Canada, John Harvill, Md. Arid Hasan, Monowar Hasan, Sadid A. Hasan, Maram Hasanain, Peter Hase, Taku Hasegawa, Chikara Hashimoto, Md. Mahadi Hassan, Sabit Hassan, Hans Ole Hatzel, Annette Hautli-Janisz, William N. Havard, Adi Haviv, Hiroaki Hayashi, Yoshihiko Hayashi, Amir Hazem, Ben He, Bin He, Guoxiu He, Jacqueline He, Jianfeng He, Jiangen He, Jie He, Jinzheng He, Kai He, Keqing He, Kun He, Liang He, Lihong He, Shizhu He, Tianxing He, Wanwei He, Wei He, Xingwei He, Xuanli He, Xuehai He, Yifan He, Yun He, Zexue He, Behnam Hedayatnia, Ming Shan Hee, Benjamin Heinzerling, William Barr Held, Leonhard Hennig, Lucas Torroba Hennigen, Yu-Jung Heo, Freddy Heppell, Daniel Hershcovich, Christian Heumann, Gerhard Heyer, Derrick Higgins, Stefan Hillmann, Tsutomu Hirao, Arnav Hiray, Namgyu Ho, Cuong Hoang, Julia Hockenmaier, Chris Hokamp, Pavan S Holur, Nils Holzenberger, Ukyo Honda, Giwon Hong, Jenny Hong, Pengyu Hong, Zhiqing Hong, Mark Hopkins, Ales Horak, Sho Hoshino, Tom Hosking, Md Mosharaf Hossain, Mohammad Javad Hosseini, Rui Hou, Wenjun Hou, Yifan Hou, Yu Hou, Yupeng Hou, Oumaima Hourrane, Alexander Hoyle, Jennifer Hsia, Cheng-Yu Hsieh, Chao-Chun Hsu, I-Hung Hsu, Yi-Li Hsu, Anwen Hu, Dou Hu, Guangneng Hu, Guimin Hu, Han Hu, Hanxu Hu, Jinyi Hu, Lijie Hu, Linmei Hu, Mengting Hu, Minda Hu, Po Hu, Songbo Hu, Vincent Tao Hu, Wei Hu, Xiang Hu, Xiaodan Hu, Xiaoyu Hu, Xuming Hu, Yibo Hu, Yuchen Hu, Yushi Hu, Yutong Hu, Zhe Hu, Zhiqiang Hu, Zhiwei Hu, Zhiyuan Hu, Ziniu Hu, Hang Hua, Wenyue Hua, Chao-Wei Huang, Chenyang Huang, Fei Huang, Haifeng Huang, Hen-Hsen Huang, James Y. Huang, Jen-tse Huang, Jerry Huang, Jiangping Huang, Jiaxin Huang, Jie Huang, Jimin Huang, Jin-Xia Huang, Kuan-Hao Huang, Kung-Hsiang Huang, Qiushi Huang, Quzhe Huang, Tenghao Huang, Xiaolei Huang, Xijie Huang, Xinting Huang, Yi-Ting Huang, Yinya Huang, Yufang Huang, Zhiqi Huang, Zhongqiang Huang, Zijian Huang, Ziyang Huang, John S Hudzina, Binyuan Hui, Wenyang Hui, Zheng Hui, Chia-Chien Hung, Kyunghoon Hur, Tin Van Huynh, Dae Yon Hwang, Truong Son Hy, Dongmin Hyun, Katharina Hämmerl, Ali Hürriyetoğlu

Ignacio Iacobacci, Adrian Iftene, Ryu Iida, Nikolai Ilinykh, Kenji Imamura, Joseph Marvin Imperial, Hirofumi Inaguma, Koji Inoue, Takashi Inui, Tatsuya Ishigaki, Shotaro Ishihara, Etsuko Ishii, Aminul Islam, Tunazzina Islam, Hayate Iso, Masaru Isonuma, Takumi Ito, Hamish Ivison, Tomoya Iwakura, Ran Iwamoto, Kenichi Iwatsuki, Arun Iyer, Roshni Iyer, Vivek Iyer, Peter Izsak

Cassandra L Jacobs, Labiba Jahan, Aashi Jain, Raghav Jain, Rishabh Jain, Sarthak Jain, Miloš Jakubíček, Shoaib Jameel, Abhik Jana, Eugene Jang, Hyeju Jang, Peter Jansen, Sujay Kumar Jauhar, Tommi Jauhiainen, Sébastien Jean, Nicolaas Paul Jedema, Fran Jelenić, Sophie Jentzsch, Sungho Jeon, Young-Seob Jeong, Kevin Jesse, Akshita Jha, Ananya Harsh Jha, Prince Jha, Harsh Jhamtani, Bin Ji, Kaixuan Ji, Seunghyun Ji, Shaoxiong Ji, Shengpeng Ji, Ziwei Ji, Chen Jia, Mengzhao Jia, Qi Jia, Zixia Jia, Xiangru Jian, Chao Jiang, Chengyue Jiang, Cong Jiang, Huiqiang Jiang, Jinhao Jiang, Jiyue Jiang, Junfeng Jiang, Jyun-Yu Jiang, Liting Jiang, Ming Jiang, Minhao Jiang, Nan-Jiang Jiang, Pengcheng Jiang, Ridong Jiang, Song Jiang, Tianyu Jiang, Wenbin Jiang, Xiang Jiang, Xiaotong Jiang, Yuxin Jiang, Zhihua Jiang, Zhuoren Jiang, Zhuoxuan Jiang, Cathy Jiao, Fangkai Jiao, Pengfei Jiao, Wenxiang Jiao, Yizhu Jiao, Cheng Jiayang, Di Jin, Li Jin, Peng Jin, Qiao Jin, Shuning Jin, Woojeong Jin, Xiaomeng Jin, Yiping Jin, Yiqiao Jin, Zhuoran Jin, Zijian Jin, Ishan Jindal, Baoyu Jing, Liqiang Jing, Hwiyeol Jo, Arnaud Joly, Erik Jones, Kenneth Joseph, Abhinav Joshi, Aditya Joshi, Brihi Joshi, Nitish Joshi, Omkar Jayant Joshi, Rishabh Joshi, Jaap Jumelet, Gurusha Juneja, Juho Jung, Minjoon Jung, Juraj Juraska, Prathyusha Jwalapuram

Jad Kabbara, Mohsinul Kabir, Kazuma Kadowaki, Ivana Kajic, Mihir Kale, Oren Kalinsky, Katikapalli Subramanyam Kalyan, Danial Kamali, Ehsan Kamalloo, Amita Kamath, Nishant Kambhatla, Hiroshi Kanayama, Kamil Kanclerz, Arun Kandoor, Gi-Cheon Kang, Minki Kang, Xiaoxi Kang, Yujin Kang, Yash Kankanampati, Nithish Kannen, Ryuji Kano, Tapas Kanungo, Debanjana Kar, Pinar Karagoz, Giannis Karamanolakis, Siddharth Karamcheti, Younes Karimi, Payam Karisani,

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Karn, Marzena Karpinska, Pradeep Karuturi, Siva Rajesh Kasa, Omid Kashefi, Abhay Kashyap, Abhinav Ramesh Kashyap, Zdeněk Kasner, Aly M. Kassem, Nora Kassner, Uri Katz, Simerjot Kaur, Noriaki Kawamae, Efsun Kayi, Hideto Kazawa, Ashkan Kazemi, Amirhossein Kazemnejad, Zixuan Ke, Akhil Kedia, Sedrick Scott Keh, Amr Keleg, Neha Nayak Kennard, Casey Kennington, Baber Khalid, Salam Khalifa, Sammy Khalife, Aditi Khandelwal, Shima Khanehzar, Simran Khanuja, Subhendu Khatuya, Vivek Khetan, Md Tawkat Islam Khondaker, Kyung Seo Ki, Bugeun Kim, Chaehyeong Kim, Dohee Kim, Gangwoo Kim, Gunhee Kim, Gyuhak Kim, Gyuwan Kim, Haven Kim, Hazel Kim, Hongjin Kim, Hyounghun Kim, Hyunjae Kim, Hyunwoo Kim, Jiho Kim, Joo-Kyung Kim, Joshua Yee Kim, Jung-jae Kim, Junho Kim, Junyeob Kim, Kang-Min Kim, Kangil Kim, Kyungho Kim, Minsoo Kim, Minsoo Kim, Minsu Kim, Seungone Kim, Sungdong Kim, Taehwan Kim, Taeuk Kim, Takyoung Kim, Yeachan Kim, Yekyung Kim, YoungBin Kim, Youngwoo Kim, Yunsu Kim, Zae Myung Kim, Yasutomo Kimura, Milton King, Tracy Holloway King, Christo Kirov, Denis Kiselev, Shun Kiyono, Christopher Klamm, Julien Kloetzer, Miyoung Ko, Goro Kobayashi, Thomas H Kober, Jan Kocon, Prashant Kodali, Svetla Peneva Koeva, Mare Koit, Kanako Komiya, Grzegorz Kondrak, Cunliang Kong, Lingkai Kong, Selcuk Kopru, Michalis Korakakis, Mandy Barrett Korpusik, Katsunori Kotani, Ana Kotarcic, Suraj Nandkishor Kothawade, Fajri Koto, Alexander Kotov, Manolis Koubarakis, Vasiliki Kougia, Mahnaz Koupaee, Venelin Kovatchev, Md Kowsher, Ivan Koychev, Matthias Kraus, Simon Krek, Brigitte Krenn, Amrith Krishna, Kalpesh Krishna, Kundan Krishna, Adit Krishnan, Canasai Kruengkrai, Udo Kruschwitz, Wojciech Maciej Kryscinski, Da Kuang, Marek Kubis, Andrei Kucharavy, Seth Kulick, Ashish Kulkarni, Atharva Kulkarni, Mayank Kulkarni, Ashutosh Kumar, Nitesh Kumar, Rahul Kumar, Revant Kumar, Sachin Kumar, Shankar Kumar, Shanu Kumar, Shivani Kumar, Sujit Kumar, Vishwajeet Kumar, Vivek Kumar, Sadhana Kumaravel, Anoop Kunchukuttan, Tuhin Kundu, Maria Kunilovskaya, Tatsuki Kuribayashi, Mikko Kurimo, Kemal Kurniawan, Guy Kushilevitz, Beong-woo Kwak, Haewoon Kwak, Jin Myung Kwak, Sunjun Kweon, Arne Köhn

Philippe Laban, Sofie Labat, Matthieu Labeau, Yanis Labrak, Aritra Kumar Lahiri, Allison Lahnala, Huiyuan Lai, Kenneth Lai, Viet Dac Lai, Wen Lai, Surafel Melaku Lakew, Kushal Lakhotia, Yash Kumar Lal, Tsz Kin Lam, Wai Lam, Hemank Lamba, Vasileios Lampos, Gerasimos Lampouras, Wuwei Lan, Yihuai Lan, Yunshi Lan, Hao Lang, David Langlois, Maurice Langner, Mateusz Lango, Mirella Lapata, Issam H. Laradji, Stefan Larson, Kornel Laskowski, Leo Laugier, Alexandra Lavrentovich, Dawn Lawrie, Dung D. Le, Hung Le, Phong Le, Phuong-Hang Le, Thang Le, Joseph Le Roux, Kevin Leach, Changmin Lee, Chia-Hsuan Lee, Deokjae Lee, Dong-Ho Lee, Dong-Hyun Lee, Donghun Lee, Dongkyu Lee, Gibbeum Lee, Hojin Lee, Hwanhee Lee, Hyungyung Lee, Hyunju Lee, I-Ta Lee, Jae Sung Lee, Jae Hee Lee, Jaeseong Lee, Ji-Ung Lee, Jihwan Lee, Jinsik Lee, Jooyoung Lee, Jun-Min Lee, Jungseob Lee, Keeheon Lee, Koanho Lee, Lung-Hao Lee, Mingyu Lee, Minwoo Lee, Nayeon Lee, Nayeon Lee, Sangyong Lee, Yoonjoo Lee, Young-Suk Lee, Yuanyuan Lei, Camelia Lemnaru, Sicong Leng, Pietro Lesci, Ran Levy, Sharon Levy, Anqi Li, Bangqi Li, Baoli Li, Bei Li, Belinda Z. Li, Bin Li, Bin Li, Bo Li, Bobo Li, Chen Li, Cheng Li, Cheng-Te Li, Chengming Li, Chong Li, Chuang Li, Dawei Li, Dongfang Li, Dongyang Li, Guanlin Li, Hao Li, Haochen Li, Haonan Li, Haoqi Li, Haoran Li, Huao Li, Jiacheng Li, Jialu Li, Jiangnan Li, Jiangtong Li, Jiaqi Li, Jiaxuan Li, Jiazhao Li, Jing Li, Jingjing Li, Jinpeng Li, Jiyi Li, Junhui Li, Junyi Li, Keyi Li, Kun Li, Lei Li, Li Erran Li, Linjing Li, Linyang Li, Mark Junjie Li, Mengze Li, Miao Li, Mingchen Li, Mingda Li, Na Li, Peifeng Li, Peng Li, Peng Li, Qi Li, Qi Li, Qian Li, Qintong Li, Rongsheng Li, Ru Li, Rui Li, Ruifan Li, Ruizhe Li, Runjia Li, Runnan Li, Sha Li, Shanshan Li, Shaobo Li, Sheng Li, Shengjie Li, Shichen Li, Shiyang Li, Shuangyin Li, Shujun Li, Shuyang Li, Shuyue Stella Li, Si Li, Siheng Li, Tianjian Li, Tianyi Li, Wei Li, Wei Li, Wei Li, Weixian Waylon Li, Wenchang Li, Wenxi Li, Wenyan Li, Xia Li, Xiang Li, Xiang Li, Xiang Lisa Li, Xiang Li, Xiaonan Li, Xiaoya Li, Xin Li, Xingxuan Li, Xinjian Li, Xintong Li, Xinxin Li, Xiuxing Li, Yafu Li, Yang Li, Yanyang Li, Yanzeng Li, Yanzhou Li, Yaoyiran Li, Yichuan Li, Yinghui Li, Yingjie Li, Yingya Li, Yitian Li, Yiyuan Li, Yizhi Li, Yu Li, Yuan-Fang Li, Yuanxi Li, Yue Li, Yufei Li, Yuncong Li, Zekun Li, Zhaohui Li, Zhaoyi Li, Zhixin Li, Ziheng

Li, Zizhong Li, Yixin Lian, Chao-Chun Liang, Davis Liang, Di Liang, Hongru Liang, Ke Liang, Sheng Liang, Xinnian Liang, Yunlong Liang, Yuzhi Liang, Zhengzhong Liang, Zhenwen Liang, Baohao Liao, I-Bin Liao, Siyu Liao, Anna Liednikova, Veronica Liesaputra, Gilbert Lim, Jia Peng Lim, Jungwoo Lim, Kwan Hui Lim, Tomasz Limisiewicz, Peerat Limkonchotiwat, Haitao Lin, Haokun Lin, Hongzhan Lin, Inna Wanyin Lin, Kevin Lin, Kevin Qinghong Lin, Li Lin, Lucy H. Lin, Nankai Lin, Peiqin Lin, Qika Lin, Sheng-Chieh Lin, Ting-En Lin, Victoria Lin, Wei Lin, Weizhe Lin, Wenye Lin, Ying-Jia Lin, Zhenxi Lin, Stephan Linzbach, Tal Linzen, Gili Lior, Pierre Lison, Onkar Rupesh Litake, Diane Litman, Robert Litschko, Aiwei Liu, Alisa Liu, Ao Liu, Bing Liu, Boyang Liu, Chen Cecilia Liu, Chi-Liang Liu, Chunhua Liu, Congcong Liu, Danni Liu, Dexi Liu, Emmy Liu, Fangchao Liu, Fenglin Liu, Guangliang Liu, Guisheng Liu, Han Liu, Hengyu Liu, Hui Liu, Hui Liu, Ji Liu, Jiangming Liu, Jiawei Liu, Jiduan Liu, Jie-Jyun Liu, Junhao Liu, Junpeng Liu, Lei Liu, Lemao Liu, Linqing Liu, Meizhen Liu, Meng Liu, Ming Liu, Minqian Liu, Nayu Liu, Nelson F. Liu, Peng Liu, Qianchu Liu, Qiang Liu, Qin Liu, Shuaiqi Liu, Shuqi Liu, Song Liu, Tianyu Liu, Wei Liu, Wei Liu, Weiyu Liu, Wenqiang Liu, Xiang Liu, Xiangyang Liu, Xiao Liu, Xiaoyuan Liu, Xiaoze Liu, Xin Liu, Xiping Liu, Xubo Liu, Xuye Liu, Yang Liu, Yang Liu, Ye Liu, Yezi Liu, Yihong Liu, Yinhong Liu, Yixin Liu, Yong Liu, Yongbin Liu, Yonghao Liu, Yongkang Liu, Yuanxin Liu, Yuanxing Liu, Yue Liu, Zeming Liu, Zeyu Liu, Zhe Liu, Zhenghao Liu, Zhengyuan Liu, Zhijian Liu, Zhu Liu, Zitao Liu, Zixuan Liu, Ziyi Liu, Zuozhu Liu, Robert Lo, Robert L. Logan IV, Lajanugen Logeswaran, Quanyu Long, Shangbang Long, Wanqiu Long, Sampath Lonka, Lucelene Lopes, Marcos Lopes, Jaime Lorenzo-Trueba, Chao Lou, Renze Lou, Natalia V Loukachevitch, Holy Lovenia, Brian Lu, Di Lu, Hongyuan Lu, Jiaying Lu, Jing Lu, Jinghui Lu, Jinliang Lu, Junru Lu, Peng Lu, Weiming Lu, Wenpeng Lu, Xiaolei Lu, Xinyuan Lu, Xuesong Lu, Yao Lu, Yichao Lu, Yu Lu, Yu Lu, Yuxuan Lu, Nurul Lubis, Li Lucy, Cheng Luo, Guoqing Luo, Haoran Luo, Haozheng Luo, Hongyin Luo, Jiaming Luo, Jinqi Luo, Junyu Luo, Ling Luo, Linhao Luo, Man Luo, Renqian Luo, Xiao Luo, Ziyang Luo, Jordi Luque, Kelvin Luu, Shangwen Lv, Zheqi Lv, Chen Lyu, Chenyang Lyu, Hanjia Lyu, Qing Lyu, Weimin Lyu, Xinglin Lyu, Yiwei Lyu, Yougang Lyu, Ziyu Lyu, Samuel Läubli

Meryem M'hamdi, Chenkai Ma, Cong Ma, Dan Ma, Fukun Ma, Huifang Ma, Jing Ma, Kaixin Ma, Lianbo Ma, Mingyu Derek Ma, Nianzu Ma, Ruotian Ma, Siliang Ma, Suyu Ma, Tengfei Ma, Xiao Ma, Xinbei Ma, Xinyin Ma, Xueguang Ma, Xutai Ma, Yingwei Ma, Yubo Ma, Yukun Ma, Zhixin Ma, Ziqiao Ma, Rooweither Mabuya, Jakub Macina, Aman Madaan, Avinash Madasu, Mounica Maddela, Margot Madina, Brielen Madureira, Rahmad Mahendra, Ayush Maheshwari, Adnan Mahmood, Wolfgang Maier, Peter Makarov, Aaron Maladry, Prodromos Malakasiotis, Bhavitvya Malik, Valentin Malykh, Hieu Man, Marta Marchiori Manerba, Pranav Maneriker, Pranav Mani, Irene L Manotas, Saab Mansour, Ramesh Manuvinakurike, Kelong Mao, Qianren Mao, Wenji Mao, Zhendong Mao, Zhiming Mao, Kelly Marchisio, Piotr Mardziel, Katerina Margatina, Alex Marin, Stella Markantonatou, Antonis Maronikolakis, Santiago Marro, Eugenio Martinez-Camara, Juan Martinez-Romo, Bruno Martins, Pedro Henrique Martins, Claudia Marzi, Mihai Masala, Laura Mascarell, Tessa Masis, Sarah Masud, Sandeep Mathias, Sérgio Matos, Yoshitomo Matsubara, Yuichiroh Matsubayashi, Takuya Matsuzaki, Evgeny Matusov, Kaushal Kumar Maurya, Amir Mazaheri, Sahisnu Mazumder, John Philip McCrae, David D. McDonald, Euan McGill, Denis Jered McInerney, Bridget T. McInnes, Zoran Medić, Alexander Mehler, Nikhil Mehta, Sanket Vaibhav Mehta, Stephen Meisenbacher, Clara Meister, Dheeraj Mekala, Merve Unlu Menevse, Telmo Menezes, Chuan Meng, Rui Meng, Rui Meng, Yu Meng, Yuanliang Meng, Zhao Meng, Rakesh R Menon, Elena Merdjanovska, Paola Merlo, Md Messal Monem Miah, Lin Miao, Yisong Miao, Alessio Miaschi, Timothee Mickus, Lesly Miculicich, Margot Mieskes, Jeremiah Milbauer, Filip Miletic, Aleksandra Miletić, Alice Millour, Hye-Jin Min, Zeping Min, Hideya Mino, Andrei Mircea, Santiago Miret, Seyed Abolghasem Mirroshandel, Roshanak Mirzaee, Abhijit Mishra, Prakamya Mishra, Shubhanshu Mishra, Swaroop Mishra, Kanishka Misra, Mitch Paul Mithun, Ekata Mitra, Ashish Mittal, Vibhu O. Mittal, Yasuhide Miura, So Miyagawa, Yusuke Miyao, Moran Mizrahi, Fengran Mo, Tong Mo, Wentao Mo, Yijun Mo, Daichi Mochihashi, Daniela Moctezuma, Ali Modarressi, Sandip Modha, Aditya Mogadala, Nikita Moghe, Mahmoud Mo-

hammadi, Alireza Mohammadshahi, Hosein Mohebbi, Behrang Mohit, Tasnim Mohiuddin, Luis Gerardo Mojica, Negar Mokhberian, Diego Molla, Nicholas Monath, Erfan Moosavi Monazzah, Sneha Mondal, Ali Montazeralghaem, Manuel Montes, Hyeonseok Moon, Jong Hak Moon, Lori Moon, Ray Mooney, Samraj Moorjani, Goncalo Mordido, Antonio Moreno-Ortiz, Antonio Moreno-Sandoval, Yusuke Mori, Véronique Moriceau, Gaku Morio, Makoto Morishita, Somaye Moslemnejad, Xiangyang Mou, Basel Mousi, Maximilian Mozes, Frank Martin Mtumbuka, Hamdy Mubarak, Aashiq Muhamed, Shashank Mujumdar, Anjishnu Mukherjee, Rajdeep Mukherjee, Sandeep Sricharan Mukku, Sai Munikoti, Dragos Stefan Munteanu, Saliha Muradoglu, Koji Murakami, Masayasu Muraoka, Kenton Murray, Rudra Murthy, Karthik Murugadoss, Emir Muñoz, Alberto Muñoz-Ortiz, Agnieszka Mykowiecka, Sheshera Mysore

Seung-Hoon Na, Farah Nadeem, Nona Naderi, Sreyashi Nag, Mayank Nagda, Aakanksha Naik, Saeed Najafi, Tetsuji Nakagawa, Yuta Nakashima, Christoforos Nalmpantis, Sungjin Nam, Marcin Namysl, Guoshun Nan, Linyong Nan, Abhilash Nandy, Tarek Naous, Diane Napolitano, Jason Naradowsky, Sharan Narasimhan, Arun Balajiee Lekshmi Narayanan, Yaswanth Narsupalli, Shahrzad Naseri, Vivi Nastase, Anandhavelu Natarajan, Costanza Navarretta, Tapas Nayak, Mojtaba Nayyeri, Carina Suzana Negreanu, Joshua Nemecek, Poli Nemkova, Graham Neubig, Günter Neumann, Mariana Neves, Benjamin Newman, Sina Bagheri Nezhad, Dianwen Ng, Axel-Cyrille Ngonga Ngomo, Chien Van Nguyen, Duc-Vu Nguyen, Hoang H Nguyen, Huy V. Nguyen, Huyen Nguyen, Kiet Van Nguyen, Long HB Nguyen, Nhung T.H. Nguyen, Thanh-Tung Nguyen, Tin Trung Nguyen, Truc-Vien T. Nguyen, Trung Hieu Nguyen, Tu Nguyen, Tung Nguyen, Van Bach Nguyen, Hoang-Quoc Nguyen-Son, Jingwei Ni, Jinjie Ni, Minheng Ni, Pin Ni, Shiwen Ni, Zhaoheng Ni, Iftitahu Ni'mah, Massimo Nicosia, Ercong Nie, Hongyi Nie, Lunyiu Nie, Ping Nie, Shaoliang Nie, Yixin Nie, Yuxiang Nie, Zhijie Nie, Malvina Nikandrou, Sofia Nikiforova, Irina Nikishina, Dmitry Nikolaev, Vassilina Nikoulina, Lasguido Nio, Kosuke Nishida, Noriki Nishida, Masaaki Nishino, Toru Nishino, Hao Niu, Jingcheng Niu, Runliang Niu, Xing Niu, Tadashi Nomoto, Kolby Nottingham, Jekaterina Novikova, Krenare Pireva Nuci, Pierre Nugues, Nasheen Nur, Sarana Nutanong, Claire Nédellec, Aurélie Névéol

Alexander O'Connor, Jose Ochoa-Luna, Stephan Oepen, Bahadorreza Ofoghi, Maciej Ogrodniczuk, Kelechi Ogueji, Jungwoo Oh, Minsik Oh, Mayumi Ohta, Kiyonori Ohtake, Atul Kr Ojha, Yui Oka, Tsuyoshi Okita, Inez Okulska, Eda Okur, Hugo Gonçalo Oliveira, Montse Cuadros Oller, Kaustubh Olpadkar, Ali Omrani, Byung-Won On, Brian David Ondov, Donovan Ong, Yasumasa Onoe, Andreas Opedal, Juri Opitz, Matan Orbach, Riccardo Orlando, Aitor Ormazabal, Yohei Oseki, Wolfgang Otto, Hiroki Ouchi, Jessica Ouyang, Yawen Ouyang, Risako Owan

Vishakh Padmakumar, Karthik Padthe, Artidoro Pagnoni, Vardaan Pahuja, Liu Pai, Santanu Pal, Sayantan Pal, Vaishali Pal, Alexis Palmer, Huitong Pan, Jiayi Pan, Liangming Pan, Youcheng Pan, Zhufeng Pan, Alexander Panchenko, Saurabh Kumar Pandey, Jianhui Pang, Lu Pang, Sheena Panthaplackel, Isabel Papadimitriou, Konstantinos Papakostas, Nikos Papasarantopoulos, Paolo Papotti, Emerson Cabrera Paraiso, Letitia Parcalabescu, Antonio Pareja-Lora, Tanmay Parekh, Shantipriya Parida, Pierre-Henri Paris, Chan Young Park, Eunkyung Park, Hyunwoo Park, Jae Sung Park, Jun-Hyung Park, Jungsoo Park, Kunwoo Park, Lucy Park, Seo Yeon Park, Shinwoo Park, Sungjin Park, Woohyun Park, Youngja Park, Panupong Pasupat, Arkil Patel, Maitreya Patel, Raj Nath Patel, Sapan Kirit Patel, Barun Patra, Braja Patra, Jasabanta Patro, Lincy Pattanaik, Parth Patwa, Siddharth Patwardhan, Indraneil Paul, Adam Pauls, Silviu Paun, Lucas Pavanelli, Ellie Pavlick, John Pavlopoulos, Vera Pavlova, Sarah Ruth Brogden Payne, Pavel Pecina, Jiahuan Pei, Jiaxin Pei, Zhengqi Pei, Stephan Peitz, Olga Pelloni, Boci Peng, Han Peng, Hao Peng, Haoyuan Peng, Min Peng, Puyuan Peng, Qiwei Peng, Qiyao Peng, Siyao Peng, Wei Peng, Xi Peng, Xutan Peng, Yifan Peng, Martin Pereira, Carla Perez-Almendros, Gabriele Pergola, Charith Peris, Francesco Periti, Daniel J Perry, Stanislav Peshterliev, Slav Petrov, Pavel Petrushkov, Sandro Pezzelle, Thang M. Pham, Thang Chau Phan, Jason Phang, Fred Philippy, Bornali Phukon, Matúš Pikuliak, Juan Pino, Dhivya Piraviperumal, Telmo Pires, Flammie A Pirinen, Jakub Pisko-

rski, Priya Pitre, Flor Miriam Plaza del Arco, Joan Plepi, Bryan A. Plummer, Lahari Poddar, Massimo Poesio, Marco Polignano, Simone Paolo Ponzetto, Ian Porada, Beatrice Portelli, Sahitya Potluri, Christopher Potts, Aniket Pramanick, Animesh Prasad, Archiki Prasad, Radityo Eko Prasojo, Adithya Pratapa, Gabriele Prato, Judita Preiss, Priyanshu Priya, Michal Ptaszynski, Dongqi Pu, Giulia Pucci, Ratish Puduppully, Rajkumar Pujari, Stephen Pulman, Sukannya Purkayastha, Alberto Purpura, Valentina Pyatkin, Juan Antonio Pérez-Ortiz

Ehsan Qasemi, Ji Qi, Jiexing Qi, Mengnan Qi, Wang Qi, Chen Qian, Hongjin Qian, Yong Qian, Yujie Qian, Yushan Qian, Linbo Qiao, Qiao Qiao, Shuofei Qiao, Yaqiong Qiao, Chengwei Qin, Chuan Qin, Jinghui Qin, Kechen Qin, Libo Qin, Yanxia Qin, Yujia Qin, Chen Qiu, Haoyi Qiu, Jielin Qiu, Linlu Qiu, Long Qiu, Mengyang Qiu, Shuwen Qiu, Xihe Qiu, Xin Ying Qiu, Moreno La Quatra

Ella Rabinovich, Daniele Paolo Radicioni, Luca Ragazzi, Preethi Raghavan, Mizanur Rahman, Sajjadur Rahman, Sunny Rai, Md Nishat Raihan, Lisa Raithel, Nishant Raj, Sara Rajaee, Kanagasabai Rajaraman, Shahab Raji, Heri Ramampiaro, Owen Rambow, Juan Ramirez-Orta, Jitenkumar Babubhai Rana, Leonardo Ranaldi, Surangika Ranathunga, Priya Rani, Dongning Rao, Ahmad Rashid, Hannah Rashkin, Vikas Raunak, Andrea Amelio Ravelli, Shauli Ravfogel, Manikandan Ravikiran, Srinivas Ravishankar, Bhanu Pratap Singh Rawat, Arijit Ray, Avik Ray, Soumya Ray, Shaina Raza, Anastasia Razdaibiedina, Evgeniia Razumovskaia, Traian Rebedea, Gabor Recski, Michael Regan, Aishwarya Naresh Reganti, Georg Rehm, Markus Reiter-Haas, Navid Rekabsaz, Da Ren, HaoPeng Ren, Liliang Ren, Pengjie Ren, Shuhuai Ren, Yuxin Ren, Steven J Rennie, Ashwathy T. Revi, Eugénio Ribeiro, Leonardo F. R. Ribeiro, Mattia Rigotti, Matīss Rikters, Parker Riley, Fabio Rinaldi, Ruty Rinott, Anthony Rios, Yara Rizk, Mathieu Roche, Alvaro Rodrigo, Melissa Roemmele, Mahdin Rohmatillah, Paul Roit, Lina Maria Rojas-Barahona, Roland Roller, Julia Romberg, Julien Romero, Kevin Ros, Domenic Rosati, Jan Rosendahl, Guy D. Rosin, Joe Cheri Ross, Robert Ross, Guy Rotman, Paul Rottger, Mozhdeh Rouhsedaghat, Dmitri Roussinov, Bryan R. Routledge, Aniruddha Roy, Kashob Kumar Roy, Shamik Roy, Soumyadeep Roy, Sudipta Singha Roy, Sumegh Roychowdhury, Benjamin Rozonoyer, Rimvydas Rubavicius, Daniel Onoro Rubio, Koustav Rudra, Amina Mardiyyah Rufai, Federico Ruggeri, Ramon Ruiz-Dolz, Mukund Rungta, Thomas Ruprecht, Alexander M Rush, Irene Russo, Hee Jung Ryu, Susanna Rücker

Ramaneswaran S, Hadeel Saadany, Zaina J. Z. Saadeddin, Arkadiy Saakyan, Masoud Jalili Sabet, Caroline Sabty, Mobashir Sadat, Arka Sadhu, Philipp Sadler, Sahar Sadrizadeh, Mehrnoosh Sadrzadeh, Marzieh Saeidi, Mustafa Safdari, Alsu Sagirova, Monjoy Saha, Sougata Saha, Swarnadeep Saha, Tanay Kumar Saha, Saurav Sahay, Sovan Kumar Sahoo, Oscar Sainz, Sakriani Sakti, Mohammadreza Salehi, Elizabeth Salesky, Vishal Vivek Saley, Avneesh Saluja, Pranay Reddy Samala, Niloofar Safi Samghabadi, Farhan Samir, Abhilasha Sancheti, Ramses J Sanchez, Vicente Ivan Sanchez Carmona, Anushka Sandesara, Jivnesh Sandhan, Sashank Santhanam, Andrea Santilli, Diana Santos, Bishal Santra, Sebastin Santy, Maarten Sap, Irina Saparina, Abulhair Saparov, Maya Sappelli, Xabier Saralegi, Chayan Sarkar, Rajdeep Sarkar, Rupak Sarkar, Ritesh Sarkhel, Parth Sarthi, Gabriele Sarti, Felix Sasaki, Minoru Sasaki, Shota Sasaki, Ryohei Sasano, MSVPJ Sathvik, Giorgio Satta, Danielle Saunders, Rohit Saxena, Michael Saxon, Kevin Scaria, Frank Schilder, David Schlangen, Viktor Schlegel, Jörg Schlötterer, Helmut Schmid, Fabian David Schmidt, Robin M. Schmidt, Martin Schmitt, Tyler Schnoebelen, Florian Schottmann, Hendrik Schuff, William Schuler, Claudia Schulz, Elliot Schumacher, Raphael Schumann, Diarmuid O Seaghdha, Anastasiia Sedova, Kyle Seelman, Nasredine Semmar, Indira Sen, Sailik Sengupta, Shubhashis Sengupta, Jaehyung Seo, Seungmin Seo, Royal Sequiera, Sofia Serrano, Agam Shah, Ankit Shah, Muhammad A Shah, G. M. Shahariar Shibli, Cory Shain, Igor Shalyminov, Erfan A Shams, Chao Shang, Mingyue Shang, Wenbo Shang, Chen Shani, Hanyin Shao, Jie Shao, Yijia Shao, Yujie Shao, Natalie Shapira, Ori Shapira, Abhishek Sharma, Aditya Sharma, Ashish Sharma, Drishti Sharma, Karishma Sharma, Piyush Sharma, Prawaal Sharma, Roshan Sharma, Shivam Sharma, Serge Sharoff, Shuaijie She, Artem Shelmanov, Hua Shen, Jiaming Shen, Jocelyn

J Shen, Lingfeng Shen, Shiqi Shen, Siqi Shen, Tianhao Shen, Xin Shen, Ying Shen, Yongliang Shen, Yuming Shen, Zejiang Shen, Zhengyuan Shen, Emily Sheng, Jiawei Sheng, Qiang Sheng, Quan Z. Sheng, Zhecheng Sheng, Ashish Shenoy, Tom Sherborne, Akshay Krishna Sheshadri, Chen Shi, Jihao Shi, Kaize Shi, Ning Shi, Peng Shi, Shuming Shi, Tao Shi, Tianze Shi, Xiao Shi, Zhan Shi, Zhengxiang Shi, Tomohide Shibata, Gyu-Ho Shin, Kazutoshi Shinoda, Takahiro Shinozaki, Prashant Shiralkar, Harry Shomer, Ziyi Shou, Mohit Shridhar, Ritvik Shrivastava, Dong Shu, Kai Shu, Raphael Shu, Yuxuan Shu, Ruihao Shui, Zeren Shui, KaShun Shum, Chenglei Si, Jiasheng Si, Shuzheng Si, Anthony Sicilia, A.B. Siddique, Melanie Siegel, Ankur Sikarwar, Sandipan Sikdar, Max Silberztein, João Silva, Kanishka Silva, Stefano Silvestri, Robert Sim, Patrick Simianer, Abhishek Singh, Harman Singh, Ishika Singh, Jyotika Singh, Mukul Singh, Shubhankar Singh, Telem Joyson Singh, Sneha Singhania, Koustuv Sinha, Olivier Siohan, Amy Siu, Steven Skiena, Victor Skobov, Aviv Slobodkin, Noah A. Smith, Artem Sokolov, Elena Sokolova, Luca Soldaini, Amir Soleimani, Aina Garí Soler, Ilia Sominsky, Pia Sommerauer, Junyoung Son, Youngseo Son, Sheetal S. Sonawane, Feifan Song, Haiyue Song, Hyun-Je Song, Kaitao Song, Linqi Song, Nirui Song, Ran Song, Rui Song, Yifan Song, Zhenqiao Song, Nikita Soni, Sandeep Soni, Alexey Sorokin, Dmitry Sotnikov, Anna Sotnikova, Sajad Sotudeh, Marlo Souza, Gerasimos Spanakis, Timo Spinde, Cesare Spinoso-Di Piano, Andreas Spitz, Makesh Narsimhan Sreedhar, Arvind Krishna Sridhar, Sharath Nittur Sridhar, Mukund Srinath, Gokul Srinivasagan, Balaji Vasan Srinivasan, Pranesh Srinivasan, Tejas Srinivasan, Aarohi Srivastava, Ankit Kumar Srivastava, Aseem Srivastava, Ieva Staliunaite, Dominik Stammbach, Karolina Stanczak, Marija Stanojevic, Gabriel Stanovsky, David Stap, Katherine Stasaski, Julius Steen, Shane Steinert-Threlkeld, Georg Stemmer, Elias Stengel-Eskin, Andreas Stephan, Zachary Stine, Alessandro Stolfo, Shane Storks, Tomek Strzalkowski, Phillip Benjamin Ströbel, Sara Stymne, Sebastian Stüker, Hsuan Su, Jinyan Su, Miao Su, Qiang Su, Ruolin Su, Xiangdong Su, Xin Su, Xuefeng Su, Ying Su, Yixuan Su, Pedro Ortiz Suarez, Nishant Subramani, Prasanna Lakkur Subramanyam, Smitha Muthya Sudheendra, Hiroaki Sugiyama, Kazunari Sugiyama, Yoshi Suhara, Xuhui Sui, Elior Sulem, Md Arafat Sultan, Albert Sun, Changzhi Sun, Chengjie Sun, Chenkai Sun, Fei Sun, Guangzhi Sun, Haipeng Sun, Hao Sun, Hao Sun, Hong Sun, Jian Sun, Kaiser Sun, Kun Sun, Le Sun, Ming Sun, Peijie Sun, Qiushi Sun, Renliang Sun, Rui Sun, Shichao Sun, Simeng Sun, Tianxiang Sun, Weiwei Sun, Weixuan Sun, Xiaoyang Sun, Yi Sun, Yutao Sun, Zengkui Sun, Zequn Sun, Zewei Sun, Zhiqing Sun, Zhoujian Sun, Mujeen Sung, Yoo Yeon Sung, Hanna Suominen, Marek Suppa, Abhijit Suresh, Allmin Pradhap Singh Susaiyah, Andrias Susanto, Lintang Sutawika, Benjamin Suter, Mirac Suzgun, Sarathkrishna Swaminathan, Sandesh Swamy, Munira Syed, Stan Szpakowicz, Mario Sänger, Joan Andreu Sánchez, Ricardo Muñoz Sánchez, Víctor M. Sánchez-Cartagena

Santosh T.Y.S.S, Jeniya Tabassum, Oyvind Tafjord, Chang-Yu Tai, Dima Taji, Yu Takagi, Sho Takase, Zeerak Talat, George Tambouratzis, Aleš Tamchyna, Chao-Hong Tan, Haochen Tan, Liling Tan, Minghuan Tan, Qingyu Tan, Xingwei Tan, Ryota Tanaka, Gongbo Tang, Haoyu Tang, Hui Tang, Liyan Tang, Ruixiang Tang, Shuai Tang, Tianyi Tang, Wei Tang, Xiangru Tang, Xuemei Tang, Xunzhu Tang, Yihong Tang, Zheng Tang, Zhiwen Tang, Zineng Tang, Guanhong Tao, Mingxu Tao, Wei Tao, Sandeep Tata, Marta Tatu, Selma Tekir, Serra Sinem Tekiroglu, Eric S. Tellez, Irina Temnikova, Zhiyang Teng, Davide Testa, Alberto Testoni, Martin Teuffenbach, Katherine Thai, Khushboo Thaker, Urmish Thakker, Himanshu Thakur, Surendrabikram Thapa, Avijit Thawani, Anton Frederik Thielmann, Camilo Thorne, Ran Tian, Xuetao Tian, Yijun Tian, Yuanhe Tian, Yufei Tian, Yuhang Tian, Zhiliang Tian, Jörg Tiedemann, Zoran Tiganj, Abhisek Tiwari, Nidhi Tiwari, Prayag Tiwari, Soham Dinesh Tiwari, Hung Quoc To, Amalia Todirascu, Sebastian-Antonio Toma, Nadi Tomeh, Nicholas Tomlin, Cagri Toraman, Shubham Toshniwal, Samia Touileb, Yannick Toussaint, Benjamin Towle, Khanh Quoc Tran, Khiem Vinh Tran, Lucas Vinh Tran, Son Quoc Tran, Thi Hong Hanh Tran, Long Hai Trieu, Bayu Distiawan Trisedya, Harsh Trivedi, Sergey Troshin, Adam Tsakalidis, Dimitrios Tsarapatsanis, Bo-Hsiang Tseng, Ioannis Tsiamas, Eleftheria Tsipidi, Masaaki Tsuchida, Geng Tu, Jingxuan Tu, Yunbin Tu, Yi-Lin Tuan, Marco Turchi

Irina Ualiyeva, Rutuja Ubale, Ana Sabina Uban, Solomon Ubani, Oseremen Oscar Uduehi, Adrian Ulges, Eddie L. Ungless, Apoorva Upadhyaya, Gorka Urbizu, Ashok Urlana, Asahi Ushio, Saiteja Utpala

Jyothir S V, Venktesh V, Saujas Vaduguru, Ashwini Vaidya, Nidhi Vakil, Marco Valentino, Jannis Vamvas, Tim Van de Cruys, Chris Van der Lee, David Vandyke, Natalia Vanetik, Daniel Varab, Vasudha Varadarajan, Francielle Vargas, Neeraj Varshney, Shikhar Vashishth, Siddharth Vashishtha, Aditya Srikanth Veerubhotla, Akshaj Kumar Veldanda, Aswathy Velutharambath, Saranya Venkatraman, Gaurav Verma, Rakesh M Verma, Siddharth Verma, Yannick Versley, Iñaki San Vicente, Jesús Vilares, Dan Vilenchik, Danae Sanchez Villegas, Juraj Vladika, Nikos Voskarides, Ali Vosoughi, Pavlos Vougiouklis, Trang Vu, Yogarshi Vyas, Menno van Zaanen, Pius von Däniken, Spencer McIntosh von der Ohe

Manya Wadhwa, Hiromi Wakaki, David Wan, Hai Wan, Stephen Wan, Xingchen Wan, Yao Wan, Yixin Wan, Yu Wan, Bailin Wang, Baoxin Wang, Baoxun Wang, Benyou Wang, Bethany Yixin Wang, Bin Wang, Bingqing Wang, Boshi Wang, Boxin Wang, Chao Wang, Chao Wang, Chen Wang, Chengyu Wang, Chuan-Ju Wang, Cong Wang, Cunxiang Wang, Fei Wang, Gengyu Wang, Hai Wang, Haobo Wang, Haoyu Wang, Haoyu Wang, Haoyu Wang, Heng Wang, Hongfei Wang, Hongru Wang, Huimin Wang, Jiaan Wang, Jian Wang, Jianing Wang, Jianyu Wang, Jianzong Wang, Jiaqi Wang, Jie Wang, Jin Wang, Jin Wang, Jingjing Wang, Jue Wang, Jun Wang, Jun Wang, Junjie Wang, Ke Wang, Keheng Wang, Lei Wang, Liang Wang, Lijie Wang, Likang Wang, Lingzhi Wang, Longshaokan Wang, Peifeng Wang, Pidong Wang, Ping Wang, Qiang Wang, Qingyun Wang, Qiqi Wang, Rui Wang, Rui Wang, Rui Wang, Ruibo Wang, Saizhuo Wang, Shaobo Wang, Shi Wang, Shoujin Wang, Shuhe Wang, Siyin Wang, Song Wang, Suge Wang, Tao Wang, Tianduo Wang, Tiannan Wang, Wei Wang, Wei Wang, Wei Wang, Weiqi Wang, Weiyue Wang, Wenbo Wang, Wenhui Wang, Wenjin Wang, Wenxuan Wang, Wenya Wang, Xiangdong Wang, Xiao Wang, Xiaolin Wang, Xiaozhi Wang, Xin Wang, Xindi Wang, Xing Wang, Xintong Wang, Xinyu Wang, Xu Wang, Xuan Wang, Xuancong Wang, Xun Wang, Yang Wang, Yanshan Wang, Yaqing Wang, Ye Wang, Ye Wang, Yibo Wang, Yichen Wang, Yikun Wang, Yile Wang, Yimu Wang, Yinggui Wang, Yining Wang, Yiran Wang, Yiwei Wang, Yong Wang, Yu Wang, Yu Wang, Yue Wang, Yueqian Wang, Yun-Cheng Wang, Yuxuan Wang, Zekun Wang, Zengzhi Wang, Zhaowei Wang, Zhen Wang, Zheng Wang, Zhenhailong Wang, Zhichun Wang, Zhiguang Wang, Zhilin Wang, Zhiqiang Wang, Zhongqing Wang, Zhuoer Wang, Zhuoyi Wang, Zifeng Wang, Zihan Wang, Zihan Wang, Zihan Wang, Zihao Wang, Zijian Wang, Zijie Wang, Zilong Wang, Zixiao Wang, Ziyao Wang, Zuhui Wang, Zun Wang, Prashan Wanigasekara, Yusuke Watanabe, William Watson, Bonnie L. Webber, Leon Weber-Genzel, Tharindu Cyril Weerasooriya, Jingxuan Wei, Lingwei Wei, Penghui Wei, Tianxin Wei, Victor Junqiu Wei, Wei Wei, Shira Wein, Leonie Weissweiler, Orion Weller, Simon Wells, Bingbing Wen, Bingyang Wen, Haoyang Wen, Jiaxin Wen, Liang Wen, Yuqiao Wen, Zhihao Wen, Zhihua Wen, Zhiyuan Wen, Rongxiang Weng, Yixuan Weng, Michael Wiegand, John Frederick Wieting, Thilini Wijesiriwardene, Ethan Wilcox, Rodrigo Wilkens, Ronald Wilson, Dawid Wisniewski, Emilia Wiśnios, Raymond K. Wong, Tak-Lam Wong, Alina Wróblewska, Anna Wróblewska, Anne Wu, Ben Peng Wu, Bowen Wu, Cantao Wu, Changxing Wu, Chen Wu, Chen Henry Wu, Chien-Sheng Wu, Di Wu, Di Wu, Fangzhao Wu, Han Wu, Hua Wu, Hui Wu, Jiageng Wu, Junjie Wu, Kelly Ting Wu, Lianwei Wu, Linzhi Wu, Mengxi Wu, Qiyu Wu, Shengqiong Wu, Shih-Hung Wu, Sixing Wu, Stephen Wu, Te-Lin Wu, Ting-Wei Wu, Tingting Wu, Tongtong Wu, Weibin Wu, Xian Wu, Xianchao Wu, Xiaobao Wu, Xin Wu, Xixin Wu, Yang Wu, Yangjun Wu, Yaoyao Wu, Yike Wu, Yimeng Wu, Yuanbin Wu, Yunfang Wu, Yuting Wu, Yuxia Wu, Zeqiu Wu, Zhen Wu, Zhijing Wu, Zhizheng Wu, Zhuofeng Wu, Zihao Wu, Zirui Wu

Jingbo Xia, Menglin Xia, Patrick Xia, Yu Xia, Anhao Xiang, Suncheng Xiang, Wei Xiang, Chaojun Xiao, Chenghao Xiao, Chunyang Xiao, Jinfeng Xiao, Jing Xiao, Min Xiao, Zhaomin Xiao, Zilin Xiao, Chenhao Xie, Jian Xie, Kaige Xie, Shangyu Xie, Tong Xie, Yaqi Xie, Yiqing Xie, Yubo Xie, Yuqing Xie, Zhiwen Xie, Zhongbin Xie, Zhouhang Xie, Zhuohan Xie, Xin Xin, Ying-

wei Xin, Bowen Xing, Chen Xing, Linzi Xing, Bo Xiong, Chao Xiong, Haoyu Xiong, Bokai Xu, Boyan Xu, Canwen Xu, Chen Xu, Chenchen Xu, Chunpu Xu, Dongfang Xu, Dongkuan Xu, Fan Xu, Fangyuan Xu, Guandong Xu, Guangyue Xu, Haiyang Xu, Hanzi Xu, Hongfei Xu, Hongyan Xu, Hongzhi Xu, Jialiang Xu, Jiannan Xu, Jiashu Xu, Jin Xu, Jinan Xu, Jitao Xu, Jun Xu, Kai Xu, Kang Xu, Keyang Xu, Kun Xu, Lei Xu, Liyan Xu, Lu Xu, Lvxiaowei Xu, Peng Xu, Qiongkai Xu, Shanshan Xu, Shicheng Xu, Wang Xu, Wanshi Xu, Weiwen Xu, Wenduan Xu, Wenjie Xu, Xianghong Xu, Xiao Xu, Xiaohao Xu, Xinnuo Xu, Yan Xu, Yi Xu, Yige Xu, Yiheng Xu, Zhen Xu, Zhenhui Xu, Zhiyang Xu, Zihang Xu, Fuzhao Xue, Huiyin Xue, Yun Xue

Mohit Yadav, Aditya Yadavalli, Ikuya Yamada, An Yan, Brian Yan, Chenwei Yan, Hanqi Yan, Jianhao Yan, Jun Yan, Ming Yan, Tianwei Yan, Weixiang Yan, Yang Yan, Zhaohui Yan, An Yang, Baosong Yang, Cheng Yang, Chengfu Yang, Chenyang Yang, Dejie Yang, Deqing Yang, Eugene Yang, Fan Yang, Guang Yang, Guanqun Yang, Guo Yang, Haiqin Yang, Hao Yang, Haoran Yang, Heng Yang, Jianing Yang, Jiaoyun Yang, Jun Cheng Yang, Jun Yang, Jun Yang, Kexin Yang, Liner Yang, Linyi Yang, Longfei Yang, Mingming Yang, Muqiao Yang, Muyun Yang, Nan Yang, Peng Yang, Ruichao Yang, Sen Yang, Sen Yang, Shiquan Yang, Shuai Yang, Songlin Yang, Tao Yang, Tianchi Yang, Tsung-Yen Yang, Wei Yang, Wenmian Yang, Xianjun Yang, Yahan Yang, Yilin Yang, Yizhe Yang, Yue Yang, Yujiu Yang, Zachary Yang, Zhao Yang, Zhichao Yang, Zi Yang, Zonglin Yang, Ken Yano, Barry Menglong Yao, Bingsheng Yao, Jianzhu Yao, Liang Yao, Peiran Yao, Shunyu Yao, Wenlin Yao, Yitong Yao, Zijun Yao, Bingyang Ye, Dezhi Ye, Fanghua Ye, Hai Ye, Jiasheng Ye, Jinhui Ye, Junjie Ye, Muchao Ye, Qinyuan Ye, Seonghyeon Ye, Wei Ye, Wenting Ye, Xi Ye, Akhila Yerukola, Yu Ting Yeung, Jingwei Yi, Li S. Yifei, Wen-wai Yim, Seid Muhie Yimam, Da Yin, Fan Yin, Ming Yin, Qingyu Yin, Wenbiao Yin, Wenjie Yin, Xuwang Yin, Yunting Yin, Yuwei Yin, Michael Miller Yoder, Hikaru Yokono, KiYoon Yoo, Eunseop Yoon, Hee Suk Yoon, Seunghyun Yoon, Sunjae Yoon, Susik Yoon, Issei Yoshida, Masaharu Yoshioka, Chenyu You, Haoxuan You, Weiqiu You, Yiling You, Steve Young, Paul Youssef, Bowen Yu, Changlong Yu, Cheng Yu, Dian Yu, Dong Yu, Erxin Yu, Guoxin Yu, Hang Yu, Heng Yu, Jifan Yu, Kai Yu, Le Yu, Lei Yu, Liang-Chih Yu, Mengxia Yu, Ping Yu, Shuai Yu, Simon Chi Lok Yu, Tiezheng Yu, Tong Yu, Wenhao Yu, Xiao Yu, Xiaodong Yu, Xinchen Yu, Xinyan Velocity Yu, Yue Yu, Zac Yu, Zhouliang Yu, Changsen Yuan, Chunyuan Yuan, Fangfang Yuan, Fei Yuan, Li Yuan, Lifan Yuan, Quan Yuan, Siyu Yuan, Xiaosong Yuan, Zhaoquan Yuan, Zheng Yuan, Linan Yue, Xiang Yue, Hyokun Yun

Polina Zablotskaia, Ofir Zafrir, Hamada M Zahera, Mahdi Zakizadeh, Olga Zamaraeva, Daoguang Zan, Yuan Zang, Fabio Massimo Zanzotto, Klim Zaporojets, Alessandra Zarcone, Rabih Zbib, Albin Zehe, Eric Zelikman, Yury Zemlyanskiy, Daojian Zeng, Fanhu Zeng, Guangtao Zeng, Jayden Zeng, Jiali Zeng, Qi Zeng, Qingkai Zeng, Weixin Zeng, Xianfeng Zeng, Xingshan Zeng, Yan Zeng, Yawen Zeng, Yutao Zeng, Ziqian Zeng, Deniz Zeyrek, Haolan Zhan, Hongli Zhan, Qiusi Zhan, Runzhe Zhan, Arthur Jun Zhang, Baohua Zhang, Beichen Zhang, Biao Zhang, Bowen Zhang, Bowen Zhang, Boyang Zhang, Chen Zhang, Chen Zhang, Chong Zhang, Chuheng Zhang, Dong Zhang, Dong Zhang, Dongyu Zhang, Duzhen Zhang, Fuzheng Zhang, Hainan Zhang, Hao Zhang, Hao Zhang, Hao Zhang, Haopeng Zhang, Haoran Ranran Zhang, Haotong Zhang, Hongming Zhang, Hongyu Zhang, Jianguo Zhang, Jingqing Zhang, Jipeng Zhang, Jiyang Zhang, Junwen Zhang, Kai Zhang, Kai Zhang, Kaiyan Zhang, Ke Zhang, Kechi Zhang, Kun Zhang, Le Zhang, Lei Zhang, Li Zhang, Licheng Zhang, Lingyu Zhang, Lining Zhang, Longhui Zhang, Meiru Zhang, Miaoran Zhang, Michael JQ Zhang, Mike Zhang, Muru Zhang, Nan Zhang, Qi Zhang, Qing Zhang, Ruochen Zhang, Ruohong Zhang, Shaokang Zhang, Shaolei Zhang, Shiyue Zhang, Shuai Zhang, Shujian Zhang, Shuo Zhang, Song Zhang, Songming Zhang, Tao Zhang, Tianhang Zhang, Tianlin Zhang, Tianshu Zhang, Tong Zhang, Tongtao Zhang, Wei Emma Zhang, Weixu Zhang, Wen Zhang, Wenqi Zhang, Wenqiang Zhang, Wenxuan Zhang, Wenzheng Zhang, Xiao Zhang, Xiaokun Zhang, Xiaoqiang Zhang, Xiaotong Zhang, Xin Zhang, Xin Zhang, Xinghua Zhang, Xinliang Frederick Zhang, Xinsong Zhang, Xulang Zhang, Xulong Zhang, Yan Zhang, Yan Zhang, Yangjun Zhang, Yanzhe Zhang, Yao Zhang, Yi Zhang, Yian Zhang, Yichi Zhang, Yiming Zhang,

Yin Zhang, Yizhou Zhang, Yong Zhang, Yu Zhang, Yu Zhang, Yu Zhang, Yuan Zhang, Yuanzhe Zhang, Yue Zhang, Yuhan Zhang, Yuhao Zhang, Yuhui Zhang, Yuji Zhang, Yun Zhang, Yunxiang Zhang, Yunyi Zhang, Yuqi Zhang, Yusen Zhang, Yuxiang Zhang, Zequn Zhang, Zeyu Zhang, Zhe Zhang, Zhehao Zhang, Zhengyan Zhang, Zhexin Zhang, Zhihan Zhang, Zhirui Zhang, Zhisong Zhang, Zhuo Zhang, Zhuosheng Zhang, Bing Zhao, Chao Zhao, Chenye Zhao, Fei Zhao, Guangxiang Zhao, Hai Zhao, Jeffrey Zhao, Jiahao Zhao, Jianyu Zhao, Jinming Zhao, Junchen Zhao, Kai Zhao, Kaiqi Zhao, Mengjie Zhao, Qinghua Zhao, Sanqiang Zhao, Shu Zhao, Shuai Zhao, Shuai Zhao, Tiancheng Zhao, Tianyu Zhao, Tony Z. Zhao, Wenting Zhao, Xiaoyan Zhao, Xinran Zhao, Xinyan Zhao, Xuandong Zhao, Yang Zhao, Yang Zhao, Yangyang Zhao, Yilun Zhao, Yu Zhao, Yu Zhao, Zhengyi Zhao, Zhixue Zhao, Boyuan Zheng, Ce Zheng, Changmeng Zheng, Junhao Zheng, Kai Zheng, Renjie Zheng, Rui Zheng, Shen Zheng, Siyan Zheng, Xianrui Zheng, Xiaoqing Zheng, Xinyi Zheng, Yaowei Zheng, Yinhe Zheng, Yujia Zheng, Zaixiang Zheng, Zi'ou Zheng, Zihao Zheng, Zilong Zheng, Ming Zhong, Ruiqi Zhong, Wei Zhong, Xian Zhong, Yang Zhong, Zexuan Zhong, Baohang Zhou, Ben Zhou, Bo Zhou, Dong Zhou, Guangyou Zhou, Han Zhou, Hanzhang Zhou, Houquan Zhou, Jiawei Zhou, Jiawei Zhou, Jie Zhou, Jinfeng Zhou, Jingbo Zhou, Jingyan Zhou, Junsheng Zhou, Kaitlyn Zhou, Kankan Zhou, Kun Zhou, Kyrie Zhixuan Zhou, Li Zhou, Long Zhou, Nina Zhou, Pei Zhou, Peilin Zhou, Qingyu Zhou, Shilin Zhou, Shuchang Zhou, Shuyan Zhou, Tong Zhou, Wangchunshu Zhou, Wenjie Zhou, Wenxuan Zhou, Xiabing Zhou, Xiang Zhou, Xin Zhou, Xixi Zhou, Yangqiaoyu Zhou, Yi Zhou, Yi Zhou, Yichao Zhou, Yichu Zhou, Yilun Zhou, Yu Zhou, Yucheng Zhou, Yufan Zhou, Yuhao Zhou, Yunhua Zhou, Yuxiang Zhou, Anjie Zhu, Bolin Zhu, Dawei Zhu, Fangqi Zhu, Hao Zhu, Linchao Zhu, Luyao Zhu, Peide Zhu, Qi Zhu, Rongxin Zhu, Rui-Jie Zhu, Su Zhu, Suyang Zhu, Tong Zhu, Wang Zhu, Wanzheng Zhu, Wei Zhu, Wenhao Zhu, Xi Zhu, Xiaofeng Zhu, Xuan Zhu, Xuekai Zhu, Yi Zhu, Yilun Zhu, Yongxin Zhu, Yutao Zhu, Zhihong Zhu, Zining Zhu, Honglei Zhuang, Yuan Zhuang, Yuchen Zhuang, Heike Zinsmeister, Ayah Zirikly, Yftah Ziser, Marco Zocca, Shi Zong, Bowei Zou, Hao Zou, Henry Peng Zou, Heqing Zou, Yicheng Zou, Amal Zouaq, Vilém Zouhar, Xinyu Zuo

Emily Öhman, Lilja Øvrelid, Tolúlopé Ògúnrèmí

Michal Štefánik

2

# Anti-Harassment Policy

EMNLP 2023 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 person's 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/adminwiki/index.php/Professional_Conduct_Committee*

The full policy and its implementation is defined at:

*https://www.aclweb.org/adminwiki/index.php?title=Anti-Harassment_Policy*

3

# Meal Info

**Breakfast**:
Will not be served at the Conference.

***Break**:
Coffee, tea, and light snacks will be provided late morning (approximately 10:30) and midafternoon (approximately 15:30)

**Lunch and Dinner**:
Lunch and dinner are not provided, but there are food trucks, cafes, restaurants and shops within walking distance. You can pick up a list of options at registration.

****Welcome Reception**:
Light hors d'oeuvres will be provided on Thursday Evening Dec 07, 2023, at the Welcome Reception which will be held in the Central Ballroom of the Resorts World at Sentosa Convention Center Level B2.
Welcome Reception tickets are included as part of the Full Conference Registration and can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk.
No admission without an entry ticket.

****Social Event**:
Dinner will only be provided on Saturday Evening December 9, 2023, at the Social Event which will be held in the Central Ballroom of the Resorts World at Sentosa Convention Center Level B2.
Social Event tickets are included as part of Entire Conference Registration and can be added on for Guests, Tutorial, Workshop and Exhibitors to attend at the Registration Solutions Desk.
No admission without an entry ticket.

Please note the following:
*Denotes Workshop/Tutorial/Main Conference Days
**Denotes Full Conference Days

# 4 Welcome Reception

## Welcome Reception - Thursday, Dec. 7, 2023

Join your fellow delegates for a relaxing night as we start things off on a high note with a Lion Dance performance to bring in good vibes. The reception will be spiced up with a Multi-Ethnic Dance, showcasing the Malay Dance, Chinese Fan ribbon Dance as well as Indian Semi-classical Dance. Don't miss out – come be part of the experience.

**Venue**: Resorts World at Sentosa Convention Center
**Time**: 19.00 - 21.30
**Lion Dance**: 19:00 Lion Dance, 20:00 Entertainment
**Dress code**: Smart Casual

One entry ticket will be included with each full conference registration. To get admission into the event you will need to have your name badge on your person as the QR code that is located on your badge is how the ACL Staff member(s) Scan and account for admission(s). No name badge, no entrance. Social Event tickets can be added on for Guests, Tutorials, Workshops, and Exhibitors to attend at the Registration Solutions Desk or through your Yes Events registration login.

**Location**: East & Central Ballrooms
**Directions**: From the hotel lobby head to the Convention Center. Take the escalator down to Level (B2)

**Light Hors d'oeuvres & Cash Bar**: Each Full Conference attendee will receive 1 complimentary drink ticket upon admission into the Welcome Reception

**Lion Dance**: The lion dance is a pugilistic performance dating back to more than 1,500 years. Its performance during auspicious occasions, such as the launch of new businesses and shops, is believed to bring good fortune and wealth.

# 5 Social Events

## Social Event - Saturday, Dec. 9, 2023

**Venue**: Resorts World at Sentosa Convention Center & Universal Studios Singapore

**Time**:

- 18:30 - 21:30: Buffet Dinner with a complimentary drink ticket upon admission
- 20:00 - 23:45: Social Activity Exclusive Access to EMNLP

**Dinner Details**:

- One entry ticket included with each full conference registration.
- Admission requires wearing your name badge with a QR code for scanning by ACL Staff.
- No name badge, no entrance.

**Universal Studios Access**:

- Wristbands will be administered at the conference.
- Access to Universal Studios requires the assigned wristband.
- No band, no access.

**Social Events Dinner Location**:

- West Ballroom 1 – 3 located on Level B2.

**Directions**:

- From the hotel lobby, head to the Convention Center.

- Take the escalator down to Level (B2).

**Map**:

- A map from Dinner to Universal Studios will be displayed on the TV Monitors.

Please ensure you have your name badge and wristband for seamless access to the events. Enjoy the evening!

**Universal Studios Pass** get you access from 20:00 - 23:45 to the following Exclusive usage of attraction rides in operation:

- Battlestar Galactica: HUMAN
- Transformers
- Revenge of the Mummy
- Jurassic Park Rapids Adventure
- Dino-Soarin
- Sesame Street Spaghetti Space Chase
- Puss in Boots' Giant Journey
- Enchanted Airways

**F&B outlets** in operation:

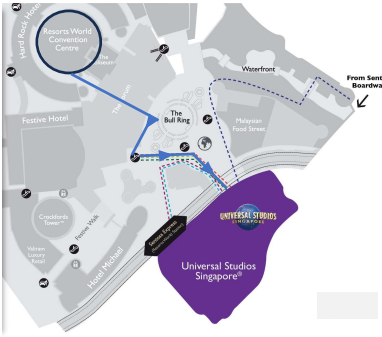
- Food kiosks: Star Snacks, Planet Yen, Cairo Market,
- USS Food Carts in operation: Phaorah, Frozen, Jungle, Galactica Treat

**Social Event Tickets**:

- Are for all Full Conference Registered Attendees.
- Guests, Tutorial, Workshop, and Exhibitors are welcome to add on the Social Event Tickets at the Registration Solutions Desk or through the Yes Events registration login.

## DIRECTIONS

Resorts World Convention Centre and Universal Studios Singapore




**Resorts World Convention Centre → Universal Studios Singapore**

- Exit Resorts World Convention Centre via B2 VIP Entrance
- Follow the Red Path to the escalator leading to USS
- Take the escalator up to L1
- Follow signage and walk towards Universal Studios Singapore

PRIVATE & CONFIDENTIAL
All information herewith is privileged/ confidential and subjected to reviews , and should not be disclosed out of the intended distribution list.

# 6 Keynotes

# Human-Centric Natural Language Processing

**Jong Park**
Korea Advanced Institute of Science and Technology (KAIST)

**Friday, Dec 08, 2023** - Room: **East & Central** - Time: **9:30-10:30**

**Abstract:** In keeping up with the three pillars of natural language processing, ever-changing principles and techniques, emerging domains of application and related resources, and people with needs for language support, our research team has paid attention to the third pillar with a special focus on human-centric diversity and minority issues, including children, experts in biology and medicine, deaf people, and people with different challenges to language use. In this talk, I wish to share our achievements along the way, recently leading to a flurry of pleasant results that make use of deep learning techniques and large language models. I conclude the talk with a cautious prognosis about what might lie beyond large language models that loom over much of what we do at the moment and may hamper diversity to a worrisome degree and even humanity in hindsight.

**Bio:** Jong Park received his BE and MSE degrees from Seoul National University and PhD degree from the University of Pennsylvania, Philadelphia. He has been working as Assistant, Associate and Full Professor at Korea Advanced Institute of Science and Technology (KAIST) since 1998. He is one of the early researchers on BioNLP, applying NLP techniques to biology and medicine. His research team at KAIST has also been working broadly on identifying emotion from text, turning spoken language into visual animation and sign language, identifying mental health issues such as mild-cognitive impairment (MCI) and clinical depression from natural language utterances, detecting abusive language, and, more recently, credibility assessment and bidirectional sign language processing. His team has received Outstanding Paper Award at ACL 2023 for the work on the Tigrinya language. He is serving as founding Editor-in-Chief of Journal of Computing Science and Engineering (JCSE) since 2007, President of Asian Federation of Natural Language Processing (AFNLP) during 2022–2024, and General Chair of IJCNLP-AACL 2023.

# From Speech to Emotion to Mood: Mental Health Modeling in Real-World Environments

**Emily Mower Provost**
University of Michigan, USA

**Saturday, Dec 09, 2023** - Room: **East & Central** - Time: **14:30-15:30**

**Abstract:** Emotions provide critical cues into our health and wellbeing. This is of particular importance in the context of mental health, where changes in emotion may signify changes in symptom severity. However, information about emotion and how it varies over time is often accessible only using survey methodology (e.g., ecological momentary assessment, EMA), which can become burdensome to participants over time. Automated speech emotion recognition systems could provide an alternative, providing quantitative measures of emotion using acoustic data captured passively from a consented individual's environment. However, speech emotion recognition systems often falter when presented with data collected from unconstrained natural environments due to issues with robustness, generalizability, and invalid assumptions. In this talk, I will discuss our journey in speech-centric mental health modeling, explaining whether, how, and when emotion recognition can be applied to natural unconstrained speech data to measure changes in mental health symptom severity.

**Bio:** Emily Mower Provost is a Professor in Computer Science and Engineering at the University of Michigan. She received her Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2010. She is a Toyota Faculty Scholar (2020) and has been awarded a National Science Foundation CAREER Award (2017), the Oscar Stern Award for Depression Research (2015), a National Science Foundation Graduate Research Fellowship (2004-2007). She is an Associate Editor for IEEE Transactions on Affective Computing and the IEEE Open Journal of Signal Processing. She has also served as Associate Editor for Computer Speech and Language and ACM Transactions on Multimedia. She has received best paper awards or finalist nominations for Interspeech 2008, ACM Multimedia 2014, ICMI 2016, and IEEE Transactions on Affective Computing. Among other organizational duties, she has been Program Chair for ACII (2017, 2021), ICMI (2016, 2018). Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of the perception and expression of human behavior.

# Academic NLP research in the Age of LLMs: Nothing but blue skies!

**Christopher Manning**
Stanford University, USA

**Sunday, Dec 10, 2023** - Room: **East & Central** - Time: **14:00-15:00**

**Abstract:** There has been a certain amount of handwringing by students – and faculty –about the prospects for academic research in the Age of Large Language Models (LLMs). Now, there is a need for universities and governments to do more to support computational research. And it is indeed the case that any major inflection point in a research field does change the best questions to concentrate your time on. But a reorientation of research is healthy, coming off a decade when students in academia had often drifted into playing the Kaggle game rather than thinking hard about fundamental ideas. A time of transformation is principally a time of opportunity: All sorts of new and exciting research questions, either not explored or only explored in very different contexts, come to the fore. These questions provide compelling opportunities for fresh and exciting work. I see many examples of this sort of work already happening and encourage many more. I will illustrate with two pieces of work from my own students: the Backpack, which explores an alternative, more interpretable architecture than the Transformer (by John Hewitt and colleagues) and Direct Preference Optimization, which explores an alternative to the Proximal Policy Optimization normally used for steering LLMs in the Reinforcement Learning from Human Feedback (RLHF) phase (by Eric Mitchell and colleagues).

**Bio:** Christopher Manning is the inaugural Thomas M. Siebel Professor in Machine Learning in the Departments of Linguistics and Computer Science at Stanford University, Director of the Stanford Artificial Intelligence Laboratory (SAIL), and an Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). His research goal is computers that can intelligently process, understand, and generate human languages. Manning was an early leader in applying Deep Learning to Natural Language Processing (NLP), with well-known research on the GloVe model of word vectors, attention, machine translation, question answering, self-supervised model pre-training, tree-recursive neural networks, machine reasoning, dependency parsing, sentiment analysis, and summarization. He also focuses on computational linguistic approaches to parsing, natural language inference and multilingual language processing, including being a principal developer of Stanford Dependencies and Universal Dependencies. Manning has coauthored leading textbooks on statistical approaches to NLP (Manning and Schütze 1999) and information retrieval (Manning, Raghavan, and Schütze, 2008), as well as linguistic monographs on ergativity

and complex predicates. His online CS224N Natural Language Processing with Deep Learning videos have been watched by hundreds of thousands of people. He is an ACM Fellow, a AAAI Fellow, and an ACL Fellow, and a Past President of the ACL (2015). His research has won ACL, Coling, EMNLP, and CHI Best Paper Awards, and an ACL Test of Time Award. He has a B.A. (Hons) from The Australian National University and a Ph.D. from Stanford in 1994, and an Honorary Doctorate from U. Amsterdam in 2023, and he held faculty positions at Carnegie Mellon University and the University of Sydney before returning to Stanford. He is the founder of the Stanford NLP group (@stanfordnlp) and manages development of the Stanford CoreNLP and Stanza software.

7

**Panel**

# Beyond Text: Inclusive Human Communication with Language Technology

**Time:** Dec. 9, 2023 - 16:00-17:00 **Location:** East & Central

**Theme:**

The scope of communication extends far beyond textual interactions. This panel converges experts from diverse fields, including sign language, multilingual conference interpretation, empathetic speech-based dialog systems, virtual reality, and embodied agents, who will discuss how advancements in language technology can help shape the future of inclusive communication, and at the same time what NLP can learn from various fields of communication.

**Panelists:**

**Ms. Lourdes de Rioja**, Freelance Conference Interpreter, Belgium
Lourdes de Rioja is a Spanish conference interpreter who has worked for the European Commission, the European Parliament, and the European Court of Justice. She is the author and producer of SCICtrain for the European Commission, editor and designer of False Friends online Dictionary, and co-creator of AIIC Conversations. Ms de Rioja has a Master's degree in Political & Corporate Communication from The George Washington University & Universidad de Navarra, Spain, and a European Master's degree in Conference Interpreting from Universidad de La Laguna, Spain. She is a member of the International Association of Conference Interpreters (AIIC) and has interpreted in various languages, including English, French, Danish, Swedish, and Catalan.

**Dr. Abraham Glasser**, Gallaudet University, USA
Dr. Abraham Glasser is an Assistant Professor in Science, Technology, Accessibility, Mathematics, and Public Health. He earned his undergraduate degree and his Ph.D. in Computing and Information Sciences from the Rochester Institute of Technology. Glasser has conducted research work for various organiza-

tions. His research interests include virtual and augmented reality, accessible technology for deaf and hard of hearing individuals – with a focus on automatic speech recognition, and he is passionate about making technology inclusive for people with disabilities. Dr. Glasser is also the reigning United States Deaf Chess Champion.

**Prof. Chengkuo Lee**, National University, Singapore
Prof Chengkuo Lee is an Associate Professor at the Dept. of Electrical and Computer Eng. of National University of Singapore, Singapore with a research background in Systems and Precision Engineering. He received his M.S. degree in Materials Science and Engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1991, and his Ph.D. degree in Precision Engineering from the University of Tokyo, Tokyo, Japan, in Jan. 1996. He has proposed augmented tactile-perception and haptic-feedback rings with multimodal sensing and feedback capabilities for Human-Machine interfaces for immersive interactions. He has also proposed a sign language recognition and communication system using a smart triboelectric glove, AI block, and the back-end VR interface. Prof. Lee has contributed more than 250 papers in peer-reviewed international journals and conferences, and 9 US patents in MEMS, NEMS, Nanophotonics and Nanotechnology fields.

**Prof. María Inés Torres**, UPV/EHU, Spain
María Inés Torres is a Professor of Computer Science at the University of the Basque Country. She founded the Pattern Recognition and Speech Technology research group in 1990, which she has been leading since then. She has conducted research related to speech technologies, including automatic speech recognition and understanding, language identification, machine translation, specific processing of Basque language, and acquisition and generation of language resources. She has published numerous papers in journals and international conferences and edited three books. Her current research interests focus on statistical approaches to deal with spoken dialog systems, being also interested in learning from human interaction to generate artificial interaction.

**Prof. Monojit Choudhury**, MBZUAI, Abu Dhabi
Monojit Choudhury is a Professor of NLP at MBZUAI, Abu Dhabi. He has a PhD and BTech in computer science and engineering from IIT Kharagpur. Prof Choudhury's research interests span various aspects of natural language processing and cognitive sciences, including multilingual models, computational ethics, and AI and NLP for social good. Prior to joining MBZUAI, he was a principal researcher at Microsoft. He also serves as a professor of practice at Plaksha University and an adjunct faculty at IIIT Hyderabad. Prof Choudhury is the general chair of the Panini Linguistics Olympiad, which is the Indian national linguistics Olympiad for high school students and the founding co-chair of the Asia Pacific Linguistics Olympiad.

# 8 Birds-of-a-Feather and Affinity Group Meetup

At EMNLP 2023, we continue the tradition of conducting birds-of-a-feather (or BoF) and Affinity Group Meeting sessions to help newbie researchers get in touch with other people working in the same areas. Here's a quick primer for people who haven't attended such sessions before!

**What do we hope to accomplish with BoF and Affinity Group Meeting sessions?**

- We want to provide junior researchers/people attending a CL conference for the first time with a platform to discuss ideas and issues with other researchers in their areas of interest. This can be a good avenue for junior researchers to get feedback on ongoing ideas, learn about relevant ongoing projects at other groups/universities, and develop a broader understanding of their field of interest.
- We want to facilitate more exchange of research ideas and enable collaborative discussions.

**How are BoF and Affinity Group Meeting sessions structured?**

- All BoF and Affinity Group Meeting sessions are in-person only.
- Each meetup will be led by the session chair(s).
- Anyone who wants to attend can join; prior signup is unnecessary.

**When are the BoF and Affinity Group meeting sessions happening?**

**Data: Dec. 08, 2023**

- **BOF-1: The role of NLP researchers in shaping and supporting a society with ubiquitous AI use**
  - Time: 14:00-15:30
  - Room: Aquarius 1

  - Session Chair(s): Maria Liakata

- **BOF-2: NLP on legal texts**
  - Time: 14:00 - 15:30
  - Room: Aquarius 2
  - Session Chair(s): Santosh Tokala
- **BOF-3: Open multilingual text collections under the umbrella of the EU Horizon HPLT project (`https://hplt-project.org/`)**
  - Time: 16:00 - 17:30
  - Room: Aquarius 1
  - Session Chair(s): Andrey Kutuzov
- **BOF-5: NLP for Climate Change**
  - Time: 16:00 - 17:30
  - Room: Aquarius 2
  - Session Chair(s): Jingwei Ni
- **AGM-1: Queer in AI social at EMNLP**
  - Time: 17:30 - 19:30
  - Room: Aquarius 4
  - Session Chair(s): Pranav A

**Data: Dec. 09, 2023**

- **BOF-4: EthioNLP**
  - Time: 9:00 - 10:30
  - Room: Aquarius 1
  - Session Chair(s): Teshome Ababu
- **BOF-6: Embeddings**
  - Time: 11:00 - 12:30
  - Room: Aquarius 1
  - Session Chair(s): Han Xiao
- **BOF-7: Recipes in Building Language Reasoners**
  - Time: 11:00 - 12:30
  - Room: Aquarius 2
  - Session Chair(s): Yiyuan Li and Wenting Zhao
- **BOF-8: ACL Mentorship: Sharing learnings from this conference & identifying promising research directions**

  - Time: 13:45 - 14:30
  - Room: Aquarius 1
  - Session Chair(s): Mentors from ACL Mentorship: Zhijing Jin, Rada Mihalcea, Mohit Bansal, Yuntian Deng, and Others

- **BOF-9: Controlled/Constrained Text Generation**
  - Time: 13:45 - 14:30
  - Room: Aquarius 2
  - Session Chair(s): Allen Roush

- **BOF-10: Arabic NLP**
  - Time: 13:45 - 14:30
  - Room: Aquarius 4
  - Session Chair(s): Khalil Mrini and Rawan N. Almatham

**Data: Dec. 10, 2023**

- **BOF-11: Discourse & Pragmatics**
  - Time: 9:00 - 10:30
  - Room: Aquarius 1
  - Session Chair(s): Valentina Pyatkin, Elias Stengel-Eskin, and Janet Liu

- **AGM-2: Christians@NLP**
  - Time: 9:00 - 10:30
  - Room: Aquarius 2
  - Session Chair(s): Philipp Heinisch

- **BOF-12: Text Simplification and Readability**
  - Time: 11:00 - 12:30
  - Room: Aquarius 1
  - Session Chair(s): Fernando Alva Manchego

- **BOF-13: Automatic Text Evaluation**
  - Time: 11:00 - 12:30
  - Room: Aquarius 2
  - Session Chair(s): Gisela Vallejo and Zhuohan Xie

Attendees are free to carry forward the discussion after the BoF and Affinity Group Meeting sessions end (or even continue in the same room if it is available!).

If you have any questions about the information listed in the schedule, please contact the D&I chairs via emnlp2023diversity@googlegroups.com.

**Additional Guidelines for Attendees**

- Your session chair(s) will invite you to introduce yourself briefly.
- The session format can vary based on the preferences of the chair(s), for example, it may be structured as an Ask Me Anything, a slide deck roundtable, a roundtable Q&A, etc. Your session chair(s) will explain the format at the beginning of the session.
- Please follow the participation guidelines set by the session chair(s) (e.g., raising your hand to ask questions, etc.)
- Please do not record the session.
- Please abide by the ACL Anti-Harassment Policy at all times.

We hope you enjoy these meetups and get a chance to make new friends who share your research interests!

# 9

# Tutorials: Wednesday, December 6, 2023

## Overview

| | | |
|---|---|---|
| 07:30 - 17:00 | ***Registration*** | |
| 09:00 - 10:30 | ***Morning Tutorials – Session 1*** | |
| | *T1: NLP+Vis: NLP Meets Visualization*<br>Shafiq Joty, Enamul Hoque, Jesse Vig | *Pisces 1* |
| | *T2: Security Challenges in Natural Language Processing Models*<br>Qiongkai Xu, Xuanli He | *Pisces 2 & 3* |
| | *T3: Designing, Evaluating, and Learning from Humans Interacting with NLP Models*<br>Tongshuang Wu, Diyi Yang, Sebastin Santy | *Leo 3 & 4* |
| 10:30 - 11:00 | ***Coffee Break*** | |
| 11:00 - 12:30 | ***Morning Tutorials – Session 2*** | |
| | *T1: NLP+Vis: NLP Meets Visualization*<br>Shafiq Joty, Enamul Hoque, Jesse Vig | *Pisces 1* |
| | *T2: Security Challenges in Natural Language Processing Models*<br>Qiongkai Xu, Xuanli He | *Pisces 2 & 3* |
| | *T3: Designing, Evaluating, and Learning from Humans Interacting with NLP Models*<br>Tongshuang Wu, Diyi Yang, Sebastin Santy | *Leo 3 & 4* |
| 12:30 - 14:00 | ***Lunch*** | |
| 14:00 - 15:30 | ***Afternoon Tutorials – Session 1*** | |
| | *T4: LLM-driven Instruction Following: Progresses and Concerns*<br>Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, Hinrich Schütze | *Leo 3 & 4* |
| | *T5: Mitigating Societal Harms in Large Language Models* | *Pisces 1* |

| | | |
|---|---|---|
| | Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, Yulia Tsvetkov | |
| | *T6: Creative Natural Language Generation*<br>Tuhin Chakrabarty, Vishakh Padmakumar,He He, Nanyun Peng | *Pisces 2 & 3* |
| 15:30 - 16:00 | ***Coffee Break*** | |
| 16:00 - 17:30 | ***Afternoon Tutorials – Session 2*** | |
| | *T4: LLM-driven Instruction Following: Progresses and Concerns*<br>Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, Hinrich Schütze | *Leo 3 & 4* |
| | *T5: Mitigating Societal Harms in Large Language Models*<br>Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, Yulia Tsvetkov | *Pisces 1* |
| | *T6: Creative Natural Language Generation*<br>Tuhin Chakrabarty, Vishakh Padmakumar,He He, Nanyun Peng | *Pisces 2 & 3* |

# Message from the Tutorial Chairs

Welcome to the Tutorials Session of EMNLP 2023.

The EMNLP tutorials session is organized to give conference attendees a comprehensive introduction by expert researchers to a variety of topics of importance drawn from our rapidly growing and changing research field.

This year, as has been the tradition over the past few years, the call, submission, reviewing and selection of tutorials were coordinated jointly for multiple conferences: EACL, ACL, and EMNLP. The committee followed a reviewing process that ensured that each of the 42 tutorial submissions received at least two reviews. The selection criteria included clarity, preparedness, novelty, timeliness, instructors' experience, likely audience, open access to the teaching materials, diversity (multilingualism, gender, age and geolocation) and the compatibility of preferred venues. A total of six tutorials were selected for EMNLP.

We would like to thank the tutorial authors for their contributions and flexibility while organising the conference in a hybrid format. Finally, we would like to thank the conference organizers for effective collaboration, and in particular to the general chair Yuji Matsumoto.

We hope you enjoy the tutorials.

EMNLP 2023 Tutorial Co-chairs
Hassan Sajjad
Qi Zhang

# T1 - NLP+Vis: NLP Meets Visualization

Shafiq Joty, Enamul Hoque and Jesse Vig
Cutting-edge
Wednesday, Dec 06, 2023 - 09:00-12:30 (Pisces 1)

https://nlp4vis.github.io/

Natural language and visualization (Vis) are two powerful modalities of human communication. The goal of this tutorial is to push forward the agenda of tightly integrating these two modalities. To this end, the tutorial will introduce NLP+Vis with a focus on two main threads of work: 1) *NLP for Vis:* How to develop and adapt state-of-the-art NLP models for solving various visualization tasks? and 2) *Vis for NLP*: How to leverage visualization techniques to interpret and explain complex NLP models effectively?

The tutorial will first motivate why NLP+Vis is an important area of research and provide an overview of research topics on combining NLP and Vis techniques. Then an overview of state-of-the-art deep learning models for NLP will be covered. Next, we will provide an overview of applying visualization techniques to help make NLP models more interpretable and explainable. In the final part, we will focus on various application tasks at the intersection of NLP and Vis. We will conclude with an interactive discussion of future challenges for NLP+Vis applications. The audience will include researchers interested in applying NLP for visualizations as well as others who focus more generally at the intersection of machine learning and visualization.

---

**Shafiq Joty**, Salesforce Research, USA
email: sjoty@salesforce.com
website: https://raihanjoty.github.io/
Bio. Shafiq Joty is a Research Director at Salesforce Research, and is also an Associate Professor (on leave) at NTU, Singapore. His work has primarily focused on developing language analysis tools and NLP applications. A significant part of his current research focuses on multilingual (machine translation, cross-lingual transfer), multimodal (visual-language learning, NLP+Vis, Code+NLP) NLP, interpretability and robustness of NLP models. His research contributed to 17 patents and more than 110 papers in top-tier NLP and ML conferences and journals including ACL, EMNLP, NAACL, NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CL and JAIR. Shafiq served (or will serve) as a PC chair of SIGDIAL'23, an S/AC for ICLR-23, ACL'22, EMNLP'21, ACL'19-21, EMNLP'19, NAACL'21 and EACL'21 and an AE for ACL-RR. He gave tutorials at IEEE Vis'22, ACL'19, ICDM'18 and COLING'18, and taught deep learning for NLP,[1] a graduate-level NLP course, and an undergraduate NLP course at NTU.

[1] https://ntunlpsg.github.io/ce7455_deep-nlp-20/

**Enamul Hoque**, York University, Canada
email: enamulh@yorku.ca
website: https://www.yorku.ca/enamulh/
Bio. Enamul Hoque is an Associate Professor at York University where he directs the Intelligent Visualization Lab. Previously, he was a postdoctoral fellow in Computer Science at Stanford University. He received the Ph.D. degree in Computer Science from the University of British Columbia. His research focuses on combining information visualization and human-computer interaction with natural language processing to address the challenges of the information overload problem. Recently, he has worked on developing natural language interfaces for visualizations, automatic chart question answering, chart retrieval and chart summarization. He has also worked on developing visual text analytics to support the user's task of exploring and analyzing conversations. Since his research is uniquely positioned at the intersection of information visualization, NLP, and HCI, he publishes at the major venues in each of these areas such as IEEE Vis, ACL, EMNLP, CHI, and UIST. He serves as an Area Chair for the ACL Rolling Review (2021-) and as a program committee member (2018-) for the IEEE Vis. He has also been teaching the graduate-level Information Visualization course at York University for the past 3 years.

**Jesse Vig**, Salesforce Research, USA
email: jesse.vig@gmail.com
website: https://jessevig.com/
Bio. Jesse Vig is a lead research scientist at Salesforce Research working on NLP, explainable AI, and HCI. Much of his research has explored novel interpretability methods, ranging from causal analysis of language models to attention interpretation in protein sequence models. He developed the BertViz[2] library for visualizing attention in Transformer models, as well as the SummVis and ProVis visualization tools. His work has appeared in NeurIPS, ICLR, IUI, UIST, ACL, NAACL, FAccT, and WWW, as well as the VISxAI and BlackBoxNLP workshops. Vig's research has been recognized with a Best Paper award at the Intelligent User Interfaces conference.

[2] https://github.com/jessevig/bertviz

# T2 - Security Challenges in Natural Language Processing Models

Qiongkai Xu, Xuanli He
Cutting-Edge
Wednesday, December 6, 2023 - 9:00-12:30 (Pisces 2 & 3)

https://emnlp2023-nlp-security.github.io

Large-scale natural language processing models have been developed and integrated into numerous applications, given the advantage of their remarkable performance. Nonetheless, the security concerns associated with these models prevent the widespread adoption of these black-box machine learning models. In this tutorial, we will dive into three emerging security issues in NLP research, i.e., backdoor attacks, private data leakage, and imitation attacks. In this tutorial, these threats will be introduced in accordance with their threatening usage scenarios, attack methodologies, and defense technologies.

---

**Qiongkai Xu**, Macquarie University & the University of Melbourne, Australia
email: qiongkai.xu@mq.edu.au
website: https://xuqiongkai.github.io
Dr. Qiongkai Xu is a Lecturer at Macquarie University, having earned his PhD from the Australian National University and previously served as a research fellow at the University of Melbourne. His research primarily focuses on Natural Language Processing, Privacy & Security, Machine Learning and Data Mining. Recently, his attention has been directed towards auditing machine learning models, specifically in two areas: 1) identifying and addressing privacy and security issues in ML/NLP models and their applications and 2) developing comprehensive evaluation theory and methods for ML/NLP models from various perspectives.

**Xuanli He**, University College London, UK
email: xuanli.he@ucl.ac.uk
website: https://xlhex.github.io
Dr. Xuanli He is a Research Fellow at University College London, having earned his PhD from Monash University. His recent research lies in an intersection between deep learning and natural language processing, with an emphasis on robustness and security in NLP models, including privacy leakage and protection, backdoor attack and defense, and imitation attack and defense. He has published more than 20 papers in top-tier machine learning and natural language processing conferences.

# T3 - Designing, Evaluating, and Learning from Humans Interacting with NLP Models

Tongshuang Wu, Diyi Yang and Sebastin Santy

Cutting-Edge

Wednesday, December 6, 2023 - 9:00-12:30 (Leo 3 & 4)

https://nlp-hci.github.io/tutorial/

The rapid advancement of natural language processing (NLP) research has led to various applications spanning a wide range of domains that require models to interact with humans — e.g., chatbots responding to human inquiries, machine translation systems assisting human translators, designers prompting Large Language Models for co-creation or prototyping AI-infused applications, etc. In these cases, humans interaction is key to the success of NLP applications; any potential misconceptions or differences might lead to error cascades at the subsequent stages. Such interaction involves a lot of design choices around models, e.g. the sensitivity of interfaces, the impact of design choice and evaluation questions, etc.

This tutorial aims to provide a systematic and up-to-date overview of key considerations and effective approaches for studying human-NLP model interactions. Our tutorial will focus specifically on the scenario where end users – lay people and domain experts who have access to NLP models but are less familiar with NLP techniques — use or collaborate with deployed models.

Throughout the tutorial, we will use five case studies (on classifier-assisted decision making, machine-aided translation, dialog systems, and prompting) to cover three major themes: (1) how to conduct human-in-the-loop usability evaluations to ensure that models are capable of interacting with humans; (2) how to design user interfaces (UIs) and interaction mechanisms that provide end users with easy access to NLP models; (3) how to learn and improve NLP models through the human interactions. We will use best practices from HCI to ground our discussion, and will highlight current challenges and future directions.

---

**Sherry Tongshuang Wu**, Carnegie Mellon University, USA
email: sherryw@cs.cmu.edu
website: http://cs.cmu.edu/~sherryw
Sherry is an assistant professor at the Human-Computer Interaction Institute, Carnegie Mellon University. Her primary research investigates how humans (AI experts, lay users, domain experts) interact with (debug, audit, and collaborate) AI systems. Sherry has organized two workshops at NLP and HCI conferences: Shared Stories and Lessons Learned workshop at EMNLP 2022 and Trust and Reliance in AI-Human

Teams at CHI 2022 and 2023. Sherry and Diyi have co-developed a new course on Human-Centered NLP that has been offered at both CMU and Stanford.

**Diyi Yang**, Stanford University, USA
email: diyiy@cs.stanford.edu
website: https://cs.stanford.edu/~diyiy/
Diyi is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on human-centered natural language processing and computational social science. Diyi has organized multiple workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, NLG Evaluation workshop at EMNLP 2021, and Shared Stories and Lessons Learned workshop at EMNLP 2022. She also gave a tutorial at ACL 2022 on Learning with Limited Data, and a tutorial at EACL 2023 on Summarizing Conversations at Scale.

**Sebastin Santy**, University of Washington, USA
email: ssanty@cs.washington.edu
website: https://sebastinsanty.com/
Sebastin is a third-year PhD student at the Paul G. Allen School of Computer Science & Engineering, University of Washington. He works on problems at the intersection of NLP and HCI and specifically his research focuses on uncovering design biases in NLP systems and building natural language user interfaces. He has previously worked on multilingual NLP and machine translation.

# T4 - LLM-driven Instruction Following: Progresses and Concerns

Wenpeng Yin

Qinyuan Ye

Pengfei Liu

Xiang Ren

Hinrich Schütze

Cutting-edge
Wednesday, Dec 6, 2023 - 14:00-17:30 (Leo 3 & 4)

www.wenpengyin.org/publications/instruction-following-emnlp23

The progress of natural language processing (NLP) is primarily driven by machine learning that optimizes a system on a large-scale set of task-specific labeled examples. This learning paradigm limits the ability of machines to have the same capabilities as humans in handling new tasks since humans can often solve unseen tasks with a couple of examples accompanied by task instructions. In addition, we may not have a chance to prepare task-specific examples of large-volume for new tasks because we cannot foresee what task needs to be addressed next and how complex to annotate for it. Therefore, task instructions act as a novel and promising resource for supervision.

This tutorial targets researchers and practitioners who are interested in AI and ML technologies for NLP generalization in a low-shot scenario. In particular, we will present a diverse thread of instruction-driven NLP studies that try to answer the following questions: (i) What is task instruction? (ii) How is the process of creating datasets and evaluating systems conducted? (iii) How to encode task instructions? (iv) When and why do some instructions work better? (v) What concerns remain in LLM-driven instruction following? We will discuss several lines of frontier research that tackle those challenges and will conclude the tutorial by outlining directions for further investigation.

---

**Wenpeng Yin**, Penn State University, USA
email: wenpeng@psu.edu
website: www.wenpengyin.org

Wenpeng is an Assistant Professor in the Department of Computer Science and Engineering at Penn State University. His research focuses on NLP with three sub-areas: (i) learning from task instructions; (ii) information extraction; (iii) NLP for education, bioinformatics, etc. Dr. Yin has presented the tutorial "Indirectly Supervised Natural Language Processing" at ACL'23, and the tutorial "Learning from Task Instructions" at KONVENS'23.

**Qinyuan Ye**, University of Southern California, USA
email: qinyuany@usc.edu
website: yeqy.xyz
Qinyuan is a fifth-year Ph.D. student at the University of Southern California, advised by Prof. Xiang Ren. Her research interest lies in natural language processing. In particular, she is interested in approaches that reduce human annotation efforts, including methods leveraging distant supervision, high-level human supervision (e.g., explanations, instructions), and meta-learning.

**Pengfei Liu**, Shanghai Jiaotong University, China
email: pengfei@sjtu.edu.cn
website: http://pfliu.com
Pengfei is an associate professor at Shanghai Jiaotong University and leads the Generative Artificial Intelligence Research Lab (GAIR). His research topics currently focus on information extraction, text generation, language pre-training, and NLP system evaluation. He won the Best Demo Paper award in ACL 2021 and the Outstanding Demo Paper award in ACL 2022.

**Xiang Ren**, University of Southern California, USA
email: xiangren@usc.edu
website: https://shanzhenren.github.io
Xiang is an Associate Professor in Computer Science and the Andrew and Erna Viterbi Early Career Chair at USC. Ren's research seeks to build generalizable NLP systems that can handle a wide variety of language tasks and situations. He works on new algorithms and datasets to make NLP systems cheaper to develop and maintain, arm machine models with common sense, and improve model's transparency and reliability to build user trust. His research work has received several best paper awards in top NLP and AI conference venues. Ren has been awarded an NSF CAREER Award, multiple faculty research awards from Google, Facebook, Amazon, JP Morgan and Sony, and the 2018 ACM SIGKDD Doctoral Dissertation Award. He was named Forbes' Asia 30 Under 30 in 2019. Ren has presented a number of tutorials, such as Knowledge-Augmented Methods for Natural Language Processing at ACL 2022, Scalable Construction and Reasoning of Massive Knowledge Bases at NAACL 2018, and other related tutorials at WWW'18, CIKM'17, etc.

**Hinrich Schütze**, LMU Munich, Germany
email: hinrich@hotmail.com
website: https://schuetze.cis.lmu.de
Hinrich is Chair of Computational Linguistics and co-director of the Center of Information and Language Processing at Ludwig-Maximilians-Universität München (LMU Munich), Germany. He was the President of the Association for Computational Linguistics in 2020, and General Chair of ACL 2013. In 2022, Prof. Schütze was elected as ACL Fellow. Prior to joining LMU Munich, he was a Professor of Theoretical Computational Linguistics at the University of Stuttgart. Hinrich holds a Ph.D. in computational linguistics from Stanford University.

# T5 - Mitigating Societal Harms in Large Language Models

Vidhisha Balanchandran, Sachin Kumar, Antonios Anastasopoulos, Lucille Njoo

Cutting Edge

Wednesday, Dec 6, 2023 - 14-17:30 (Pisces 1)

https://llm-harm-mitigation.github.io/tutorial

As with all language, text generated by language models can be harmful, or used to bring about harm. Automating language generation with large language models adds both an element of scale and also more subtle or emergent undesirable tendencies to the generated text. With an ever increasing number of user-facing applications being built and deployed on top of large language models, there are risks of societal harms that such applications can cause. It is of utmost importance for NLP practitioners to be aware of these risks as well as ways to mitigate them.

In this tutorial, we will provide a systematic and up-to-date review of potential risks of harms and social issues surrounding language generation systems and discuss methods to mitigate them. We will cover harms ranging from discrimination, toxicity and bias, to privacy, to misinformation and factuality. Our primary focus will be on how to systematically identify risks, and how to eliminate them at various stages of NLP pipeline, from data collection, model development, model adaptation, inference, to application deployment. Through this tutorial, we aim to equip NLP researchers and practitioners with a suite of practical tools for mitigating safety risks from language generation models, while highlighting current challenges and future research directions.

---

**Vidhisha Balachandran**, Carnegie Mellon University, USA
email: vbalacha@andrew.cmu.edu
website: https://vidhishanair.github.io/
Vidhisha Balachandran is a PhD candidate at the Language Technologies Institute at Carnegie Mellon University. Her research focuses on understanding reliability concerns in language systems and developing methods to address them. Her work spans multiple research areas like interpretability, factuality and knowledge integration in NLP. She has published in NLP/ML venues like EMNLP, NAACL ICLR and EACL, organized a workshop on interpreting and evaluating LLMs in COLING 2022 and co-taught a previous version of this tutorial with Sachin at The Web Conference 2022.

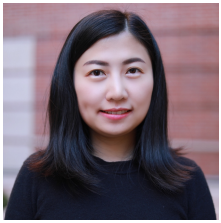
**Sachin Kumar**, Allen Institute for AI, USA
email: sachin@allenai.org
website: http://shocheen.com

Sachin is a postdoctoral researcher at Allen Institute for AI and an incoming assistant professor at the Ohio State University (Fall 2024). He obtained his Ph.D. in Language Technologies at Carnegie Mellon University. His research broadly revolves around Machine Learning and Natural Language Processing (NLP) with a particular focus on user-adaptable and controllable models. His work has been published in ML and NLP venues such as ICLR, NeurIPS, ACL, EMNLP, and EACL. Vidhisha and he previously co-taught a version of this tutorial at The Web Conference 2022.

**Antonios Anastasopolous**, George Mason University, USA
email: antonis@gmu.edu
Antonios Anastasopoulos is an Assistant Professor in Computer Science at George Mason University. He received his PhD in Computer Science from the University of Notre Dame and then did a postdoc at Languages Technologies Institute at Carnegie Mellon University. His research is on natural language processing with a focus on low-resource settings, endangered languages, and cross-lingual learning, and is currently funded by the National Science Foundation, the National Endowment for the Humanities, the DoD, Google, Amazon, and Meta. Antonis co-taught a tutorial on Endangered Languages and NLP in COLING 2020.

**Lucille Njoo**, University of Washington, USA
email: lnjoo@cs.washington.edu
Lucille Njoo is a third-year PhD student at the Paul G. Allen School of Computer Science and Engineering at the University of Washington, exploring the intersection of NLP, ethics, and computational social science. She works on identifying societal harms in NLP models and also focuses on developing ways to use language technologies in real-world, high-stakes scenarios where language is intertwined with social context and pragmatics.

# T6 - Creative Natural Language Generation



Tuhin Chakrabarty, Vishakh Padmakumar, He He, Nanyun Peng

Cutting-Edge

Wednesday, December 6, 2023 - 14:00-17:30 (Pisces 2 & 3)

https://emnlp2023-creative-nlg.github.io

Large language models such as GPT-3, GPT4, Claude etc., have advanced the state of the art in several natural language generation tasks such as text summarization and machine translation. However, when it comes to open-ended tasks with a focus on creativity such as generating stories, poetry, or various forms of figurative language, these state-of-the-art language models are often found to be inadequate.

This tutorial aims to bring awareness of the important and emerging research area of open-domain creative generation, with a focus on language generation while also touching on multi-modal generation (e.g., image captioning, visual metaphors). It targets natural language processing (NLP) and artificial intelligence (AI) researchers as well as creative writing practitioners who are interested in building systems that are capable of emulating as well as augmenting human creativity.

In particular, we will review recent studies on creative language generation both at the sentence level as well as longer forms of text. We will provide the audiences with a holistic view of 1) the importance and challenges of building creative language generation systems; 2) how we incorporate content planning, domain knowledge, and creativity-specific heuristics for different forms of creative language generation such as story, poetry, humor, metaphors, etc 3) how can we build better evaluation methods for creative text generation in standalone as well as interactive settings? In particular, how could the recent advancement of AI shape the future workforce for creativity? We will conclude the tutorial by outlining future research directions in this area.

---

**Tuhin Chakrabarty**, Columbia University, USA
email: tuhin.chakr@cs.columbia.edu
website: https://tuhinjubcse.github.io/

Bio. Tuhin Chakrabarty is a Ph.D. candidate in Computer Science at Columbia University and a part of the Natural Language Processing group, where he is advised by Smaranda Muresan. His research is supported by the Columbia Center of Artificial Intelligence & Technology (CAIT) and Amazon Science Ph.D. Fellowship. He was also a fellow at The New York Times R&D team working on Natural Language Generation. His overarching research question centers around how we can use large language models for creativity. He has published several papers in various NLP conferences and journals including ACL, NAACL, TACL and EMNLP.

**Vishakh Padmakumar**, New York University, USA
email: vishakh@nyu.edu
website: https://vishakhpk.github.io/
Bio. Vishakh Padmakumar is a Ph.D. student in Data Science at New York University advised by He He. His research is broadly in the field of natural language processing and human-AI collaboration with a focus on collaborative text generation for creative writing tasks and other interactive settings. Prior to this, he was a Graduate Research Associate at the NYU Center for Social Media and Politics working on political stance classification and multimodel content sharing in online disinformation campaigns. He has published papers at several NLP and machine learning venues including ACL, EMNLP, and ICML and was the chair of the ACL 2023 Student Research Workshop.

**Nanyun Peng**, University of California Los Angeles, USA
email: violetpeng@cs.ucla.edu
website: https://vnpeng.net/
Bio. Nanyun Peng is an Assistant Professor in the Department of Computer Science at the University of California Los Angeles. She received her Ph.D. in Computer Science from Johns Hopkins University. Her research focuses on the generalizability of NLP technologies, with applications to creative language generation, low-resource information extraction, and zero-shot cross-lingual transfer. Her works have won the Outstanding Paper Award at NAACL 2022, the Best Paper Award at AAAI 2022 Deep Learning on Graphs workshop, and have been featured an IJCAI 2022 early career spotlight.She has given a tutorial at NAACL 2018 on information extraction.

**He He**, New York University, USA
email: hhe@cs.nyu.edu
website: https://hhexiy.github.io/
Bio: He He is an Assistant Professor of Computer Science and the Center for Data Science at New York University. She is affiliated with the CILVR Lab, the Machine Learning for Language Group, and the Alignment Research Group. Her research focuses on building intelligent systems that can communicate with humans effectively and enable individuals to achieve their goals. Today's systems are often opaque, brittle, and difficult to control, which limits their usefulness in human-centered applications. To make them our trustworthy collaborators, her research aims to (i) understand the computational foundation of generalization in novel scenarios, and (ii) build interactive systems that align with users' goals. She has given a tutorial at EMNLP 2021 on robustness and adversarial examples in NLP.

# 10

# Workshops: December 6 & 7, 2023

## Overview

Prior day (Dec. 5) Registration will be held from 18:00 - 21:00.
During the days of the workshops (Dec. 6 & 7), **Registration** will be held from 07:30.
Do not forget the Welcome Reception on Dec. 07, 19:00-21:30.

### Wednesday, December 6, 2023


| | | |
|---|---|---|
| West 1 | **W1** - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only) | p.71 |
| West 3 | **W2** - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023) | p.74 |
| Central 1 | **W3** - The Eighth Conference on Machine Translation (WMT23) | p.75 |
| Central 3 | **W4** - GenBench: The first workshop on generalisation (benchmarking) in NLP | p.79 |
| Virgo 3 | **W5** - The 4th International Workshop on Computational Approaches to Historical Language Change (LChange'23) | p.82 |
| West 2 | **W6** - The 4th New Frontiers in Summarization Workshop (NewSumm) | p.84 |
| Virgo 1 | **W7** - The 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS) | p.85 |
| Virgo 2 | **W8** - The Pattern-based Approaches to NLP in the Age of Deep Learning Workshop (Pan-DL) | p.87 |
| Leo 1 | **W9** - The Seventh Widening NLP Workshop (WiNLP 2023) | p.88 |
| Pisces 4 | **W10** - Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP) | p.91 |
| Leo 2 | **W11** - Natural Language Generation, Evaluation, and Metric (GEM) | p.92 |

## Thursday, December 7, 2023


| | | |
|---|---|---|
| West 1 | **W1** - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only) | p.93 |
| West 3 | **W2** - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023) | p.96 |
| Central 1 | **W3** - The Eighth Conference on Machine Translation (WMT23) | p.97 |
| Central 3 | **W12** - The 10th Workshop on Argument Mining (ArgMining) | p.101 |
| Virgo 1 & 2 | **W13** - The Big Picture: Crafting a Research Narrative (BigPicture) | p.102 |
| West 2 | **W14** - BlackboxNLP 2023: The 6th Workshop on Analysing and Interpreting Neural Networks for NLP | p.103 |
| Virgo 3 | **W15** - The Sixth Workshop on Computational Approaches to Linguistic Code Switching | p.104 |
| Pisces 4 | **W16** - The Natural Legal Language Processing Workshop 2023 (NLLP) | p.106 |
| Pisces 1 | **W17** - The First Arabic Natural Language Processing Conference (ArabicNLP 2023) | p.109 |
| Leo 3 & 4 | **W18** - The Third Workshop on Multi-lingual Representation Learning (MRL) | p.116 |
| Leo 1 & 2 | **W19** - Novel Ideas in Learning to Learn through Interaction (NILLI) | p.117 |
| Pisces 2& 3 | **W20** - The First Bangla Language Processing Workshop (BLP) | p.118 |

# W1 - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only)

**Organizers:**
Jing Jiang, David Reitter, Shumin Deng

https://www.conll.org/2023
Venue: West 1
**Wednesday, December 6, 2023**

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 2023. Registrations for CoNLL can be made through EMNLP (workshop 1).

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 09:10 - 10:30 | ***Keynote 1 (Preslav Nakov)*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Oral Session 1*** |
| | *Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics*<br>Yuhan Zhang, Edward Gibson and Forrest Davis |
| | *ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind*<br>Xiaomeng Ma, Lingyu Gao and Qihui Xu |
| | *The Zipfian Challenge: Learning the statistical fingerprint of natural languages*<br>Christian Bentz |
| | *On the Effects of Structural Modeling for Neural Semantic Parsing*<br>Xiang Zhang, Shizhu He, Kang Liu and Jun Zhao |
| 12:30 - 13:45 | ***Lunch Break*** |
| 13:45 - 15:15 | ***Poster Session 1*** |
| | *Humans and language models diverge when predicting repeating text*<br>Aditya Vaidya, Javier Turek and Alexander Huth |
| | *Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum*<br>Urban Knuples, Diego Frassinelli and Sabine Schulte im Walde |
| | *ArchBERT: Bi-Modal Understanding of Neural Architectures and Natural Languages*<br>Mohammad Akbari, Saeed Ranjbar Alvar, Behnam Kamranian, Amin Banitalebi-Dehkordi and Yong Zhang |
| | *A Comparative Study on Textual Saliency of Styles from Eye Tracking, Annotations, and Language Models*<br>Karin de Langis and Dongyeop Kang |
| | *PROPRES: Investigating the Projectivity of Presupposition with Various Triggers and Environments*<br>Daiki Asami and Saku Sugawara |
| | *A Minimal Approach for Natural Language Action Space in Text-based Games* |

Dongwon Ryu, Meng Fang, Gholamreza Haffari, Shirui Pan and Ehsan Shareghi

*Structural Ambiguity and its Disambiguation in Language Model Based Parsers: the Case of Dutch Clause Relativization*
Gijs Wijnholds and Michael Moortgat

*On the utility of enhancing BERT syntactic bias with Token Reordering Pretraining*
Yassir El Mesbahi, Atif Mahmud, Abbas Ghaddar, Mehdi Rezagholizadeh, Phillippe Langlais and Prasanna Parthasarathi

*Quirk or Palmer: A Comparative Study of Modal Verb Frameworks with Annotated Datasets*
Risako Owan, Maria Gini and Dongyeop Kang

*Enhancing Code-mixed Text Generation Using Synthetic Data Filtering in Neural Machine Translation*
Dama Sravani and Radhika Mamidi

*Quantifying Information of Tokens for Simple and Flexible Simultaneous Machine Translation*
DongHyun Lee, Minkyung Park and Byung-Jun Lee

*Towards Better Evaluation of Instruction-Following: A Case-Study in Summarization*
Ondrej Skopek, Rahul Aralikatte, Sian Gooding and Victor Carbune

*Syntactic Inductive Bias in Transformer Language Models: Especially Helpful for Low-Resource Languages?*
Luke Gessler and Nathan Schneider

*Attribution and Alignment: Effects of Local Context Repetition on Utterance Production and Comprehension in Dialogue*
Aron Molnar, Jaap Jumelet, Mario Giulianelli and Arabella Sinclair

*[BabyLM Challenge] Baby Llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty*
Inar Timiryasov and Jean-Loup Tastet

*[BabyLM Challenge] BabyLM Challenge: Curriculum learning based on sentence complexity approximating language acquisition*
Miyu Oba, Akari Haga, Akiyo Fukatsu and Yohei Oseki

*[BabyLM Challenge] Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior?*
Grusha Prasad, Oliver Smith, Anzi Wang and Aryaman Datta Chobey

*[BabyLM Challenge] CogMemLM: Human-Like Memory Mechanisms Improve Performance and Cognitive Plausibility of LLMs*
Lukas Thoma, Ivonne Weyers, Erion Çano, Stefan Schweter, Jutta L Mueller and Benjamin Roth

*[BabyLM Challenge] McGill BabyLM Shared Task Submission: The Effects of Data Formatting and Structural Biases*
Ziling Cheng, Rahul Aralikatte, Ian Porada, Cesare Spinoso-Di Piano and Jackie CK Cheung

*[BabyLM Challenge] On the effect of curriculum learning with developmental data for grammar acquisition*
Mattia Opper, J Morrison and Siddharth N

*[BabyLM Challenge] ToddlerBERTa: Exploiting BabyBERTa for Grammar Learn- ing and Language Understanding*
Ömer Veysel Çağatan

15:15 - 15:30 ***Coffee Break***

15:30 - 17:00 ***Oral Session 2***

*The Validity of Evaluation Results: Assessing Concurrence Across Compositionality Benchmarks*
Kaiser Sun, Adina Williams and Dieuwke Hupkes

*Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning*
Lucas Weber, Elia Bruni and Dieuwke Hupkes

*Med-HALT: Medical Domain Hallucination Test for Large Language Models*
Ankit pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu

*Revising with a Backward Glance: Regressions and Skips during Reading as Cognitive Signals for Revision Policies in Incremental Processing*
Brielen Madureira, Pelin Çelikkol and David Schlangen

# W2 - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)

**Organizers:**
Maciej Ogrodniczuk, Sameer Pradhan, Vincent Ng, Massimo Poesio

https://sites.google.com/view/crac2023/
Venue: West 3
**Wednesday, December 6, 2023**

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, including both coreference and types of anaphora such as bridging references resolution and discourse deixis, can be presented.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W3 - The Eighth Conference on Machine Translation (WMT23)

**Organizers:**
Philipp Koehn, Barry Haddow, Tom Kocmi, Christof Monz

http://www2.statmt.org/wmt23/
Venue: Central 1
**Wednesday, December 6, 2023**

The long-standing Conference on Machine Translation (building on the earlier Workshop on Statistical Machine Translation) brings together researchers from the area of machine translation and features selected research papers to be presented at the conference. The conference also features a large number of shared tasks: a general translation task (former news task), a terminology translation task, a literary translation task, a word-level autocompletion task, a sign language translation task, a biomedical translation task, a indic languages translation task, an African languages translation task, a metrics evaluation task, a quality estimation task, a task to introduce novel machine translation testsuites, an automatic post-editing task, and a parallel data curation task.

| | |
|---|---|
| 08:45 - 09:00 | ***Opening Remarks*** |
| 09:00 - 10:30 | ***Session 1 — Shared Task Overview Papers I*** |
| 09:00-09:30 | *Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet*<br>Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda and Roman Grundkiewicz |
| 09:30-09:45 | *Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System*<br>Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro and Lana Yeganova |
| 09:45-10:00 | *Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs*<br>Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, C h a o - H o n g Liu and Yufeng Ma |
| 10:00-10:15 | *Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)*<br>Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina E s p a ñ a - B o n e t, Anne Göhring and Roman Grundkiewicz |
| 10:15-10:30 | *Findings of the WMT 2023 Shared Task on Parallel Data Curation*<br>Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda and Philipp Koehn |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Session 2 — Shared Task Overview Posters I*** |
| 11:00 - 12:30 | ***General Translation Task*** |
| 11:00-12:30 | *Samsung R&D Institute Philippines at WMT 2023*<br>Jan Christian Blaise Cruz |
| 11:00-12:30 | *NAIST-NICT WMT'23 General MT Task Submission*<br>Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli and Taro Watanabe |

| | |
|---|---|
| 11:00-12:30 | *CUNI at WMT23 General Translation Task: MT and a Genetic Algorithm*<br>Josef Jon, Martin Popel and Ondřej Bojar |
| 11:00-12:30 | *SKIM at WMT 2023 General Translation Task*<br>Keito Kudo, Takumi Ito, Makoto Morishita and Jun Suzuki |
| 11:00-12:30 | *KYB General Machine Translation Systems for WMT23*<br>Ben Li, Yoko Matsuzaki and Shivam Kalkar |
| 11:00-12:30 | *Yishu: Yishu at WMT2023 Translation Task*<br>Luo Min, Yixin Tan and Qiulin Chen |
| 11:00-12:30 | *PROMT Systems for WMT23 Shared General Translation Task*<br>Alexander Molchanov and Vladislav Kovalenko |
| 11:00-12:30 | *AIST AIRC Submissions to the WMT23 Shared Task*<br>Matiss Rikters and Makoto Miwa |
| 11:00-12:30 | *MUNI-NLP Submission for Czech-Ukrainian Translation Task at WMT23*<br>Pavel Rychly and Yuliia Teslia |
| 11:00-12:30 | *Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings*<br>Yangjian Wu and Gang Hu |
| 11:00-12:30 | *Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task*<br>Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie and Lizhi Lei |
| 11:00-12:30 | *UvA-MT's Participation in the WMT 2023 General Translation Shared Task*<br>Di Wu, Shaomu Tan, David Stap, Ali Araabi and Christof Monz |
| 11:00-12:30 | *Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters*<br>Hui Zeng |
| 11:00-12:30 | *IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task*<br>Wenbo Zhang |
| 11:00-12:30 | *GTCOM and DLUT's Neural Machine Translation Systems for WMT23*<br>Hao Zong |
| 11:00 - 12:30 | ***Test Suites*** |
| 11:00-12:30 | *RoCS-MT: Robustness Challenge Set for Machine Translation*<br>Rachel Bawden and Benoît Sagot |
| 11:00-12:30 | *Multifaceted Challenge Set for Evaluating Machine Translation Performance*<br>Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei and Hao Yang |
| 11:00-12:30 | *Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?*<br>Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina L a p s h i n o v a - K o l t u n s k i, Sergei Bagdasarov and Sebastian Möller |
| 11:00-12:30 | *IIIT HYD's Submission for WMT23 Test-suite Task*<br>Ananya Mukherjee and Manish Shrivastava |
| 11:00-12:30 | *Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES*<br>Beatrice Savoldi, Marco Gaido, Matteo Negri and Luisa Bentivogli |
| 11:00 - 12:30 | ***Biomedical Translation Task*** |
| 11:00-12:30 | *Biomedical Parallel Sentence Retrieval Using Large Language Models*<br>Sheema Firdous and Sadaf Rauf |
| 11:00-12:30 | *The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task* |

| | |
|---|---|
| | Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie and Lizhi Lei |
| 11:00-12:30 | *Investigating Techniques for a Deeper Understanding of Neural Machine Translation (NMT) Systems through Data Filtering and Fine-tuning Strategies*<br>Lichao Zhu, Maria Zimina, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski and J e a n - B a p t i s t e Yunès |
| 11:00 - 12:30 | ***Literary Translation Task*** |
| 11:00-12:30 | *MAX-ISI System at WMT23 Discourse-Level Literary Translation Task*<br>Li An, Linghao Jin and Xuezhe Ma |
| 11:00-12:30 | *The MAKE-NMTVIZ System Description for the WMT23 Literary Task*<br>Fabien Lopez, Gabriela González, Damien Hansen, Mariam Nakhle, Behnoosh Namdarzadeh, Nicolas Ballier, Marco Dinarelli, Emmanuelle E s p e r a n ç a - R o d i e r, Sui He and Sadaf Mohseni |
| 11:00-12:30 | *DUTNLP System for the WMT2023 Discourse-Level Literary Translation*<br>Anqi Zhao, Kaiyu Huang, Hao Yu and Degen Huang |
| 11:00-12:30 | *HW-TSC's Submissions to the WMT23 Discourse-Level Literary Translation Shared Task*<br>Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo and Lizhi Lei |
| 11:00-12:30 | *TJUNLP:System Description for the WMT23 Literary Task in Chinese to English Translation Direction*<br>Shaolin Zhu and Deyi Xiong |
| 11:00 - 12:30 | ***African Languages Translation Task*** |
| 11:00-12:30 | *Machine Translation for Nko: Tools, Corpora, and Baseline Results*<br>Moussa Doumbouya, Baba Diané, Solo Cissé, Djibrila Diané, Abdoulaye Sow, Séré Doumbouya, Daouda Bangoura, Fodé Bayo, Ibrahima Conde and Kalo Diané |
| 11:00 - 12:30 | ***Sign Language Translation Task*** |
| 11:00-12:30 | *TTIC's Submission to WMT-SLT 23*<br>Marcelo S a n d o v a l - C a s t a n e d a, Yanhong Li, Bowen Shi, Diane Brentari, Karen Livescu and Gregory Shakhnarovich |
| 11:00-12:30 | *KnowComp Submission for WMT23 Sign Language Translation Task*<br>Baixuan Xu, Haochen Shi, Tianshi Zheng, Qing Zong, Weiqi Wang, Zhaowei Wang and Yangqiu Song |
| 11:00 - 12:30 | ***Parallel Data Curation Task*** |
| 11:00-12:30 | *A Fast Method to Filter Noisy Parallel Data WMT2023 Shared Task on Parallel Data Curation*<br>N g u y e n - H o a n g M i n h - C o n g, Nguyen Vinh and Nguyen L e - M i n h |
| 11:00-12:30 | *A Sentence Alignment Approach to Document Alignment and Multi-faceted Filtering for Curating Parallel Sentence Pairs from Web-crawled Data*<br>Steinthor Steingrimsson |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Session 3 — Research Papers on Document-Level Translation and Use of Large Language Models*** |
| 14:00-14:15 | *Document-Level Language Models for Machine Translation*<br>Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi and Hermann Ney |
| 14:15-14:30 | *Identifying Context-Dependent Translations for Evaluation Set Production*<br>Rachel Wicks and Matt Post |
| 14:30-14:45 | *Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist*<br>Marzena Karpinska and Mohit Iyyer |
| 14:45-15:00 | *ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages*<br>Nathaniel Robinson, Perez Ogayo, David R. Mortensen and Graham Neubig |

| | |
|---|---|
| 15:00-15:15 | *Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA*<br>Xuan Zhang, Navid Rajabi, Kevin Duh and Philipp Koehn |
| 15:15-15:30 | *Towards Effective Disambiguation for Machine Translation with Large Language Models*<br>Vivek Iyer, Pinzhen Chen and Alexandra Birch |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Session 4 — Research Papers on Translation Modelling*** |
| 16:00-16:15 | *A Closer Look at Transformer Attention for Multilingual Translation*<br>Jingyi Zhang, Gerard De Melo, Hongfei Xu and Kehai Chen |
| 16:15-16:30 | *Bridging the Gap between Position-Based and Content-Based Self-Attention for Neural Machine Translation*<br>Felix Schmidt and Mattia Di Gangi |
| 16:30-16:45 | *Visual Prediction Improves Zero-Shot Cross-Modal Machine Translation*<br>Tosho Hirasawa, Emanuele Bugliarello, Desmond Elliott and Mamoru Komachi |
| 16:45-17:00 | *The Gender-GAP Pipeline: A Gender-Aware Polyglot Pipeline for Gender Characterisation in 55 Languages*<br>Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi, Christophe Ropers, Eric Smith, Adina Williams, Luke Zettlemoyer, Pierre Andrews and Marta R. C o s t a - J u s s à |
| 17:00-17:15 | *Towards Better Evaluation for Formality-Controlled English-Japanese Machine Translation*<br>Edison M a r r e s e - T a y l o r, Pin Chen Wang and Yutaka Matsuo |
| 17:15-17:30 | *There's No Data like Better Data: Using QE Metrics for MT Data Filtering*<br>J a n - T h o r s t e n Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska and Markus Freitag |

# W4 - GenBench: The first workshop on generalisation (benchmarking) in NLP

**Organizers:**
Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, Elia Bruni

https://genbench.org/workshop/
Venue: Central 3
**Wednesday, December 6, 2023**

The ability to generalise well is often mentioned as one of the primary desiderata for models of natural language processing (NLP). However, how generalisation should be defined and evaluated, or when it is particularly important, is a far from trivial question. The GenBench workshop on generalisation (benchmarking) in NLP aims to provide a platform to discuss challenging questions related to generalisation in NLP and establish a shared platform for state-of-the-art generalisation testing. We invited submitters to contribute work discussing generalisation in NLP and also held a collaborative benchmarking task, for which we called for submissions of challenging generalisation tests.

In this first edition of the workshop, we have 10 archival papers in our main track, 7 archival papers for our collaborative benchmarking track, and 6 extended abstracts. The workshop also provides a platform for the authors of 29 EMNLP findings paper related the workshop's topic to present their work as a poster at the workshop. In addition to poster sessions, we furthermore have three exciting invited speakers – Adina Williams, Anna Rogers and Tatsunori Hashimoto. They will talk about challenges in evaluating LLMs, how to consider emergent properties from the perspective of generalisation, and evaluating generalisation in the era of instruction tuning, respectively. We will end the day with an exciting panel in which we discuss challenging questions related to generalisation.

The workshop would not have been possible without the dedication of the programme committee, whom we would like to thank for their contributions. We would also like to thank Amazon for their sponsorship of 5000 dollars, which we used to fund one of our invited speakers, to grant travel awards to allow participants that could otherwise not have attended to participate in the workshop, and to grant two awards, to the best submitted paper and best submitted benchmark. Lastly, we are grateful to our invited speakers, Adina Williams, Anna Rogers, and Tatsunori Hashimoto, for contributing to our programme.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:00 | ***Keynote 1 by Anna Rogers: A sanity check on emergent properties*** |
| 10:00 - 11:15 | ***Poster Session 1*** |
| 10:00-11:15 | *Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models*<br>Jirui Qi, Raquel Fernández and Arianna Bisazza |
| 10:00-11:15 | *Temporal Generalizability in Multimodal Misinformation Detection*<br>Nataliya Stepanova and Björn Ross |
| 10:00-11:15 | *Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Context*<br>Michael Ginn and Alexis Palmer |
| 10:00-11:15 | *Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains* |

| | |
|---|---|
| | Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner and Carolin Lawrence |
| 10:00-11:15 | *The ICL Consistency Test*<br>Lucas Weber, Elia Bruni and Dieuwke Hupkes |
| 10:00-11:15 | *Generalizability and Robustness of Large Language Models Detecting Alzheimer's Disease from Speech*<br>Jekaterina Novikova |
| 10:00-11:15 | *Syntax-Guided Transformers: Elevating Compositional Generalization and Grounding in Multimodal Environments*<br>Danial Kamali and Parisa Kordjamshidi |
| 10:00-11:15 | *Inductive Bias Is in the Eye of the Beholder*<br>Michael Wilson and Robert Frank |
| 10:00-11:15 | *On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation*<br>Anssi Moisio, Mathias Creutz and Mikko Kurimo |
| 10:30 - 11:00 | ***Morning Coffee Break*** |
| 11:15 - 12:00 | ***Keynote 2 by Adina Williams: Evaluation after the LLM boom: frustrations, fallacies, and the future*** |
| 12:00 - 12:30 | ***CBT Spotlights*** |
| 12:00-12:08 | *GenCodeSearchNet: A Benchmark Test Suite for Evaluating Generalization in Programming Language Understanding*<br>Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler and Ansgar Scherp |
| 12:08-12:15 | *Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study*<br>Maike Züfle, Verna Dankers and Ivan Titov |
| 12:15-12:23 | *On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation*<br>Anssi Moisio, Mathias Creutz and Mikko Kurimo |
| 12:23-12:30 | *Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models*<br>Jirui Qi, Raquel Fernández and Arianna Bisazza |
| 12:30 - 14:00 | ***Lunch break*** |
| 14:00 - 14:45 | ***Keynote 3 by Tatsunori Hashimoto: Understanding generalization for instruction following and black-box language models*** |
| 14:45 - 15:30 | ***Oral presentations*** |
| 14:45-15:00 | *Evaluating Neural Language Models as Cognitive Models of Language Acquisition*<br>Hector Javier Vazquez Martinez, Annika Lea Heuser, Charles Yang and Jordan Kodner |
| 15:00-15:15 | *Robust Code Summarization*<br>Debanjan Mondal, Abhilasha Lodha, Ankita Sahoo and Beena Kumari |
| 15:15-15:30 | *Cross-Lingual Data Augmentation For Thai Question-Answering*<br>Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet and Peerat Limkonchotiwat |
| 15:30 - 16:00 | ***Afternoon Coffee Break*** |
| 16:00 - 17:00 | ***Poster session 2 (hybrid)*** |
| 16:00-17:00 | *90% F1 Score in Relation Triple Extraction: Is it Real?*<br>Pratik Saini, Samiran Pal, Tapas Nayak and Indrajit Bhattacharya |
| 16:00-17:00 | *mSCAN: A Dataset for Multilingual Compositional Generalisation Evaluation*<br>Amélie Reymond and Shane Steinert-Threlkeld |
| 16:00-17:00 | *GQG: Generalized Quantifier Generalization - A Dataset for Evaluating Quantifier Semantics Understanding in Language Models*<br>Leroy Zhifei Wang and Shane Steinert-Threlkeld |
| 16:00-17:00 | *Fighting Bias with Bias: Promoting Model Robustness by Amplifying Dataset Biases* |

| | |
|---|---|
| | Yuval Reif and Roy Schwartz |
| 16:00-17:00 | *GenCodeSearchNet: A Benchmark Test Suite for Evaluating Generalization in Programming Language Understanding*<br>Andor Diera, Abdelhalim Dahou, Lukas Galke, Fabian Karl, Florian Sihler and Ansgar Scherp |
| 16:00-17:00 | *Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study*<br>Maike Züfle, Verna Dankers and Ivan Titov |
| | *Blackbird Language Matrices Tasks for Generalization*<br>Paola Merlo, Chunyang Jiang, Giuseppe Samo and Vivi Nastase |
| 16:00-17:00 | *In-Context Learning for Text Classification with Many Labels*<br>Aristides Milios, Siva Reddy and Dzmitry Bahdanau |
| 16:00-17:00 | *Shifted PAUQ: Distribution shift in text-to-SQL*<br>Oleg Somov and Elena Tutubalina |
| 17:00 - 17:30 | ***Pannel*** |
| 17:30 - 17:45 | ***Closing Remarks and Best Paper Award*** |

# W5 - The 4th International Workshop on Computational Approaches to Historical Language Change (LChange'23)

**Organizers:**
Nina Tahmasebi, Syrielle Montariol, Haim Dubossarsky, Andrey Kutuzov, Simon Hengchen, David Alfter, Francesco Periti, Pierluigi Cassotti

https://www.changeiskey.org/event/2023-emnlp-lchange/
Venue: Virgo 3
**Wednesday, December 6, 2023**

The LChange workshop is an avenue on state-of-the-art computational methodologies, theories and digital text resources on exploring the time-varying nature of human language. The aim of this workshop is three-fold. First, we want to provide pioneering researchers who work on computational methods, evaluation, and large-scale modelling of language change an outlet for disseminating cutting-edge research on topics concerning language change. We particularly support discussion on the evaluation of computational methodologies for uncovering language change. Second, we want to bring together domain experts across disciplines by connecting researchers in historical linguistics with those that develop and test computational methods for detecting semantic change and laws of semantic change; and those that need knowledge (of the occurrence and shape) of language change, for example, in digital humanities and computational social sciences where text mining is applied to diachronic corpora subject to e.g., lexical semantic change. Third, the detection and modelling of language change using diachronic text and text mining raise fundamental theoretical and methodological challenges for future research.

| | |
|---|---|
| 09:15 - 09:30 | ***Introduction*** |
| 09:30 - 10:30 | ***Keynote Mario Giulianelli*** - Chair: Andrey Kutuzov |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:00 | ***Session 1*** - Chair: Pierluigi Cassotti |
| 11:00-11:20 | *EvoSem: A database of polysemous cognate sets*<br>Mathieu Dehouck, Alex François, Siva Kalyan, Martial Pastor and David Kletz |
| 11:20-11:40 | *Semantic Shifts in Mental Health-Related Concepts*<br>Naomi Baes, Nick Haslam and Ekaterina Vylomova |
| 11:40-12:00 | *Scent and Sensibility: Perception Shifts in the Olfactory Domain*<br>Teresa Paccosi, Stefano Menini, Elisa Leonardelli, Ilaria Barzon and Sara Tonelli |
| 12:00 - 13:30 | ***Lunch Break*** |
| 13:30 - 14:30 | ***Keynote Gemma Boleda*** - Chair: Syrielle Montariol |
| 14:30 - 15:30 | ***Session 2*** - Chair: Bill Noble |
| 14:30-14:50 | *Political dogwhistles and community divergence in semantic change*<br>Max Boholm and Asad B. Sayeed |
| 14:50-15:10 | *Automating Sound Change Prediction for Phylogenetic Inference: A Tukanoan Case Study*<br>Kalvin Chang, Nathaniel Romney Robinson, Anna Cai, Ting Chen, Annie Zhang and David R Mortensen |
| 15:10-15:30 | *Domain-Adapting BERT for Attributing Manuscript, Century and Region in Pre-Modern Slavic Texts* |

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus and Elena Renje

15:30 - 16:30 ***Poster Session***

*ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection*
Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic and Chu-Ren Huang

*Changing usage of Low Saxon auxiliary and modal verbs*
Janine Siewert, Martijn Wieling and Yves Scherrer

*Towards Detecting Lexical Change of Hate Speech in Historical Data*
Sanne Hoeken, Sophie Jasmin Spliethoff, Silke Schwandt, Sina Zarrieß and Özge Alacam

*From Diachronic to Contextual Lexical Semantic Change: Introducing Semantic Difference Keywords (SDKs) for Discourse Studies*
Isabelle Gribomont

*Multi-lect automatic detection of Swadesh list items from raw corpus data in East Slavic languages*
Ilia Afanasev

*Literary Intertextual Semantic Change Detection: Application and Motivation for Evaluating Models on Small Corpora*
Jackson Ehrenworth and Katherine A. Keith

*A longitudinal study about gradual changes in the Iranian Online Public Sphere pre and post of 'Mahsa Moment': Focusing on Twitter*
Sadegh Jafari, Amin Fathi, Abolfazl Hajizadegan, Amirmohammad Kazemeini and Sauleh Eetemadi

16:30 - 17:30 ***Session 3*** - Chair: Andrey Kutuzov

16:30-16:50 *Representing and Computing Uncertainty in Phonological Reconstruction*
Johann-Mattis List, Nathan Hill, Robert Forkel and Frederic Blum

16:50-17:10 *Anchors in Embedding Space: A Simple Concept Tracking Approach to Support Conceptual History Research*
Jetske Adams, Martha Larson, Jaap Verheul and Michael Boyden

17:10-17:30 *GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages*
Christin Beck and Marisa Köllner

17:30 - 17:45 ***Closing Remarks***

# W6 - The 4th New Frontiers in Summarization Workshop (NewSumm)

**Organizers:**
Yue Dong, Wen Xiao, Lu Wang, Fei Liu, Giuseppe Carenini

https://newsumm.github.io/2023/
Venue: West 2
**Wednesday, December 6, 2023**

NewSumm workshop, a key forum in its fourth edition, aims to develop intelligent systems for producing concise, fluent, and accurate summaries in natural language processing. It unites experts from diverse fields like summarization, language generation, and psycholinguistics to explore automatic summarization's critical aspects. The comprehensive agenda addresses innovative paradigms, multilingual setups, novel evaluation methods, and future research directions. This edition, following successful predecessors at EMNLP 2017, 2019, and 2021, received 31 paper submissions with a 42% acceptance rate. It features five esteemed speakers, including Kathleen McKeown, Jackie Cheung, Rui Zhang, Iz Beltagy, and Chenguang Zhu, representing a broad spectrum of expertise in the field. The workshop aims to build a cohesive research community and develop new tools and resources for academia, industry, and government.

| | |
|---|---|
| 08:50 - 09:00 | ***Opening Remarks*** |
| 09:00 - 09:45 | ***Keynote I - Kathleen McKeown (Columbia University)*** |
| 09:45 - 10:30 | ***Keynote II - Jackie Cheung (McGill University)*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 11:45 | ***Keynote III - Rui Zhang (Penn State University)*** |
| 11:45 - 12:30 | ***Keynote IV - Iz Beltagy (Allen Institute for AI)*** |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 14:45 | ***Keynote V - Chenguang Zhu (Zoom)*** |
| 14:45 - 15:30 | ***Lignting Talk (Workshop papers and Findings papers)*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Afternoon Session II - Poster Session (Workshop papers and Findings papers)*** |

# W7 - The 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS)

**Organizers:**
Liling Tan; Geeticka Chauhan; Dmitrijs Milajevs; Jeremy Gwinnup; Elijah Rippeth

https://nlposs.github.io/2023/index.html
Venue: Virgo 1
**Wednesday, December 6, 2023**

The primary objective of this workshop is to further the sharing of insights on the engineering and community aspects of creating, developing, and maintaining NLP open source software (OSS), which we seldom talk about in scientific publications. Our secondary goal is to promote synergies between different open source projects and encourage cross-software collaborations and comparisons.

| | |
|---|---|
| 09:00 - 09:15 | ***Opening Remarks*** |
| 09:15 - 10:15 | ***Invited Talk 1*** |
| | *trlX: A Framework for Large Scale Open Source RLHF*<br>Louis Castricato |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 11:30 | ***Lightning Session 1*** |
| 11:30 - 12:15 | ***Poster Session 1*** |
| | *Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models*<br>Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang and Han Xiao |
| | *Deepparse : A State-Of-The-Art Library for Parsing Multinational Street Addresses*<br>David Beauchemin and Marouane Yassine |
| | *PyThaiNLP: Thai Natural Language Processing in Python*<br>Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip and Can Udomcharoenchaikit |
| | *Zelda Rose: a tool for hassle-free training of transformer models*<br>Loïc Grobol |
| | *Kani: A Lightweight and Highly Hackable Framework for Language Model Applications*<br>Andrew Zhu, Liam Dugan, Alyssa Hwang and Chris Callison-Burch |
| | *Beyond the Repo: A Case Study on Open Source Integration with GECToR*<br>Sanjna Kashyap, Zhaoyang Xie, Kenneth Steimel and Nitin Madnani |
| | *EDGAR-CRAWLER: Finding Needles in the Haystack of Financial Documents*<br>Lefteris Loukas, Manos Fergadiotis and Prodromos Malakasiotis |
| | *Two Decades of the ACL Anthology: Development, Impact, and Open Challenges*<br>Marcel Bollmann, Nathan Schneider, Arne Köhn and Matt Post |
| | *nanoT5: Fast & Simple Pre-training and Fine-tuning of T5 Models with Limited Resources*<br>Piotr Nawrot |
| | *AWARE-TEXT: An Android Package for Mobile Phone Based Text Collection and On-Device Processing*<br>Salvatore Giorgi, Garrick Sherman, Douglas Bellew, Sharath Chandra Guntuku, Lyle Ungar and Brenda Curtis |

*SOTASTREAM: A Streaming Approach to Machine Translation Training*
Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain and Marcin Junczys-Dowmunt

*An Open-source Web-based Application for Development of Resources and Technologies in Underresourced Languages*
Siddharth Singh, Shyam Ratan, Neerav Mathur and Ritesh Kumar

*Rumour Detection in the Wild: A Browser Extension for Twitter*
Andrej Jovanovic and Björn Ross

12:15 - 13:45 ***Lunch Break***

13:45 - 14:45 ***Invited Talk 2***

*SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models*
David Ong and Peerat Limkonchotiwat

14:45 - 15:15 ***Lightning Session 2***

15:15 - 15:30 ***Coffee Break***

15:30 - 16:15 ***Poster Session 2***

*GPT4All: An Ecosystem of Open Source Compressed Language Models*
Yuvanesh Anand, Zach Nussbaum, Adam Treat, Aaron Miller, Richard Guo, Benjamin M Schmidt, Brandon Duderstadt and Andriy Mulyar

*LaTeX Rainbow: Open Source Document Layout Semantic Annotation Framework from LaTeX to PDF*
Changxu Duan and Sabine Bartsch

*DeepZensols: A Deep Learning Natural Language Processing Framework for Experimentation and Reproducibility*
Paul Landes, Barbara Di Eugenio and Cornelia Caragea

*Improving NER Research Workflows with SeqScore*
Constantine Lignos, Maya Kruse and Andrew Rueda

*torchdistill Meets Hugging Face Libraries for Reproducible, Coding-free Deep Learning Studies: A Case Study on NLP*
Yoshitomo Matsubara

*Using Captum to Explain Generative Language Models*
Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano and Narine Kokhlikyan

*nerblackbox: A High-level Library for Named Entity Recognition in Python*
Felix Stollenwerk

*News Signals: An NLP Library for Text and Time Series*
Chris Hokamp, Demian Gholipour Ghalandari and Parsa Ghaffari

*PyTAIL: An Open Source Tool for Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data*
Shubhanshu Mishra and Jana Diesner

*GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings*
Bang Fu and Di Feng

*The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation*
Dung Nguyen Manh, Nam Le Hai, Anh T. V. Dau, Anh Minh Nguyen, Khanh Nghiem, Jin Guo and Nghi D. Q. Bui

16:15 - 17:15 ***Invited Talk 3***

*Towards Explainable and Accessible AI*
Brandon Duderstadt and Yuvanesh Anand

17:15 - 17:30 ***Closing Remarks***

# W8 - The Pattern-based Approaches to NLP in the Age of Deep Learning Workshop (Pan-DL)

**Organizers:**
Enrique Noriega, Gus Hahn-Powell, Mihai Surdeanu, Clayton Morrison, Rebecca Sharp, Dayne Frietag, Ellen Riloff, Laura Chiticariu

https://pan-dl.github.io/2023/about

Venue: Virgo 2

**Wednesday, December 6, 2023**

This workshop will focus on all aspects of pattern-based approaches, including their application, representation, and interpretability, as well as their strengths and weaknesses relative to state-of-the-art machine learning approaches. It will also explore ways of combining the strengths of pattern-based, deep learning and other statistical methods.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W9 - The Seventh Widening NLP Workshop (WiNLP 2023)

**Organizers:**
Bonaventure F. P. Dossou, Shaily Bhatt, Sunipa Dev, Tirthankar Ghosal, Hatem Haddad, Haley Lepp, Fatemehsadat Mireshghallah Surangika Ranathunga, Alexandra Schofield, Isidora Chara Tourni, Weijia Xu, Atnafu Lambebo Tonja, Mukund Rungta

https://www.winlp.org/
Venue: Leo 1
**Wednesday, December 6, 2023**

The WiNLP workshop aims to foster an inclusive and diverse ACL environment by highlighting the work of underrepresented groups (URG) or anyone who self-identifies within an underrepresented demographic. The 2023 iteration of the workshop will build on the successes of prior iterations (2017-2022) with a focus on diversity in scientific background, discipline, training, obtained degrees, and seniority. Additionally, this iteration will target women and nonbinary researchers, the queer community, researchers outside the U.S. and Europe, and neurodiverse researchers in NLP. The full-day event will include a call for abstracts, a combination of invited talks, a panel discussion, oral presentations, and poster sessions. The workshop provides an opportunity for junior members of the community to showcase their work and connect with senior mentors for feedback and career advice. WiNLP also offers recruitment opportunities with leading industrial and academic labs. Most importantly, the workshop provides an accepting space that lowers structural barriers that make it difficult for URGs to join and collaborate with their NLP colleagues. The opportunity to present at the workshop is intended for URGs, and allies are encouraged to attend and support speakers. For more details on the vision, mission, and activities of WiNLP, visit our website above.

| | |
|---|---|
| 08:30 - 09:30 | ***Poster Session (Virtual)*** |
| 09:30 - 09:45 | ***Welcome*** |
| 09:45 - 10:30 | ***Invited Talk by Monojit Choudhury*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:00 | ***Panel Discussion: Misinformation in LLMs*** |
| 12:00 - 14:00 | ***Launch*** |
| 14:00 - 15:30 | ***Poster Session (In-Person)*** |
| | *STD-SQL: Seq2Seq Model Infused with SQL Grammar Structures with Text-to-SQL*<br>Jingyao Tang, Chenghu Zhou, Xinbing Wang and Zhouhan Lin |
| | *AmEn: Amharic-English Large Parallel Corpus for Machine Translation*<br>Tadesse Destaw Belay, Atnafu Lambebo Tonja, Seid Muhie Yimam and Abinew Ali Ayele |
| | *Am-QuAD: Amharic Question Answering Dataset*<br>Tilahun Abedissa Taffa, Ricardo Usbeck and Yaregal Assabie |
| | *Adversarial Robustness of Transformer-based Arabic Offensive Language Detectors*<br>Maged Abdelaty |
| | *Synthetic Data Augmentation for Low-Resource NMT A Case Study on Moroccan dialect*<br>Kamel Gaanoun and Imade Benelallam |
| | *FonMTL: Towards Multitask Learning for the Fon Language* |

Bonaventure F. P. Dossou, Iffanice Houndayi, Pamely Zantou and Gilles Hacheme

*Is it an Easy Task to Accurately Detect Automatically Generated Academic Content?*
Vijini Liyanage

*Usable Amharic Document for Natural Language Processing*
Michael Melese Woldeyohannis

*Error Analysis in Amharic Hate Speech Detection Task*
Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay and Chris Biemann

*YORC: Yoruba Reading Comprehension dataset*
Anuoluwapo Aremu, Jesujoba Alabi and David Ifeoluwa Adelani

*Bilingual Hate Speech Detection on Social Media Amharic and Afaan Oromo*
Teshome Mulugeta Ababu, Michael Melese Woldeyohannis and Emuye Getane

*Evaluating Large Language Models in Low-Resource Settings*
Bontu Balcha and Hellina Hailu Nigatu

*How good are Commercial Large Language Models on African Languages?*
Jessica Ojo and Kelechi Ogueji

*Do You Speak Basquenglish? Assessing Low-resource Multilingual Proficiency of Pretrained Language Models*
Inigo Parra

*Amharic Fake News Detection on social media Using. Pretrained Language Model*
Menbere Hailu Worku and Michael Melese Woldeyohannis

*An Exploratory Analysis of Differential Linguistic Features of Depression in Adolescents and Adults via Social Media*
Charlotte Rosario

*Measuring analogy resolution performance of clinical language models using a multilingual knowledge graph*
Fabian Villena and Jocelyn Dunstan

*Don't Overlook the Grammatical Gender: Bias Evaluation for Hindi-English Machine Translation*
Pushpdeep Singh

*Morpheme-based Bi-directional Machine Translation using Deep Learning*
Mengistu Negia and Rahel Tamiru

*Debiasing Stereotyped Models via Model Editing*
Xin Xu and Ningyu Zhang

*A Multi-way Parallel Named Entity Annotated Corpus for English, Tamil and Sinhala*
Surangika Ranathunga, Rashmi Galappaththi, Ayodya Dandeniya and Malithi Samaraweera

*Numerical Masking: Enhancing Numerical Proficiency in Language Models for Mathematical Word Problem Solving*
Nilesh Srivastava, K y u n g - M i n Kim and Seongchan Kim

*Ambient Adventures: Teaching ChatGPT on Developing Complex Stories*
Zexin Chen, Eric Zhou, Kenneth Eaton, Xiangyu Peng and Mark Riedl

*RoCode: A Dataset for Measuring Code Intelligence from Problem Definitions in Romanian*
Adrian Cosma, I o a n - B o g d a n Iordache and Paolo Rosso

*Social Media Portrayals of Happy Moments Among Depressed Individuals*
A n a - M a r i a Bucur, Berta Chulvi, Adrian Cosma and Paolo Rosso

*Hate Speech and Offensive Content Identification For Low-Resource Languages*
Mohammadmostafa Rostamkhani and Sauleh Eetemadi

*Error Analysis of Tigrinya-English Machine Translation Systems*
Negasi Abadi, Nuredin Ali Abdelkadir and Asmelash Teka Hadgu

*Mind What You Measure For: A Study on Reliability of Prompt-Based Bias Measurement*
Ruyuan Zuo and Jieyu Zhao

*Efficient domain adaptation while minimizing energy and hardware resource consumption.*
Hernan Maina, Nicolas Wolovick and Luciana Benotti

*Analyzing Text Generation of Large Language Models in Diverse English Dialects*
Victoria Graf, Dan Friedman and Danqi Chen

*Bridging Nations: Quantifying the Role of Multilinguals in Communication on Social Media*
Julia Mendelsohn, Sayan Ghosh, David Jurgens and Ceren Budak

*Analyzing Gender Accuracy and Gender Quality in Multilingual Machine Translation with Large Language Models*
Sarah Zhang, Lily Chen and William Zhang

*Digital and Historical Exclusivity in Femiinine Linguistics: From Nushu to Xiaohongshu*
Ruyuan Wan and Lingbo Tong

*StackOverflowVQA: Stack Overflow Visual Question Answering Dataset*
Motahhare Mirzaei, Mohammad Javad Pirhadi and Sauleh Eetemadi

*Regulating LLMs in AI: An analysis of whether LLMs are sensitive to transphobic content*
Anthony Li and Suzanna Sia

*Assessing Few-shot and Zero-shot Learning with CLIP Model for Visual Question Answering*
Ghazal Zamaninejad, Shaghayegh Mobasher and Sauleh Eetemadi

*FigurativeQA: Answering Yes/No Questions from Figurative Contexts*
Geetanjali Rakshit and Jeffrey Flanigan

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

16:00 - 16:45 ***Fireside Chat: Prof. Thamar Solorio and Dr. Katharina von der Wense***

16:45 - 17:00 ***Closing Remarks***

# W10 - Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP)

**Organizers:**
Malihe Alikhani, Ashiwarya Padmakumar, Xin Wang, Yue Fan, Mert Inan

https://splu-robonlp-2023.github.io/
Venue: Pisces 4
**Wednesday, December 6, 2023**

We wish the workshop to be the first step in building a community of researchers from different areas of NLP, both applied and theoretical, who are interested in pattern-based approaches and who use them in their work (e.g., industry practitioners and domain experts).

| | |
|---|---|
| 09:00 - 10:00 | ***Keynote 1*** |
| 10:00 - 10:30 | ***Session 1*** |
| 10:30 - 11:00 | ***Break*** |
| 11:00 - 12:00 | ***Session 2*** |
| 12:00 - 13:00 | ***Lunch Break*** |
| 13:00 - 13:30 | ***Lightning talks for Findings papers*** |
| 13:30 - 14:30 | ***Keynote 2*** |
| 14:30 - 15:30 | ***Panel Session*** |
| 16:00 - 15:30 | ***Break*** |
| 16:00 - 17:30 | ***Poster Session*** |

# W11 - Natural Language Generation, Evaluation, and Metric (GEM)

**Organizers:**
Khyathi Raghavi Chandu, Elizabeth Clark, Kaustubh Dhole, Sebastian Gehrmann, João Sedoc, Alex Wang, Enrico Santus, Hooman Sedghamiz

https://gem-benchmark.com/workshop
Venue: Leo 2
**Wednesday, December 6, 2023**

Natural language generation is one of the most active research areas within NLP and its barrier of entry has reduced dramatically. While applying supervised state of the art models to new data sets is becoming easier, the evaluation of models is becoming more challenging as models can produce completely fluent but meaningless or subtly flawed output. This leads to a disconnect between real-world needs of generation models and published research. Most of the disconnect can be bridged via in-depth evaluation and documentation of both data and models. To that end, the GEM workshop has three core goals: (1) Encourage the development of (semi-) automatic model audits and improved human evaluation strategies, (2) Popularize model evaluations in languages beyond English, (3) Provide a platform for discussions around evaluations to bridge the gap between industry and academia.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W1 - The SIGNLL Conference on Computational Natural Language Learning (CoNLL) (in-person-only)

**Organizers:**
Jing Jiang, David Reitter, Shumin Deng

https://www.conll.org/2023
Venue: West 1
**Thursday, December 7, 2023**

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 2023. Registrations for CoNLL can be made through EMNLP (workshop 1).

| | |
|---|---|
| 09:10 - 10:30 | ***Keynote 2 (Mohit Bansal)*** |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Oral Session 3*** |
| | *ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science*<br>Bram van Dijk, Max van Duijn, Suzan Verberne and Marco Spruit |
| | *HNC: Leveraging Hard Negative Captions towards Models with Fine-Grained Visual-Linguistic Comprehension Capabilities*<br>Esra Dönmez, Pascal Tilli, Hsiu-Yu Yang, Ngoc Thang Vu and Carina Silberer |
| | *Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests*<br>Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit and Peter vanderPutten |
| | *A Block Metropolis-Hastings Sampler for Controllable Energy-based Text Generation*<br>Jarad Forristal, Fatemehsadat Mireshghallah, Greg Durrett and Taylor Berg-Kirkpatrick |
| 12:30 - 13:45 | ***Lunch Break*** |
| 13:45 - 15:15 | ***Poster Session 2*** |
| | *How Fragile is Relation Extraction under Entity Replacements?*<br>Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan and Muhao Chen |
| | *JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models*<br>Yuiga Wada, Kanta Kaneda and Komei Sugiura |
| | *MuLER: Detailed and Scalable Reference-based Evaluation*<br>Taelin Karidi, Leshem Choshen, Gal Patel and Omri Abend |
| | *The Impact of Familiarity on Naming Variation: A Study on Object Naming in Mandarin Chinese*<br>Yunke He |
| | *PSST! Prosodic Speech Segmentation with Transformers*<br>Nathan Roll, Calbert Graham and Simon Todd |
| | *Alignment via Mutual Information*<br>Shinjini Ghosh, Yoon Kim, Ramon Fernandez Astudillo, Tahira Naseem and Jacob Andreas |

*Challenging the "One Single Vector per Token" Assumption*
Mathieu Dehouck

*Strategies to Improve Low-Resource Agglutinative Languages Morphological Inflection*
Gulinigeer Abudouwaili, Wayit Ablez, Kahaerjiang Abiderexiti, Aishan Wumaier and Nian Yi

*Exploring Transformers as Compact, Data-efficient Language Models*
Clayton Fields and Casey Kennington

*Tree-shape Uncertainty for Analyzing the Inherent Branching Bias of Unsupervised Parsing Models*
Taiga Ishii and Yusuke Miyao

*Future Lens: Anticipating Subsequent Tokens from a Single Hidden State*
Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace and David Bau

*Cross-Document Event Coreference Resolution: Instruct Humans or Instruct GPT?*
Jin Zhao, Nianwen Xue and Bonan Min

*Implications of Annotation Artifacts in Edge Probing Test Datasets*
Sagnik Ray Choudhury and Jushaan Kalra

*REFER: An End-to-end Rationale Extraction Framework for Explanation Regularization*
Mohammad Reza Ghasemi Madani and Pasquale Minervini

*[BabyLM Challenge] A surprisal oracle for active curriculum language modeling*
Xudong Hong, Sharid Loáiciga and Asad B. Sayeed

*[BabyLM Challenge] Baby's CoThought: Leveraging Large Language Models for Enhanced Reasoning in Compact Models*
Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer and Ercong Nie

*[BabyLM Challenge] Byte-ranked Curriculum Learning for BabyLM Strict-small Shared Task 2023*
Justin DeBenedetto

*[BabyLM Challenge] ChapGTP, ILLC's Attempt at Raising a BabyLM: Improving Data Efficiency by Automatic Task Formation*
Jaap Jumelet, Michael Hanna, Marianne De Heer Kloots, Anna Langedijk, Charlotte Pouw and Oskar van der Wal

*[BabyLM Challenge] GPT-wee: How Small Can a Small Language Model Really Get?*
Bastian Bunzeck and Sina Zarrieß

*[BabyLM Challenge] Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings*
David Samuel

*[BabyLM Challenge] Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks*
Clayton Fields, Osama Natouf, Andrew McMains, Catherine Henry and Casey Kennington

15:15 - 15:30 ***Coffee Break***

15:30 - 15:50 ***[BabyLM Challenge] Welcome and Findings Overview***

15:50 - 17:10 ***[BabyLM Challenge] Paper Session***

15:50-16:10 *Strict / Strict-Small Track Winner: Not all layers are equally as important: Every Layer Counts BERT*
Lucas Georges Gabriel Charpentier and David Samuel

16:10-16:30 *Loose Track Winner: Towards more Human-like Language Models based on Contextualizer Pretraining Strategy*
Chenghao Xiao, G Thomas Hudson and Noura Al Moubayed

16:30-16:50 *Outstanding Paper Award 1: Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures*
Julius Steuer, Marius Mosbach and Dietrich Klakow

16:50-17:10 *Outstanding Paper Award 2: CLIMB – Curriculum Learning for Infant-inspired Model Building*

| | |
|---|---|
| | Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery and Lisa Beinborn |
| 17:10 - 17:20 | ***[BabyLM Challenge] Closing Remarks*** |
| 17:20 - 17:35 | ***Best Paper Awards and Closing*** |

# W2 - The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)

**Organizers:**
Maciej Ogrodniczuk, Sameer Pradhan, Vincent Ng, Massimo Poesio

https://sites.google.com/view/crac2023/
Venue: West 3
**Thursday, December 7, 2023**

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, including both coreference and types of anaphora such as bridging references resolution and discourse deixis, can be presented.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W3 - The Eighth Conference on Machine Translation (WMT23)

**Organizers:**
Philipp Koehn, Barry Haddow, Tom Kocmi, Christof Monz

http://www2.statmt.org/wmt23/
Venue: Central 1
**Thursday, December 7, 2023**

The long-standing Conference on Machine Translation (building on the earlier Workshop on Statistical Machine Translation) brings together researchers from the area of machine translation and features selected research papers to be presented at the conference. The conference also features a large number of shared tasks: a general translation task (former news task), a terminology translation task, a literary translation task, a word-level autocompletion task, a sign language translation task, a biomedical translation task, a indic languages translation task, an African languages translation task, a metrics evaluation task, a quality estimation task, a task to introduce novel machine translation testsuites, an automatic post-editing task, and a parallel data curation task.

| | |
|---|---|
| 09:00 - 10:30 | ***Session 5 — Shared Task Overview Papers II*** |
| 09:00-09:15 | *Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent*<br>Markus Freitag, Nitika Mathur, C h i - K i u Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch and Craig Stewart |
| 09:15-09:30 | *Findings of the WMT 2023 Shared Task on Quality Estimation*<br>Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. De Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan and Fatemeh Azadi |
| 09:30-09:45 | *Findings of the Word-Level AutoCompletion Shared Task in WMT 2023*<br>Lemao Liu, Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Shuming Shi, Taro Watanabe and Chengqing Zong |
| 09:45-10:00 | *Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies*<br>Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou and Yuchen Eleanor Jiang |
| 10:00-10:15 | *Findings of the WMT 2023 Shared Task on Automatic Post-Editing*<br>Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri and Marco Turchi |
| 10:15-10:30 | *Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation*<br>Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure and Sandeep Kumar Dash |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:30 | ***Session 6 — Shared Task System Description Posters II*** |
| 11:00 - 12:30 | ***Metrics Task*** |
| 11:00-12:30 | *ACES: Translation Accuracy Challenge Sets at WMT 2023*<br>Chantal Amrhein, Nikita Moghe and Liane Guillou |
| 11:00-12:30 | *Challenging the State-of-the-art Machine Translation Metrics from a Linguistic Perspective*<br>Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz and Sebastian Möller |
| 11:00-12:30 | *Tokengram_F, a Fast and Accurate Token-based chrF++ Derivative*<br>Sören Dreano, Derek Molloy and Noel Murphy |

| | |
|---|---|
| 11:00-12:30 | *Embed_Llama: Using LLM Embeddings for the Metrics Shared Task*<br>Sören Dreano, Derek Molloy and Noel Murphy |
| 11:00-12:30 | *eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings*<br>Muhammad Elnokrashy and Tom Kocmi |
| 11:00-12:30 | *Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics*<br>Thamme Gowda, Tom Kocmi and Marcin J u n c z y s - D o w m u n t |
| 11:00-12:30 | *MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task*<br>Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh and Markus Freitag |
| 11:00-12:30 | *GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4*<br>Tom Kocmi and Christian Federmann |
| 11:00-12:30 | *Metric Score Landscape Challenge (MSLC23): Understanding Metrics' Performance on a Wider Landscape of Translation Quality*<br>C h i - K i u Lo, Samuel Larkin and Rebecca Knowles |
| 11:00-12:30 | *MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task*<br>Ananya Mukherjee and Manish Shrivastava |
| 11:00-12:30 | *Quality Estimation Using Minimum Bayes Risk*<br>Subhajit Naskar, Daniel Deutsch and Markus Freitag |
| 11:00-12:30 | *Evaluating Metrics for Document-context Evaluation in Machine Translation*<br>Vikas Raunak, Tom Kocmi and Matt Post |
| 11:00-12:30 | *Semantically-Informed Regressive Encoder Score*<br>Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger and Alexander Panchenko |
| 11:00-12:30 | *Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task*<br>Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao and Zhao Yanqing |
| 11:00 - 12:30 | ***Quality Estimation Task*** |
| 11:00-12:30 | *Unify Word-level and Span-level Tasks: NJUNLP's Participation for the WMT2023 Quality Estimation Shared Task*<br>Xiang Geng, Zhejian Lai, Yu Zhang, Shimin Tao, Hao Yang, Jiajun Chen and Shujian Huang |
| 11:00-12:30 | *HW-TSC 2023 Submission for the Quality Estimation Shared Task*<br>Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang and Hao Yang |
| 11:00-12:30 | *Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task*<br>Ricardo Rei, Nuno M. Guerreiro, Josã© Pombal, Daan Van Stigt, Marcos Treviso, Luisa Coheur, José G. C. De Souza and André Martins |
| 11:00-12:30 | *SurreyAI 2023 Submission for the Quality Estimation Shared Task*<br>Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan and Tharindu Ranasinghe |
| 11:00-12:30 | *MMT's Submission for the WMT 2023 Quality Estimation Shared Task*<br>Yulong Wu, Viktor Schlegel, Daniel Beck and Riza B a t i s t a - N a v a r r o |
| 11:00-12:30 | *IOL Research's Submission for WMT 2023 Quality Estimation Shared Task*<br>Zeyu Yan |
| 11:00 - 12:30 | ***Word-Level Autocompletion Task*** |
| 11:00-12:30 | *SJTU-MTLAB's Submission to the WMT23 Word-Level Auto Completion Task*<br>Xingyu Chen and Rui Wang |
| 11:00-12:30 | *PRHLT's Submission to WLAC 2023*<br>Angel Navarro, Miguel Domingo and Francisco Casacuberta |
| 11:00-12:30 | *KnowComp Submission for WMT23 Word-Level AutoCompletion Task*<br>Yi Wu, Haochen Shi, Weiqi Wang and Yangqiu Song |

| | |
|---|---|
| 11:00 - 12:30 | ***Terminology Translation Task*** |
| 11:00-12:30 | *Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting*<br>Nikolay Bogoychev and Pinzhen Chen |
| 11:00-12:30 | *Lingua Custodia's Participation at the WMT 2023 Terminology Shared Task*<br>Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout and Raheel Qadar |
| 11:00-12:30 | *Domain Terminology Integration into Machine Translation: Leveraging Large Language Models*<br>Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John Kelleher, Rejwanul Haque and Andy Way |
| 11:00-12:30 | *OPUS-CAT Terminology Systems for the WMT23 Terminology Shared Task*<br>Tommi Nieminen |
| 11:00-12:30 | *VARCO-MT: NCSOFT's WMT'23 Terminology Shared Task Submission*<br>Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell and Yeonsoo Lee |
| 11:00 - 12:30 | ***Automatic Postediting Task*** |
| 11:00-12:30 | *HW-TSC's Participation in the WMT 2023 Automatic Post Editing Shared Task*<br>Jiawei Yu, Min Zhang, Zhao Yanqing, Xiaofeng Zhao, Yuang Li, Su Chang, Yinglu Li, Ma Miaomiao, Shimin Tao and Hao Yang |
| 11:00 - 12:30 | ***Indic Languages Translation Task*** |
| 11:00-12:30 | *Neural Machine Translation for English - Manipuri and English - Assamese*<br>Goutam Agrawal, Rituraj Das, Anupam Biswas and Dalton Thounaojam |
| 11:00-12:30 | *GUIT-NLP's Submission to Shared Task: Low Resource Indic Language Translation*<br>Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Prof. Shikhar Kumar Sarma and Kishore Kashyap |
| 11:00-12:30 | *NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023*<br>Raj Dabre, Jay Gala and Pranjal Chitale |
| 11:00-12:30 | *Machine Translation Advancements for Low-Resource Indian Languages in WMT23: CFILT-IITB's Effort for Bridging the Gap*<br>Pranav Gaikwad, Meet Doshi, Sourabh Deoghare and Pushpak Bhattacharyya |
| 11:00-12:30 | *Low-Resource Machine Translation Systems for Indic Languages*<br>Ivana Kvapilíková and Ondřej Bojar |
| 11:00-12:30 | *MUNI-NLP Systems for Low-resource Indic Machine Translation*<br>Edoardo Signoroni and Pavel Rychly |
| 11:00-12:30 | *NITS-CNLP Low-Resource Neural Machine Translation Systems of English-Manipuri Language Pair*<br>Kshetrimayum Boynao Singh, Avichandra Singh Ningthoujam, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay and Thoudam Doren Singh |
| 11:00-12:30 | *IACS-LRILT: Machine Translation for Low-Resource Indic Languages*<br>Dhairya Suman, Atanu Mandal, Santanu Pal and Sudip Naskar |
| 11:00-12:30 | *IOL Research Machine Translation Systems for WMT23 Low-Resource Indic Language Translation Shared Task*<br>Wenbo Zhang |
| 12:30 - 14:00 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Session 7 — Panel on Large Language Models and Machine Translation*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Session 8 — Research Papers on Evaluation I*** |
| 16:00-16:15 | *Trained MT Metrics Learn to Cope with Machine-translated References*<br>Jannis Vamvas, Tobias Domhan, Sony Trenous, Rico Sennrich and Eva Hasler |
| 16:15-16:30 | *Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level*<br>Daniel Deutsch, Juraj Juraska, Mara Finkelstein and Markus Freitag |

16:30-16:45 *Automating Behavioral Testing in Machine Translation*
Javier Ferrando, Matthias Sperber, Hendra Setiawan, Dominic Telaar and Saša Hasan

16:45-17:00 *One Wide Feedforward Is All You Need*
Telmo Pires, António Vilarinho Lopes, Yannick Assogba and Hendra Setiawan

17:00-17:15 *A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography*
Noëmi Aepli, Chantal Amrhein, Florian Schottmann and Rico Sennrich

17:15-17:30 *The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation*
Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag and Orhan Firat

# W12 - The 10th Workshop on Argument Mining (ArgMining)

**Organizers:**
Milad Alshomary, Chung-Chi Chen, Smaranda Muresan, Joonsuk Park, Julia Romberg

https://argmining-org.github.io/2023/
Venue: Central 3
**Thursday, December 7, 2023**

The ArgMining workshop series is the premier research forum that drives the exploration of argument mining tasks in all domains of discourse. Since its inception in 2014, it has been held annually for nine consecutive years at major NLP conferences: ACL (2014, 2016, 2019), NAACL (2015), EMNLP (2017, 2018, 2021), and COLING (2020, 2022). Seeing the continuous emergence of argument mining research in the last years, this workshop provides a follow-on forum for the previous editions and the many recent relevant events. The workshop will take place in a hybrid setting and feature a panel session reflecting on the past and the future of argument mining in celebration of the 10th anniversary of the workshop series, a keynote speech, paper presentations, as well as two shared tasks.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W13 - The Big Picture: Crafting a Research Narrative (BigPicture)

**Organizers:**
Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, Noah A. Smith

https://www.bigpictureworkshop.com/
Venue: Virgo 1 & 2
**Thursday, December 7, 2023**

The Big Picture Workshop provides a dedicated venue for exploring and distilling broader NLP research narratives. All research exists within a larger context, and progress is made by standing on the shoulders of giants: building on the foundations laid by earlier researchers. In light of rapid publication rates and concise paper formats, it has become increasingly difficult, however, to recognize the larger story to which a paper is connected. The Big Picture Workshop invites researchers to reflect on how their individual contributions fit within the overall research landscape and what stories they are telling with their bodies of research. The goals of the workshop are to enhance communication and understanding between different lines of work, highlight how works connect and build on each other, generate insights that are difficult to glean without combining and reconciling different research narratives, encourage broader collaboration and awareness of prior work in the NLP community, and facilitate understanding of trajectories and insights within the field of NLP.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W14 - BlackboxNLP 2023: The 6th Workshop on Analysing and Interpreting Neural Networks for NLP

**Organizers:**
Yonatan Belinkov, Najoung Kim, Sophie Hao, Arya McCarthy, Jaap Jumelet, Hosein Mohebbi

https://blackboxnlp.github.io/
Venue: West 2
**Thursday, December 7, 2023**

Abstract.

09:00 - 09:00 ***Please refer to the above link for the program of the workshop***

# W15 - The Sixth Workshop on Computational Approaches to Linguistic Code Switching

**Organizers:**
Genta Indra Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, Kalika Bali

https://code-switching.github.io/2023
Venue: Virgo 3
**Thursday, December 7, 2023**

Bilingual and multilingual speakers often mix languages when they communicate with other multilingual speakers in what is usually known as code-switching (CS). CS can occur on various language levels including inter-sentential, intra-sentential, and even morphological. Practically, it presents long-standing challenges for language technologies, such as machine translation, ASR, language generation, information retrieval and extraction, and semantic processing. Models trained for one language can quickly break down when there is input mixed in from another. The recent breakthough on using multilingual pre-trained language models (LMs) have shown possibility to yield subpar performance on CS data. Considering the ubiquitous nature of CS in informal text such as newsgroups, tweets threads, and other forms of social media communication, and the number of multilingual speakers worldwide that use these platforms, addressing the challenge of processing CS data continues to be of great practical value. This workshop aims to bring together researchers interested in technology for mixed language data, in either spoken or written form, and increase community awareness of the different efforts developed to date in this space.

| | |
|---|---|
| 09:05 - 09:00 | ***Opening Remarks*** |
| 09:05 - 10:35 | ***Paper Oral Presentation 1*** |
| 09:05-09:20 | *Towards Real-World Streaming Speech Translation for Code-Switched Speech*<br>Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng and Aashish Agarwal |
| 09:20-09:35 | *Text-Derived Language Identity Incorporation for End-to-End Code-Switching Speech Recognition*<br>Qinyi Wang and Haizhou Li |
| 09:35-09:50 | *TongueSwitcher: Fine-Grained Identification of German-English Code-Switching*<br>Igor Sterner and Simone Teufel |
| 09:50-10:05 | *CONFLATOR: Incorporating Switching Point based Rotatory Positional Encodings for Code-Mixed Language Modeling*<br>Mohsin Ali Mohammed, Sai Teja Kandukuri, Neeharika Gupta, Parth Patwa, Anubhab Chatterjee, Vinija Jain, Aman Chadha and Amitava Das |
| 10:05-10:20 | *Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages*<br>Zheng Xin Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio and Alham Fikri Aji |
| 10:20-10:35 | *Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching*<br>Tolulope Ogunremi, Christopher Manning and Dan Jurafsky |

| | |
|---|---|
| 10:35 - 11:00 | ***Break*** |
| 11:00 - 11:45 | ***Invited Talk - Preethi Jyothi*** |
| 11:45 - 12:30 | ***Panel Discussion - Sudipta Kar, Genta Winata, Marina Zhukova*** |
| 14:00 - 12:30 | ***Lunch Break*** |
| 14:00 - 15:30 | ***Poster Session - Findings*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:45 | ***Invited Talk - Haizhou Li*** |
| 16:45 - 17:15 | ***Paper Oral Presentation 2*** |
| 17:00-17:15 | *Language Preference for Expression of Sentiment for Nepali-English Bilingual Speakers on Social Media*<br>Niraj Pahari and Kazutaka Shimada |
| 16:45-17:00 | *Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer*<br>Kunal Dhawan, KDimating Rekesh and Boris Ginsburg |
| 17:15 - 17:20 | ***Best Paper Announcement*** |
| 17:20 - 17:30 | ***Closing Remarks*** |

# W16 - The Natural Legal Language Processing Workshop 2023 (NLLP)

**Organizers:**
Nikolaos Aletras, Leslie Barrett, Ilias Chalkidis, Catalina Goanta, Daniel Preotiuc-Pietro, Jerry Spanakis

https://nllpw.org/workshop/
Venue: Pisces 4
**Thursday, December 7, 2023**

The Natural Legal Language Processing (NLLP) 2023 workshop, now at its fifth 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.

| | |
|---|---|
| 09:00 - 09:10 | ***Workshop Opening*** |
| 09:10 - 10:30 | ***Session 1*** - Chair: Ilias Chalkidis |
| 09:10-09:15 | *Anthropomorphization of AI: Opportunities and Risks*<br>Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan and Ashwin Kalyan |
| 09:15-09:20 | *Legal NLP Meets MiCAR: Advancing the Analysis of Crypto White Papers*<br>Carolina Camassa |
| 09:20-09:25 | *On the Potential and Limitations of Few-Shot In-Context Learning to Generate Metamorphic Specifications for Tax Preparation Software*<br>Dananjay Srinivas, Rohan Das, Saeid Tizpaz-Niari, Ashutosh Trivedi and Maria Leonor Pacheco |
| 09:35-09:40 | *NOMOS: Navigating Obligation Mining in Official Statutes*<br>Andrea Pennisi, Elvira González Hernández and Nina Koivula |
| 09:40-09:45 | *Long Text Classification using Transformers with Paragraph Selection Strategies*<br>Mohit Tuteja and Daniel González Juclà |
| 09:45-09:50 | *Italian Legislative Text Classification for Gazzetta Ufficiale*<br>Marco Rovera, Alessio Palmero Aprosio, Francesco Greco, Mariano Lucchese, Sara Tonelli and Antonio Antetomaso |
| 09:50-09:55 | *Towards Mitigating Perceived Unfairness in Contracts from a Non-Legal Stakeholder's Perspective*<br>Anmol Singhal, Preethu Rose Anish, Shirish Karande and Smita Ghaisas |
| 09:55-10:00 | *Low-Resource Deontic Modality Classification in EU Legislation*<br>Kristina Minkova, Shashank Chakravarthy and Gijs Dijck |

| | |
|---|---|
| 10:00-10:05 | *Transferring Legal Natural Language Inference Model from a US State to Another: What Makes It So Hard?*<br>Alice Kwak, Gaetano Forte, Derek Bambauer and Mihai Surdeanu |
| 10:05-10:10 | *Large Language Models for Legally Enforceable Hate Speech Detection*<br>Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan and Xiaodan Zhu |
| 10:10-10:15 | *More than Votes? Voting and Language based Partisanship in the US Supreme Court*<br>Biaoyan Fang, Trevor Cohn, Timothy Baldwin and Lea Frermann |
| 10:30 - 11:00 | ***Break*** |
| 11:00 - 12:30 | ***Session 2*** - Chair: Ilias Chalkidis |
| 11:00-11:05 | *Retrieval-Augmented Chain-of-Thought in Semi-structured Domains*<br>Vaibhav Mavi, Abulhair Saparov and Chen Zhao |
| 11:05-11:10 | *Mixed-domain Language Modeling for Processing Long Legal Documents*<br>Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li and Melanie Weber |
| 11:10-11:15 | *Large Language Models are legal but they are not: Making the case for a powerful LegalLLM*<br>Thanmay Jayakumar, Fauzan Farooqui and Luqman Farooqui |
| 11:15-11:20 | *Exploration of Open Large Language Models for eDiscovery*<br>Sumit Pai, Sounak Lahiri, Ujjwal Kumar, Krishanu Baksi, Elijah Soba, Michael Suesserman, Nirmala Pudota, Jon Foster, Edward Bowen and Sanmitra Bhattacharya |
| 11:20-11:25 | *A Comparative Study of Prompting Strategies for Legal Text Classification*<br>Ali Hakimi Parizi, Yuyang Liu, Prudhvi Nokku, Sina Gholamian and David Emerson |
| 11:25-11:30 | *A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction*<br>Ruihao Shui, Yixin Cao, Xiang Wang and Tat-Seng Chua |
| 11:40-11:45 | *SCALE: Scaling up the Complexity for Advanced Language Model Evaluation*<br>Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel Ho and Joel Niklaus |
| 11:45-11:50 | *Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US*<br>Biaoyan Fang, Trevor Cohn, Timothy Baldwin and Lea Frermann |
| 11:50-11:55 | *MultiLegalPile: A 689GB Multilingual Legal Corpus*<br>Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis and Daniel Ho |
| 11:55-12:00 | *ECtHR-PCR: A Dataset for Precedent Understanding and Prior Case Retrieval in the European Court of Human Rights*<br>Santosh T.y.s.s, Rashid Haddad and Matthias Grabmair |
| 12:00-12:05 | *AsyLex: A Dataset for Legal Language Processing of Refugee Claims*<br>Claire Barale, Mark Klaisoongnoen, Pasquale Minervini, Michael Rovatsos and Nehal Bhuta |
| 12:05-12:10 | *LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain*<br>Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer and Ilias Chalkidis |
| 12:30 - 14:00 | ***Lunch and In-Person Poster Session*** |
| 14:00 - 15:00 | ***Keynote - NLP in the Legal World - Jerrold Soh (Yong Pung How School of Law)*** - Chair: Nikolaos Aletras |
| 14:00 - 15:30 | ***Session 3*** - Chair: Catalina Goanta |
| 15:00-15:05 | *Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA*<br>Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min and Jy-Yong Sohn |
| 15:05-15:10 | *Joint Learning for Legal Text Retrieval and Textual Entailment: Leveraging the Relationship between Relevancy and Affirmation*<br>Nguyen Hai Long, Thi Hai Yen Vuong, Ha Thanh Nguyen and Xuan-Hieu Phan |
| 15:10-15:15 | *Tracing Influence at Scale: A Contrastive Learning Approach to Linking Public Comments and Regulator Responses*<br>Linzi Xing, Brad Hackinen and Giuseppe Carenini |
| 15:15-15:20 | *Legal Passage Retrieval: A pragmatic approach to legal AI* |

| | |
|---|---|
| | Robert Mahari, Dominik Stammbach, Elliott Ash and Alex Pentland |
| 15:30 - 16:00 | ***Break*** |
| 16:00 - 17:30 | ***Session 4*** - Chair: Daniel Preotiuc-Pietro |
| 16:00-16:05 | *Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset*<br>Santosh T.y.s.s, Nina Baumgartner, Matthias Stürmer, Matthias Grabmair and Joel Niklaus |
| 16:05-16:10 | *LLMs – the Good, the Bad or the Indispensable?: A Use Case on Legal Statute Prediction and Legal Judgment Prediction on Indian Court Cases*<br>Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Kumar Guha, Koustav Rudra and Kripabandhu Ghosh |
| 16:10-16:15 | *Exploiting Contrastive Learning and Numerical Evidence for Confusing Legal Judgment Prediction*<br>Leilei Gan, Baokui Li, Kun Kuang, Yating Zhang, Lei Wang, Anh Tuan Luu, Yi Yang and Fei Wu |
| 16:15-16:20 | *Multi-Defendant Legal Judgment Prediction via Hierarchical Reasoning*<br>Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang and Zhaochun Ren |
| 16:20-16:25 | *Pretrained Language Models v. Court Ruling Predictions: A Case Study on a Small Dataset of French Court of Appeal Rulings*<br>Olivia Vaudaux, Caroline Bazzoli, Maximin Coavoux, Géraldine Vial and Étienne Vergès |
| 16:25-16:30 | *Legal Judgment Prediction: If You Are Going to Do It, Do It Right*<br>Masha Medvedeva and Pauline Mcbride |
| 16:30-16:35 | *Extracting Sentencing-Related Factors from Criminal Cases in Hebrew*<br>Din Ezra, Maxim Bragilovski, Itay Razumenko, Keren Gorelik, Lior Kobi, Lior Rokach, Arnon Strum and Nir Grinberg |
| 16:40-16:45 | *Do Language Models Learn about Legal Entity Types during Pretraining?*<br>Claire Barale, Michael Rovatsos and Nehal Bhuta |
| 16:45-16:50 | *Questions about Contracts: Prompt Templates for Structured Answer Generation*<br>Adam Roegiest, Radha Chitta, Jonathan Donnelly, Maya Lash, Alexandra Vtyurina and Francois Longtin |
| 16:50-16:55 | *Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents*<br>Ramona Christen, Anastassia Shaitarova, Matthias Stürmer and Joel Niklaus |
| 16:55-17:00 | *Beyond The Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling*<br>Yan Shvartzshanider, Ananth Balashankar, Thomas Wies and Lakshminarayanan Subramanian |
| 17:00-17:05 | *Connecting Symbolic Statutory Reasoning with Legal Information Extraction*<br>Nils Holzenberger and Benjamin Van Durme |
| 17:05-17:10 | *Automatic Anonymization of Swiss Federal Supreme Court Rulings*<br>Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner and Marcel Gygli |
| 17:05-17:10 | *Automatic Anonymization of Swiss Federal Supreme Court Rulings*<br>Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner and Marcel Gygli |
| 17:10-17:15 | *Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?*<br>Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton and Genevieve Grant |
| 17:15-17:20 | *Information Extraction from Legal Wills: How Well Does GPT-4 Do?*<br>Alice Saebom Kwak, Cheonkam Jeong, Gaetano Vincent Forte, Derek Bambauer, Clayton T Morrison and Mihai Surdeanu |
| 17:30 - 17:35 | ***Closing Remarks & Best Presentation Award*** |

# W17 - The First Arabic Natural Language Processing Conference (ArabicNLP 2023)

**Organizers:**
Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, Rawan Almatham

https://wanlp2023.sigarab.org/
Venue: Pisces 1
**Thursday, December 7, 2023**

Welcome to The first Arabic Natural Language Processing Conference (ArabicNLP 2023) graduating from the Workshop for Arabic Natural Language Processing Workshop (WANLP) which had its seventh, and last, instance last year, in December 2022 within EMNLP 2022. Over the years, WANLP has developed a growing reputation as a high quality venue for researchers and engineers working on Arabic NLP, where they share and discuss their ongoing work. The first in the WANLP series was held in Doha, Qatar (EMNLP 2014), followed by Beijing, China (ACL 2015), Valencia, Spain (EACL 2017), Florence, Italy (ACL 2019), online with COLING 2020, online with EACL 2021, then finally a hybrid event in Abu Dhabi, UAE (EMNLP 2022). For this year's edition of ArabicNLP, we received a total of 80 main conference submissions and accepted 38 papers (32 long and 6 short), which brings us to an acceptance rate of 47.5%. All papers submitted to the conference were reviewed by at least three reviewers each. Out of the 80 submitted papers, there were 2 desk rejects. ArabicNLP 2023 included five shared tasks with 48 submissions in totals (i) The Nuanced Arabic Dialect Identification (NADI) with 13 submissions, (ii) ArAIEval (Persuasion Techniques and Disinformation Detection in Arabic Text) with 17 submissions, (iii) Qur'an QA with 6 submissions, (iv) WojoodNER with 8 submissions, and (v) Arabic Reverse Dictionary with 4 submissions. The shared task overview papers are included in the proceedings. The overview papers and the papers of the shared task winning systems are presented as talks during the conference. None of the shared task papers are counted toward the acceptance rate presented above. ArabicNLP 2023 also includes a panel discussing the hot topic "Arabic LLMs: Challenges and Opportunities' by leaders in the field, like Areeb Alowisheq, Tom Baldwin, and Kareem Darwish, moderated by Mona Diab. We were able to secure sponsorship funding from different institutions: King Salman Global Academy for Arabic Language, aiXplain, Lisan.ai, SCAI, Majarra, and Big IR, which we used to support student registrations. We thank all our sponsors for their generous support and their help in building up the Arabic NLP community. We would like to thank everyone who submitted a paper to the conference, as well as all the members of the Program Committee, who worked hard to provide reviews on a very tight schedule. Finally, on behalf of everyone involved, organizing committee as well as conference attendees, I would like to thank Nizar Habash and Houda Bouamor for supporting, mentoring and helping this conference be a success, being available for any request and filling any gaps that are overlooked. Hassan Sawaf, General Chair, on behalf of the conference organizers. Website of the conference: https://arabicnlp2023.sigarab.org.

| | |
|---|---|
| 09:00 - 09:20 | ***Welcome Session*** - Chair: Hassan Sawaf |
| 09:20 - 10:30 | ***Arabic Downstream NLP Tasks*** |

09:20-09:30 *In-Context Meta-Learning vs. Semantic Score-Based Similarity: A Comparative Study in Arabic Short Answer Grading*
Menna Fateen and Tsunenori Mina

09:30-09:40 *Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification*
Amr Keleg and Walid Magdy

09:40-09:50 *Nâbra: Syrian Arabic Dialects with Morphological Annotations*
Amal Nayouf, Tymaa Hasanain Hammouda, Mustafa Jarrar, Fadi A. Zaraket and Mohamad-Bassam Kurdy

09:50-10:00 *Arabic Fine-Grained Entity Recognition*
Haneen Abdallatif Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti and Muhammad Abdul-Mageed

10:00-10:10 *Cross-Dialectal Named Entity Recognition in Arabic*
Niama Elkhbir, Urchade Zaratiana, Nadi Tomeh and Thierry Charnois

10:10-10:20 *Leveraging Domain Adaptation and Data Augmentation to Improve Qur'anic IR in English and Arabic*
Vera Pavlova

10:20-10:30 *ArabIcros: AI-Powered Arabic Crossword Puzzle Generation for Educational Applications*
Kamyar Zeinalipour, Mohamed Zaky Saad, Marco Maggini and Marco Gori

10:30 - 11:00 ***Coffee Break***

11:00 - 12:30 ***LLMs and Applications***

11:00-11:10 *Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis*
Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Bin Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun and Imaan Alkhanen

11:10-11:20 *TARJAMAT: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties*
Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi and Muhammad Abdul-Mageed

11:20-11:30 *Analyzing Multilingual Competency of LLMs in Multi-Turn Instruction Following: A Case Study of Arabic*
Sabri Boughorbel and Majd Hawasly

11:30-11:40 *Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction*
Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi and Muhammad Abdul-Mageed

11:40-11:50 *Octopus: A Multitask Model and Toolkit for Arabic Natural Language Generation*
AbdelRahim Elmadany, El Moatez Billah Nagoudi and Muhammad Abdul-Mageed

11:50-12:00 *AlGhafa Evaluation Benchmark for Arabic Language Models*
Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay and Badreddine Noune

12:00-12:10 *GARI: Graph Attention for Relative Isomorphism of Arabic Word Embeddings*
Muhammad Asif Ali, Maha Alshmrani, Jianbin Qin, Yan Hu and Di Wang

12:10-12:20 *Violet: A Vision-Language Model for Arabic Image Captioning with Gemini Decoder*
Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Alcoba Inciarte and Muhammad Abdul-Mageed

12:20-12:30 *ArTST: Arabic Text and Speech Transformer*
Hawau Olamide Toyin, Amirbek Djanibekov, Ajinkya Kulkarni and Hanan Aldarmaki

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

14:00 - 14:20 ***Arabic Core NLP***

14:00-14:10 *SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks*
Mustafa Jarrar, Sanad Malaysha, Tymaa Hasanain Hammouda and Mohammed Khalilia

| | |
|---|---|
| 14:10-14:20 | *CamelParser2.0: A State-of-the-Art Dependency Parser for Arabic*<br>Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue and Nizar Habash |
| 14:20 - 14:45 | ***Shared Tasks Overview*** |
| 14:20-14:25 | *ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text*<br>Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino and Abed Alhakim Freihat |
| 14:25-14:30 | *NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task*<br>Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor and Nizar Habash |
| 14:30-14:35 | *WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task*<br>Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad and Alaa' Omar |
| 14:35-14:40 | *KSAA-RD Shared Task: Arabic Reverse Dictionary*<br>Rawan Al-Matham, Waad Alshammari, Abdulrahman AlOsaimy, Sarah Alhumoud, Asma Al Wazrah, Afrah Altamimi, Halah Alharbi and Abdullah Alaifi |
| 14:40-14:45 | *Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an*<br>Rana Malhas, Watheq Mansour and Tamer Elsayed |
| 14:45 - 15:30 | ***Panel on Arabic LLMs: Challenges and Opportunities (Areeb Alowisheq, Kareem Darwish and Perslav Nakov, moderated by Mona Diab)*** - Chair: Mona Diab |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 17:30 | ***Main Conference Posters (in-Person)*** |
| | *Enhancing Arabic Machine Translation for E-commerce Product Information: Data Quality Challenges and Innovative Selection Approaches*<br>Bryan Zhang, Salah Danial and Stephan Walter |
| | *Multi-Parallel Corpus of North Levantine Arabic*<br>Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek and Pavel Pecina |
| | *VoxArabica: A Robust Dialect-Aware Arabic Speech Recognition System*<br>Abdul Waheed, Bashar Talafha, Peter Sullivan, AbdelRahim Elmadany and Muhammad Abdul-Mageed |
| | *Arabic Topic Classification in the Generative and AutoML Era*<br>Doha Albared, Hadi Hamoud and Fadi A. Zaraket |
| | *Yet Another Model for Arabic Dialect Identification*<br>Ajinkya Kulkarni and Hanan Aldarmaki |
| 16:00 - 17:30 | ***EMNLP Findings Posters (in-Person)*** |
| | *Filtered Semi-Markov CRF*<br>Urchade Zaratiana, Nadi Tomeh, Niama El Khbir, Pierre Holat and Thierry Charnois |
| | *Automatic Pronunciation Assessment - A Review*<br>Yassine El Kheir, Ahmed Ali and Shammur Absar Chowdhury |
| | *Data Augmentation Techniques for Machine Translation of Code-Switched Texts: A Comparative Study*<br>Injy Hamed, Nizar Habash and Thang Vu |
| 16:00 - 17:30 | ***Shared Task Posters (in-Person)*** |
| | *LIPN at WojoodNER shared task: A Span-Based Approach for Flat and Nested Arabic Named Entity Recognition*<br>Niama Elkhbir, Urchade Zaratiana, Nadi Tomeh and Thierry Charnois |
| | *DetectiveRedasers at ArAIEval Shared Task: Leveraging Transformer Ensembles for Arabic Deception Detection*<br>Bryan E. Tuck, Fatima Zahra Qachfar, Dainis Boumber and Rakesh M. Verma |

*Nexus at ArAIEval Shared Task: Fine-Tuning Arabic Language Models for Propaganda and Disinformation Detection*
Yunze Xiao and Firoj Alam

*PTUK-HULAT at ArAIEval Shared Task Fine-tuned Distilbert to Predict Disinformative Tweets*
Areej Jaber and Paloma Martinez

*ReDASPersuasion at ArAIEval Shared Task: Multilingual and Monolingual Models For Arabic Persuasion Detection*
Fatima Zahra Qachfar and Rakesh M. Verma

*UL & UM6P at ArAIEval Shared Task: Transformer-based model for Persuasion Techniques and Disinformation detection in Arabic*
Salima Lamsiyah, Abdelkader El Mahdaouy, Hamza Alami, Ismail Berrada and Christoph Schommer

*UWB at Arabic Reverse Dictionary shared task: Computing the meaning of a gloss*
Stephen Taylor

*NLPeople at NADI 2023 Shared Task: Arabic Dialect Identification with Augmented Context and Multi-Stage Tuning*
Mohab Elkaref, Movina Moses, Shinnosuke Tanaka, James Barry and Geeth De Mel

*SANA at NADI 2023 shared task: Ensemble of Layer-Wise BERT-based models for Dialectal Arabic Identification*
Nada Almarwani and Samah Aloufi

*The Helsinki-NLP Submissions at NADI 2023 Shared Task: Walking the Baseline*
Yves Scherrer, Aleksandra Miletić and Olli Kuparinen

*LKAU23 at Qur'an QA 2023: Using Transformer Models for Retrieving Passages and Finding Answers to Questions from the Qur'an*
Sarah Alnefaie, Abdullah N. Alsaleh, Eric Atwell, Mohammad Alsalka and Abdulrahman Altahhan

*El-Kawaref at WojoodNER shared task: StagedNER for Arabic Named Entity Recognition*
Nehal Elkaref and Mohab Elkaref

*UM6P & UL at WojoodNER shared task: Improving Multi-Task Learning for Flat and Nested Arabic Named Entity Recognition*
Abdelkader El Mahdaouy, Salima Lamsiyah, Hamza Alami, Christoph Schommer and Ismail Berrada

16:00 - 17:30 ***Main Conference Posters (Virtual)***

*IDRISI-D: Arabic and English Datasets and Benchmarks for Location Mention Disambiguation over Disaster Microblogs*
Reem Suwaileh, Tamer Elsayed and Muhammad Imran

*HICMA: The Handwriting Identification for Calligraphy and Manuscripts in Arabic Dataset*
Anis Ismail, Zena Kamel and Reem Mahmoud

*Simplify: Automatic Arabic Sentence Simplification using Word Embeddings*
Yousef SalahEldin and Caroline Sabty

*ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic*
Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden and Sana Ghanem

*ArSarcasMoji Dataset: The Emoji Sentiment Roles in Arabic Ironic Contexts*
Shatha Ali A. Hakami, Robert Hendley and Phillip Smith

*Machine Translation of Omani Arabic Dialect from Social Media*
Khoula Al-Kharusi and Abdurahman Khalifa AAlAbdulsalam

*Automated De-Identification of Arabic Medical Records*
Veysel Kocaman, Youssef MELLAH, Hasham Ul Haq and David Talby

*Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus*
Helene Bøsei Olsen, Samia Touileb and Erik Velldal

*Investigating Zero-shot Cross-lingual Language Understanding for Arabic*
Zaid Alyafeai and Moataz Ahmed

*On Enhancing Fine-Tuning for Pre-trained Language Models*
Abir Betka, Zeyd Ferhat, Riyadh Barka, Selma Boutiba, Zineddine S. Kahhoul, Tiar M. Lakhdar, Ahmed Abdelali and Habiba Dahmani

*LANS: Large-scale Arabic News Summarization Corpus*
Abdulaziz Alhamadani, Xuchao Zhang, Jianfeng He, Aadyant Khatri and Chang-Tien Lu

*Performance Implications of Using Unrepresentative Corpora in Arabic Natural Language Processing*
Saied Alshahrani, Norah Alshahrani, Soumyabrata Dey and Jeanna Matthews

*ArTrivia: Harvesting Arabic Wikipedia to Build A New Arabic Question Answering Dataset*
Sultan Alrowili and K Vijay-Shanker

*Aswat: Arabic Audio Dataset for Automatic Speech Recognition Using Speech-Representation Learning*
Lamya Alkanhal, Abeer Alessa, Elaf Almahmoud and Rana Alaqil

*Offensive Language Detection in Arabizi*
Imene Bensalem, Meryem Ait Mout and Paolo Rosso

16:00 - 17:30 ***EMNLP Findings Posters (Virtual)***

*Arabic Mini-ClimateGPT : A Climate Change and Sustainability Tailored Arabic LLM*
Sahal Shaji Mullappilly, Abdelrahman M Shaker, Omkar Chakradhar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan and Fahad Khan

*Dolphin: A Challenging and Diverse Benchmark for Arabic NLG*
El Moatez Billah Nagoudi, AbdelRahim A. Elmadany, Ahmed Oumar El-Shangiti and Muhammad Abdul-Mageed

16:00 - 17:30 ***Shared Task Posters (Virtual)***

*AAST-NLP at ArAIEval Shared Task: Tackling Persuasion technique and Disinformation Detection using Pre-Trained Language Models On Imbalanced Datasets*
Ahmed El-Sayed, Omar Nasr and Noureldin Elmadany

*AraDetector at ArAIEval Shared Task: An Ensemble of Arabic-specific pre-trained BERT and GPT-4 for Arabic Disinformation Detection*
Ahmed Bahaaulddin, Vian Sabeeh, Hanan M. Belhaj, Serry Sibaee, Samar Ahmad, Ibrahim Khurfan and Abdullah I. Alharbi

*Frank at ArAIEval Shared Task: Arabic Persuasion and Disinformation: The Power of Pre-trained Models*
Dilshod Azizov, Jiyong Li and Shangsong Liang

*HTE at ArAIEval Shared Task: Integrating Content Type Information in Binary Persuasive Technique Detection*
Khaldi Hadjer and Taqiy Eddine Bouklouha

*Itri Amigos at ArAIEval Shared Task: Transformer vs. Compression-Based Models for Persuasion Techniques and Disinformation Detection*
Jehad Oumer, Nouman Ahmed and Natalia Flechas Manrique

*KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection in Arabic using Similar and Contrastive Representation Alignment*
Hariram Veeramani, Surendrabikram Thapa and Usman Naseem

*Legend at ArAIEval Shared Task: Persuasion Technique Detection using a Language-Agnostic Text Representation Model*
Olumide E. Ojo, Olaronke O. Adebanji, Hiram Calvo, Damian O. Dieke, Olumuyiwa E. OJO, Seye E. Akinsanya, Tolulope O. Abiola and Anna Feldman

*Mavericks at ArAIEval Shared Task: Towards a Safer Digital Space - Transformer Ensemble Models Tackling Deception and Persuasion*
Sudeep Mangalvedhekar, Kshitij Deshpande, Yash Patwardhan, Vedant Deshpande and Ravindra Murumkar

*PD-AR at ArAIEval Shared Task: A BERT-Centric Approach to Tackle Arabic Disinformation*
Pritam Deka and Ashwathy T. Revi

*Raphael at ArAIEval Shared Task: Understanding Persuasive Language and Tone, an LLM Approach*
Utsav Shukla, Manan Vyas and Shailendra Tiwari

*USTHB at NADI 2023 shared task: Exploring Preprocessing and Feature Engineering Strategies for Arabic Dialect Identification*
Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache and Rachida Djeradi

*rematchka at ArAIEval Shared Task: Prefix-Tuning & Prompt-tuning for Improved Detection of Propaganda and Disinformation in Arabic Social Media Content*
Reem Abdel-Salam

*Abed at KSAA-RD Shared Task: Enhancing Arabic Word Embedding with Modified BERT Multilingual*
Abdelrahim Qaddoumi

*Qamosy at Arabic Reverse Dictionary shared task: Semi Decoder Architecture for Reverse Dictionary with SBERT Encoder*
Serry Sibaee, Samar Ahmad, Ibrahim Khurfan, Vian Sabeeh, Ahmed Bahaaulddin, Hanan M. Belhaj and Abdullah I. Alharbi

*Rosetta Stone at KSAA-RD Shared Task: A Hop From Language Modeling To Word–Definition Alignment*
Ahmed Elbakry, Mohamed Gabr, Muhammad ElNokrashy and Badr AlKhamissi

*ANLP-RG at NADI 2023 shared task: Machine Translation of Arabic Dialects: A Comparative Study of Transformer Models*
Wiem Derouich, Sameh Kchaou and Rahma Boujelbane

*DialectNLU at NADI 2023 Shared Task: Transformer Based Multitask Approach Jointly Integrating Dialect and Machine Translation Tasks in Arabic*
Hariram Veeramani, Surendrabikram Thapa and Usman Naseem

*Frank at NADI 2023 Shared Task: Trio-Based Ensemble Approach for Arabic Dialect Identification*
Dilshod Azizov, Jiyong Li and Shangsong Liang

*ISL-AAST at NADI 2023 shared task: Enhancing Arabic Dialect Identification in the Era of Globalization and Technological Progress*
Shorouk Adel and Noureldin Elmadany

*IUNADI at NADI 2023 shared task: Country-level Arabic Dialect Classification in Tweets for the Shared Task NADI 2023*
Yash A. Hatekar and Muhammad S. Abdo

*Mavericks at NADI 2023 Shared Task: Unravelling Regional Nuances through Dialect Identification using Transformer-based Approach*
Vedant Deshpande, Yash Patwardhan, Kshitij Deshpande, Sudeep Mangalvedhekar and Ravindra Murumkar

*USTHB at NADI 2023 shared task: Exploring Preprocessing and Feature Engineering Strategies for Arabic Dialect Identification*
Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache and Rachida Djeradi

*UniManc at NADI 2023 Shared Task: A Comparison of Various T5-based Models for Translating Arabic Dialectical Text to Modern Standard Arabic*
Abdullah Khered, Ingy Yasser Abdelhalim, Nadine Abdelhalim, Ahmed Soliman and Riza Batista-Navarro

*UoT at NADI 2023 shared task: Automatic Arabic Dialect Identification is Made Possible*
Abduslam F A Nwesri, Nabila A S Shinbir and Hassan A H Ebrahem

*rematchka at NADI 2023 shared task: Parameter Efficient tuning for Dialect Identification and Dialect Machine Translation*
Reem Abdel-Salam

*AHJL at Qur'an QA 2023 Shared Task: Enhancing Passage Retrieval using Sentence Transformer and Translation*
Hessa A. Alawwad, Lujain A. Alawwad, Jamilah Alharbi and Abdullah I. Alharbi

*Al-Jawaab at Qur'an QA 2023 Shared Task: Exploring Embeddings and GPT Models for Passage Retrieval and Reading Comprehension*
Abdulrezzak Zekiye and Fadi Amroush

*GYM at Qur'an QA 2023 Shared Task: Multi-Task Transfer Learning for Quranic Passage Retrieval and Question Answering with Large Language Models*
Ghazaleh Mahmoudi, Yeganeh Morshedzadeh and Sauleh Eetemadi

*LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic*
Hariram Veeramani, Surendrabikram Thapa and Usman Naseem

*Alex-U 2023 NLP at WojoodNER shared task: AraBINDER (Bi-encoder for Arabic Named Entity Recognition)*
Mariam Hussein, Sarah Khaled, Marwan Torki and Nagwa El-Makky

*AlexU-AIC at WojoodNER shared task: Sequence Labeling vs MRC and SWA for Arabic Named Entity Recognition*
Shereen Elkordi, Noha Adly and Marwan Torki

*AlphaBrains at WojoodNER shared task: Arabic Named Entity Recognition by Using Character-based Context-Sensitive Word Representations*
Toqeer Ehsan, Amjad Ali and Ala Al-Fuqaha

*ELYADATA at WojoodNER Shared Task: Data and Model-centric Approaches for Arabic Flat and Nested NER*
Imen Laouirine, Haroun Elleuch and Fethi Bougares

*Lotus at WojoodNER Shared Task: Multilingual Transformers: Unveiling Flat and Nested Entity Recognition*
Jiyong Li, Dilshod Azizov, Hilal AlQuabeh and Shangsong Liang

*TCE at Qur'an QA 2023 Shared Task: Low Resource Enhanced Transformer-based Ensemble Approach for Qur'anic QA*
Mohammed Alaa Elkomy and Amany Sarhan

# W18 - The Third Workshop on Multi-lingual Representation Learning (MRL)

**Organizers:**
David Ifeoluwa Adelani, Duygu Ataman, Chris Emezue, Omer Goldman, Hila Gonen, Mammad Hajili, Benjamin Muller, Sebastian Ruder, Gözde Gül Şahin, Francesco Tinner, Genta Indra Winata

https://sigtyp.github.io/ws2023-mrl.html
Venue: Leo 3 & 4
**Thursday, December 7, 2023**

Abstract.

| | |
|---|---|
| 09:00 - 09:10 | ***Opening Remarks*** |
| 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*** |
| 14:30 - 15:30 | ***Best Paper Award Session*** |
| 15:30 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:50 | ***Afternoon Oral Session*** |
| 16:50 - 17:00 | ***Closing Remarks*** |
| 09:10 - 10:30 | ***Morning Oral Session*** |

# W19 - Novel Ideas in Learning to Learn through Interaction (NILLI)

**Organizers:**
Prasanna Parthasarathi, Koustuv Sinha, Khyathi Raghavi Chandu, Chinnadhurai Sankar, Adina Williams, Sarath Chandar, Marc-Alexandre Côté, Joelle Pineau

https://www.cs.mcgill.ca/~pparth2/nilli_workshop_2023

Venue: Leo 1 & 2

**Thursday, December 7, 2023**

Collaborative dialogues with automated systems through language interactions have become ubiquitous, wherein it is becoming common from setting an alarm to planning one's day through language interactions. With recent advances in dialogue research, embodied learning and using language as a mode of instruction for learning agents there is, now, a scope for realizing domains that can assume agents with primitive task knowledge and a continual interact-and-learn procedure to systematically acquire knowledge through verbal/non-verbal interactions. The research direction of building interactive learning agents facilitates the possibility of agents to have advanced interactions like taking instructions by being a pragmatic listener, asking for more samples, generating rationales for predictions, interactions to interpret learning dynamics, or even identifying or modifying a new task that can be used towards building effective learning-to-learn mechanisms. In a way, with verbal/non-verbal interactive medium this interdisciplinary field unifies research paradigms of lifelong learning, natural language processing, embodied learning, reinforcement learning, robot learning and multi-modal learning towards building interactive and interpretable AI.

| | |
|---|---|
| 08:30 - 09:08 | ***Opening Remarks*** |
| 08:35 - 09:20 | ***Invited Talk 1*** |
| 09:20 - 10:05 | ***Invited Talk 2*** |
| 10:05 - 10:50 | ***Invited Talk 3*** |
| 10:50 - 11:15 | ***Coffee Break*** |
| 11:15 - 12:00 | ***Invited Talk 4*** |
| 12:00 - 13:00 | ***Lunch Break*** |
| 13:00 - 15:30 | ***Lightning Talks (Session 1 - 18 talks)*** |
| 15:30 - 16:15 | ***Invited Talk 5*** |
| 16:15 - 18:00 | ***Lightning Talks (Session 2 - 12 talks)*** |
| 18:00 - 18:05 | ***Closing Remarks*** |

# W20 - The First Bangla Language Processing Workshop (BLP)

**Organizers:**
Firoj Alam, Sudipta Kar, Shammur Absar Chowdhury, Farig Sadeque, Ruhul Amin

https://blp-workshop.github.io/
Venue: Pisces 2& 3
**Thursday, December 7, 2023**

Bangla - a member of the Indo-Aryan language family, is ranked as the 6th most widely spoken language across the world, with 230 million native speakers from Bangladesh and India. This morphologically rich language has a long-standing literacy tradition, with diverse dialects and language dependent challenges. Bangla, with three decade of research history is still considered a low-resource language in the natural language processing (NLP) and speech community mainly due to the limited and scattered research efforts by individual researchers. These line of sparse works are not highly visible to the international research community. Therefore, this workshop aims to provide a forum for researchers to share and discuss their ongoing work with the international community. Following the success of prior local editions of the conferences in 2018 and 2019, in this first edition of the workshop, we will focus on Bangla, which is a low-resource language, and assess its current state-of-the-art and discuss strategies to make further progress in both NLP, Speech and multimodal research. Through this workshop, we plan to bring researchers together to come up with frameworks and strategies that can later support to other low-resource languages. This workshop is timely given the continued rise in research projects focusing on low-resource and multilingual studies. We particularly encourage researchers to submit their papers focusing on novel methodologies and resources that help towards the progress of Bangla and other low-resource languages. Novel methodologies include, but are not limited to, zero-shot learning, unsupervised learning, and simple yet effective methods applicable to low-computation scenarios.

| | |
|---|---|
| 09:00 - 09:20 | ***Opening Remarks*** |
| 09:20 - 09:50 | ***Invited Talk 1: NLP in Mexican Spanish: A Path Through Shared Tasks*** |
| 09:50 - 10:26 | ***Oral Presentation I (long papers)*** |
| 09:50-10:02 | *BSpell: A CNN-Blended BERT Based Bangla Spell Checker*<br>Chowdhury Rafeed Rahman, MD.Hasibur Rahman, Samiha Zakir, Mohammad Rafsan and Mohammed Eunus Ali |
| 10:02-10:14 | *BLP-2023 Task 1: Violence Inciting Text Detection (VITD)*<br>Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed and Mohammad Ruhul Amin |
| 10:02-10:14 | *BLP-2023 Task 1: Violence Inciting Text Detection (VITD)*<br>Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed and Mohammad Ruhul Amin |
| 10:14-10:26 | *BLP-2023 Task 2: Sentiment Analysis*<br>Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das and Afiyat Anjum |
| 10:30 - 11:00 | ***Coffee Break*** |
| 11:00 - 12:00 | ***Poster Session*** |
| 12:00 - 13:00 | ***Lunch Break*** |

| | |
|---|---|
| 13:00 - 14:00 | ***Oral Presentation II (long papers)*** |
| 13:00-13:12 | *Low-Resource Text Style Transfer for Bangla: Data & Models*<br>Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr Ojha and Ondrej Dusek |
| 13:12-13:24 | *Vio-Lens: A Novel Dataset of Annotated Social Network Posts Leading to Different Forms of Communal Violence and its Evaluation*<br>Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahouti, Syed Ishtiaque Ahmed, Nabeel Mohammed and Mohammad Ruhul Amin |
| 13:24-13:36 | *Pseudo-Labeling for Domain-Agnostic Bangla Automatic Speech Recognition*<br>Rabindra Nath Nandi, Mehadi Hasan Menon, Tareq Al Muntasir, Sagor Sarker, Quazi Sarwar Muhtaseem, Md. Tariqul Islam, Shammur Absar Chowdhury and Firoj Alam |
| 13:36-13:48 | *Contextual Bangla Neural Stemmer: Finding Contextualized Root Word Representations for Bangla Words*<br>Md Fahim, Amin Ahsan Ali, M Ashraful Amin and Akmmahbubur Rahman |
| 13:48-14:00 | *Crosslingual Retrieval Augmented In-context Learning for Bangla*<br>Xiaoqian Li, Ercong Nie and Sheng Liang |
| 14:00 - 14:10 | ***Break*** |
| 14:10 - 15:05 | ***Oral Presentation III (long + short papers)*** |
| 14:10-14:22 | *Investigating the Effectiveness of Graph-based Algorithm for Bangla Text Classification*<br>Farhan Noor Dehan, Md Fahim, Amin Ahsan Ali, M Ashraful Amin and Akmmahbubur Rahman |
| 14:22-14:29 | *BanglaCHQ-Summ: An Abstractive Summarization Dataset for Medical Queries in Bangla Conversational Speech*<br>Alvi Aveen Khan, Fida Kamal, Mohammad Abrar Chowdhury, Tasnim Ahmed, Md Tahmid Rahman Laskar and Sabbir Ahmed |
| 14:29-14:36 | *Offensive Language Identification in Transliterated and Code-Mixed Bangla*<br>Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos and Marcos Zampieri |
| 14:36-14:43 | *Intent Detection and Slot Filling for Home Assistants: Dataset and Analysis for Bangla and Sylheti*<br>Fardin Ahsan Sakib, A H M Rezaul Karim, Saadat Hasan Khan and Md Mushfiqur Rahman |
| 14:43-14:50 | *Assessing Political Inclination of Bangla Language Models*<br>Surendrabikram Thapa, Ashwarya Maratha, Khan Md Hasib, Mehwish Nasim and Usman Naseem |
| 14:50-14:57 | *SynthNID: Synthetic Data to Improve End-to-end Bangla Document Key Information Extraction*<br>Syed Mostofa Monsur, Shariar Kabir and Sakib Chowdhury |
| 14:57-15:04 | *BEmoLexBERT: A Hybrid Model for Multilabel Textual Emotion Classification in Bangla by Combining Transformers with Lexicon Features*<br>Ahasan Kabir, Animesh Chandra Roy and Zaima Sartaj Taheri |
| 15:05 - 15:35 | ***Invited Talk 2: Towards Transforming the Landscape of Indian language Technology*** |
| 15:35 - 16:00 | ***Coffee Break*** |
| 16:00 - 16:36 | ***Oral Presentation IV (long papers)*** |
| 16:00-16:12 | *BaTEClaCor: A Novel Dataset for Bangla Text Error Classification and Correction*<br>Nabilah Tabassum Oshin, Syed Mohaiminul Hoque, Md Fahim, Amin Ahsan Ali, M Ashraful Amin and Akmmahbubur Rahman |
| 16:12-16:24 | *Advancing Bangla Punctuation Restoration by a Monolingual Transformer-Based Method and a Large-Scale Corpus*<br>Mehedi Hasan Bijoy, Mir Fatema Afroz Faria, Mahbub E Sobhani, Tanzid Ferdoush and Swakkhar Shatabda |
| 16:24-16:36 | *Pipeline Enabling Zero-shot Classification for Bangla Handwritten Grapheme*<br>Linsheng Guo, Md Habibur Rahman Sifat and Tashin Ahmed |

| | |
|---|---|
| 16:36 - 17:15 | ***Panel Discussion*** |
| 17:15 - 17:30 | ***Industry Talk*** |
| 17:30 - 17:45 | ***Awards and Ending Remarks*** |

# 11

## Main Conference: December 8 - 10, 2023

# Main Conference Program (Overview)

## Main Conference Program (Overview): Day 1

7:30 **Registration**

9:00 - 9:30 **Opening Session (Room: East & Central)**

9:30 - 10:30 **Keynote 1 - Speaker: Jong Park**

10:30 - 11:00 Coffee break (Room: West Foyer)

| | | | |
|---|---|---|---|
| 11:00 - 12:30 | **Session 2** | **Information Extraction 1**<br>**Oral**<br>*Room: East* | **Machine Translation**<br>**Oral**<br>*Room: West 1* |
| | | **Computational Social Science and Cultural Analytics**<br>**Oral**<br>*Room: West 2* | **Dialogue and Interactive Systems 1**<br>**Oral**<br>*Room: West 3* |
| | | **Demo session 1**<br>*Room: East Foyer* | **Poster session 1**<br>*Room: East Foyer* |
| | | **Findings 1**<br>*Room: East Foyer* | **Industry 1**<br>*Room: East Foyer* |

12:30 - 14:00 Lunch break

| | | | |
|---|---|---|---|
| 14:00 - 15:30 | **Session 3** | **Discourse and Pragmatics**<br>**Oral**<br>*Room: East* | **Commonsense Reasoning**<br>**Oral**<br>*Room: Central 1* |
| | | **Efficient Methods for NLP 1**<br>**Oral**<br>*Room: Central 3* | **Ethics in NLP**<br>**Oral**<br>*Room: West 1* |
| | | **Phonology, Morphology, and Word Segmentation**<br>**Oral**<br>*Room: West 2* | **Information Extraction 2**<br>**Oral**<br>*Room: West 3* |
| | | **Demo session 2**<br>*Room: East Foyer* | **Poster session 2**<br>*Room: East Foyer* |
| | | **Findings 2**<br>*Room: East Foyer* | **Industry 2**<br>*Room: East Foyer* |
| | | **BOF & AGM**<br>*Room: Aquarius* | |

15:30 - 16:00 Coffee break (Room: West Foyer)

| | | | |
|---|---|---|---|
| 16:00 - 17:30 | **Session 4** | **Interpretability, Interactivity, and Analysis of Models for NLP 1**<br>**Oral**<br>*Room: East* | **Language Grounding to Vision, Robotics and Beyond**<br>**Oral**<br>*Room: Central 1* |
| | | **Language Modeling and Analysis of Language Models 1**<br>**Oral**<br>*Room: Central 3* | **Information Retrieval and Text Mining**<br>**Oral**<br>*Room: West 1* |
| | | **Linguistic Theories, Cognitive Modeling and Psycholinguistics**<br>**Oral**<br>*Room: West 2* | **Dialogue and Interactive Systems 2**<br>**Oral**<br>*Room: West 3* |
| | | **Demo session 3**<br>*Room: East Foyer* | **Poster session 3**<br>*Room: East Foyer* |
| | | **Findings 3**<br>*Room: East Foyer* | **Industry 3**<br>*Room: East Foyer* |
| | | **BOF & AGM**<br>*Room: Aquarius* | |

17:30 End of Day (One more BOF & AGM event at night time)

# Main Conference Program (Overview): Day 2

8:30 **Registration**

| | | | |
|---|---|---|---|
| 9:00 - 10:30 | **Session 5** | **Multilinguality and Linguistic Diversity 1**<br>**Oral**<br>*Room: East* | **Natural Language Generation 1**<br>**Oral**<br>*Room: Central 1* |
| | | **NLP Applications 1**<br>**Oral**<br>*Room: Central 3* | **Theme Track: Large Language Models and the Future of NLP 1**<br>**Oral**<br>*Room: West 1* |
| | | **Efficient Methods for NLP 2**<br>**Oral**<br>*Room: West 2* | **Human-Centered NLP**<br>**Oral**<br>*Room: West 3* |
| | | **Demo session 4**<br>*Room: East Foyer* | **Poster session 4**<br>*Room: East Foyer* |
| | | **Findings 4**<br>*Room: East Foyer* | **Industry 4**<br>*Room: East Foyer* |
| | | **BOF & AGM**<br>*Room: Aquarius* | |

10:30 - 11:00 Coffee break (Room: West Foyer)

| | | | |
|---|---|---|---|
| 11:00 - 12:30 | **Session 6** | **Interpretability, Interactivity, and Analysis of Models for NLP 2**<br>**Oral**<br>*Room: East* | **Language Modeling and Analysis of Language Models 2**<br>**Oral**<br>*Room: Central 1* |
| | | **Multilinguality and Linguistic Diversity 2**<br>**Oral**<br>*Room: Central 3* | **Natural Language Generation 2**<br>**Oral**<br>*Room: West 1* |
| | | **Question Answering**<br>**Oral**<br>*Room: West 2* | **Resources and Evaluation 1**<br>**Oral**<br>*Room: West 3* |
| | | **Demo session 5**<br>*Room: East Foyer* | **Poster session 5**<br>*Room: East Foyer* |
| | | **Findings 5**<br>*Room: East Foyer* | **Industry 5**<br>*Room: East Foyer* |
| | | **BOF & AGM**<br>*Room: Aquarius* | |

12:30 - 13:45 Lunch break

13:45 - 14:30 **Plenary: Business Meeting: All Attendees Welcome**

13:45 - 14:30 **BOF & AGM Event**

14:30 - 15:30 **Keynote 2 - Speaker: Emily Mower Provost**

15:30 - 16:00 Coffee break (Room: West Foyer)

16:00 - 17:00 **Plenary: Panel Discussion**

18:30 - 20:30 **Social Event: Dinner**

20:00 - 23:45 **Social Event: Universal Studio Singapore**

# Main Conference Program (Overview): Day 3

8:30 **Registration**

9:00 - 10:30 **Session 9**

| | |
|---|---|
| **Semantics 1**<br>**Oral**<br>*Room: East* | **Sentiment/Stylistic Analysis**<br>**Oral**<br>*Room: Central 1* |
| **Speech & Multimodality 1**<br>**Oral**<br>*Room: Central 3* | **Summarization**<br>**Oral**<br>*Room: West 1* |
| **Machine Learning for NLP**<br>**Oral**<br>*Room: West 2* | **Syntax, Parsing and their Applications**<br>**Oral**<br>*Room: West 3* |
| **Demo session 6**<br>*Room: East Foyer* | **Poster session 6**<br>*Room: East Foyer* |
| **Findings 6**<br>*Room: East Foyer* | **Industry 6**<br>*Room: East Foyer* |
| **BOF & AGM**<br>*Room: Aquarius* | |

10:30 - 11:00 Coffee break (Room: West Foyer)

11:00 - 12:30 **Session 10**

| | |
|---|---|
| **NLP Applications 2**<br>**Oral**<br>*Room: East* | **Resources and Evaluation 2**<br>**Oral**<br>*Room: Central 1* |
| **Semantics 2**<br>**Oral**<br>*Room: Central 3* | **Speech & Multimodality 2**<br>**Oral**<br>*Room: West 1* |
| **Theme Track: Large Language Models and the Future of NLP 2**<br>**Oral**<br>*Room: West 2* | **Industry Track**<br>**Oral**<br>*Room: West 3* |
| **Demo session 7**<br>*Room: East Foyer* | **Findings 7**<br>*Room: East Foyer* |
| **Industry 7**<br>*Room: East Foyer* | **BOF & AGM**<br>*Room: Aquarius* |

12:30 - 14:00 Lunch break

14:00 - 15:00 **Keynote 3 - Speaker: Christopher D. Manning**

15:00 - 15:30 Coffee break (Room: West Foyer)

15:30 - 16:15 **Plenary - Best Paper Awards**

16:15 - 17:00 **Plenary - Closing Session**

# Main Conference: Friday, December 8, 2023

## Registration
07:30-17:00 - Location: Level B2

## Session 1: Plenary - Opening Session
09:00-09:30 - Location: East & Central

## Session 1: Plenary - Keynote Speaker: Jong Park
09:30-10:30 - Location: East & Central

## Coffee Break
10:30-11:00 - Location: West Foyer

## Session 2: Oral & Poster - 11:00-12:30

**Information Extraction 1**

11:00-12:30 (East)

11:00-11:15 (East)

**Cross-Document Event Coreference Resolution on Discourse Structure**
*Xinyu Chen, Sheng Xu, Peifeng Li and Qiaoming Zhu*
Cross-document event coreference resolution (CD-ECR) is a task of clustering event mentions across multiple documents that refer to the same real-world events. Previous studies usually model the CD-ECR task as a pairwise similarity comparison problem by using different event mention features, and consider the highly similar event mention pairs in the same cluster as coreferent. In general, most of them only consider the local context of event mentions and ignore their implicit global information, thus failing to capture the interactions of long-distance event mentions. To address the above issue, we regard discourse structure as global information to further improve CD-ECR. First, we use a discourse rhetorical structure constructor to construct tree structures to represent documents. Then, we obtain shortest dependency paths from the tree structures to represent interactions between event mention pairs. Finally, we feed the above information to a multi-layer perceptron to capture the similarities of event mention pairs for resolving coreferent events. Experimental results on the ECB+ dataset show that our proposed model outperforms several baselines and achieves the competitive performance with the start-of-the-art baselines.

11:15-11:30 (East)
**Anaphor Assisted Document-Level Relation Extraction**
*Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang and Yongyi Mao*
Document-level relation extraction (DocRE) involves identifying relations between entities distributed in multiple sentences within a document. Existing methods focus on building a heterogeneous document graph to model the internal structure of an entity and the external interaction between entities. However, there are two drawbacks in existing methods. On one hand, anaphor plays an important role in reasoning to identify relations between entities but is ignored by these methods. On the other hand, these methods achieve cross-sentence entity interactions implicitly by utilizing a document or sentences as intermediate nodes. Such an approach has difficulties in learning fine-grained interactions between entities across different sentences, resulting in sub-optimal performance. To address these issues, we propose an Anaphor-Assisted (AA) framework for DocRE tasks. Experimental results on the widely-used datasets demonstrate that our model achieves a new state-of-the-art performance.

11:30-11:45 (East)
**RAPL: A Relation-Aware Prototype Learning Approach for Few-Shot Document-Level Relation Extraction**
*Shiao Meng, Xuming Hu, Aiwei Liu, Shuang Li, Fukun Ma, Yawen Yang and Lijie Wen*
How to identify semantic relations among entities in a document when only a few labeled documents are available? Few-shot document-level relation extraction (FSDLRE) is crucial for addressing the pervasive data scarcity problem in real-world scenarios. Metric-based meta-learning is an effective framework widely adopted for FSDLRE, which constructs class prototypes for classification. However, existing works often struggle to obtain class prototypes with accurate relational semantics: 1) To build prototype for a target relation type, they aggregate the representations of all entity pairs holding that relation, while these entity pairs may also hold other relations, thus disturbing the prototype. 2) They use a set of generic NOTA (none-of-the-above) prototypes across all tasks, neglecting that the NOTA semantics differs in tasks with different target relation types. In this paper, we propose a relation-aware prototype learning method for FSDLRE to strengthen the relational semantics of prototype representations. By judiciously leveraging the relation descriptions and realistic NOTA instances as guidance, our method effectively refines the relation prototypes and generates task-specific NOTA prototypes. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches by average 2.61% $F_1$ across various settings of two FSDLRE benchmarks.

11:45-12:00 (East)
**Guideline Learning for In-Context Information Extraction**
*Chaoxu Pang, Yixuan Cao, Qiang Ding and Ping Luo*
Large language models (LLMs) can perform a new task by merely conditioning on task instructions and a few input-output examples, without optimizing any parameters. This is called In-Context Learning (ICL). In-context Information Extraction (IE) has recently garnered attention in the research community. However, the performance of In-context IE generally lags behind the state-of-the-art supervised expert models. We highlight a key reason for this shortfall: underspecified task description. The limited-length context struggles to thoroughly express the intricate IE task instructions and various edge cases, leading to misalignment in task comprehension with humans. In this paper, we propose a Guideline Learning (GL) framework for In-context IE which reflectively learns and follows guidelines. During the learning phrase, GL automatically synthesizes a set of guidelines based on a few error cases, and during inference, GL retrieves helpful guidelines for better ICL. Moreover, we propose a self-consistency-based active learning method to enhance the efficiency of GL. Experiments on event extraction and relation extraction show that GL can significantly improve the performance of in-context IE.

12:00-12:15 (East)
**Preserving Knowledge Invariance: Rethinking Robustness Evaluation of Open Information Extraction**
*Ji Qi, Chuchun Zhang, Xiaozhi Wang, Kaisheng Zeng, Jifan Yu, Jinxin Liu, Lei Hou, Juanzi Li and Xu Bin*
The robustness to distribution changes ensures that NLP models can be successfully applied in the realistic world, especially for information extraction tasks. However, most prior evaluation benchmarks have been devoted to validating pairwise matching correctness, ignoring the crucial validation of robustness. In this paper, we present the first benchmark that simulates the evaluation of open information extraction models in the real world, where the syntactic and expressive distributions under the same knowledge meaning may drift variously. We design and annotate a large-scale testbed in which each example is a knowledge-invariant clique that consists of sentences with structured knowledge of the same meaning but with different syntactic and expressive forms. By further elaborating the robustness metric, a model is judged to be robust if its performance is consistently accurate on the overall cliques. We perform experiments on typical models published in the last decade as well as a representative large language model, and the results show that the existing successful models exhibit a frustrating degradation, with a maximum drop of $23.43$ $F_1$ score. Our resources and code will be publicly available.

12:15-12:30 (East)
**Instruct and Extract: Instruction Tuning for On-Demand Information Extraction**
*Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji and Jiawei Han*
Large language models with instruction-following capabilities open the door to a wider group of users. However, when it comes to information extraction – a classic task in natural language processing – most task-specific systems cannot align well with long-tail ad hoc extraction use cases for non-expert users. To address this, we propose a novel paradigm, termed On-Demand Information Extraction, to fulfill the personalized demands of real-world users. Our task aims to follow the instructions to extract the desired content from the associated text and present it in a structured tabular format. The table headers can either be user-specified or inferred contextually by the model. To facilitate research in this emerging area, we present a benchmark named InstructIE, inclusive of both automatically generated training data, as well as the human-annotated test set. Building on InstructIE, we further develop an On-Demand Information Extractor, ODIE. Comprehensive evaluations on our benchmark reveal that ODIE substantially outperforms the existing open-source models of similar size.

## Machine Translation

11:00-12:30 (West 1)

---

11:00-11:15 (West 1)
**Multilingual Pixel Representations for Translation and Effective Cross-lingual Transfer**
*Elizabeth Salesky, Neha Verma, Philipp Koehn and Matt Post*
We introduce and demonstrate how to effectively train multilingual machine translation models with pixel representations. We experiment with two different data settings with a variety of language and script coverage, demonstrating improved performance compared to subword embeddings. We explore various properties of pixel representations such as parameter sharing within and across scripts to better understand where they lead to positive transfer. We observe that these properties not only enable seamless cross-lingual transfer to unseen scripts, but make pixel representations more data-efficient than alternatives such as vocabulary expansion. We hope this work contributes to more extensible multilingual models for all languages and scripts.

11:15-11:30 (West 1)
**Non-autoregressive Streaming Transformer for Simultaneous Translation**
*Zhengrui Ma, Shaolei Zhang, Shoutao Guo, Chenze Shao, Min Zhang and Yang Feng*
Simultaneous machine translation (SiMT) models are trained to strike a balance between latency and translation quality. However, training these models to achieve high quality while maintaining low latency often leads to a tendency for aggressive anticipation. We argue that such issue stems from the autoregressive architecture upon which most existing SiMT models are built. To address those issues, we propose non-autoregressive streaming Transformer (NAST) which comprises a unidirectional encoder and a non-autoregressive decoder with intra-chunk parallelism. We enable NAST to generate the blank token or repetitive tokens to adjust its READ/WRITE strategy flexibly, and train it to maximize the non-monotonic latent alignment with an alignment-based latency loss. Experiments on various SiMT benchmarks demonstrate that NAST outperforms previous strong autoregressive SiMT baselines.

11:30-11:45 (West 1)
**IMTLab: An Open-Source Platform for Building, Evaluating, and Diagnosing Interactive Machine Translation Systems**
*Xu Huang, Zhirui Zhang, Ruize Gao, Yichao Du, Lemao Liu, Guoping Huang, Shuming Shi, Jiajun Chen and Shujian Huang*
We present IMTLab, an open-source end-to-end interactive machine translation (IMT) system platform that enables researchers to quickly build IMT systems with state-of-the-art models, perform an end-to-end evaluation, and diagnose the weakness of systems. IMTLab treats the whole interactive translation process as a task-oriented dialogue with a human-in-the-loop setting, in which human interventions can be explicitly incorporated to produce high-quality, error-free translations. To this end, a general communication interface is designed to support the flexible IMT architectures and user policies. Based on the proposed design, we construct a simulated and real interactive environment to achieve end-to-end evaluation and leverage the framework to systematically evaluate previous IMT systems. Our simulated and manual experiments show that the prefix-constrained decoding approach still gains the lowest editing cost in the end-to-end evaluation, while BiTIIMT achieves comparable editing cost with a better interactive experience.

11:45-12:00 (West 1)
**Integrating Language Models into Direct Speech Translation: An Inference-Time Solution to Control Gender Inflection**

*Dennis Fucci, Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri and Luisa Bentivogli*
When translating words referring to the speaker, speech translation (ST) systems should not resort to default masculine generics nor rely on potentially misleading vocal traits. Rather, they should assign gender according to the speakers' preference. The existing solutions to do so, though effective, are hardly feasible in practice as they involve dedicated model re-training on gender-labeled ST data. To overcome these limitations, we propose the first inference-time solution to control speaker-related gender inflections in ST. Our approach partially replaces the (biased) internal language model (LM) implicitly learned by the ST decoder with gender-specific external LMs. Experiments on en→es/fr/it show that our solution outperforms the base models and the best training-time mitigation strategy by up to 31.0 and 1.6 points in gender accuracy, respectively, for feminine forms. The gains are even larger (up to 32.0 and 3.4) in the challenging condition where speakers' vocal traits conflict with their gender.

12:00-12:15 (West 1)
**Crossing the Threshold: Idiomatic Machine Translation through Retrieval Augmentation and Loss Weighting**
*Emmy Liu, Aditi Chaudhary and Graham Neubig*
Idioms are common in everyday language, but often pose a challenge to translators because their meanings do not follow from the meanings of their parts. Despite significant advances, machine translation systems still struggle to translate idiomatic expressions. We provide a simple characterization of idiomatic translation and related issues. This allows us to conduct a synthetic experiment revealing a tipping point at which transformer-based machine translation models correctly default to idiomatic translations. To expand multilingual resources, we compile a dataset of ∼4k natural sentences containing idiomatic expressions in French, Finnish, and Japanese. To improve translation of natural idioms, we introduce two straightforward yet effective techniques: the strategic upweighting of training loss on potentially idiomatic sentences, and using retrieval-augmented models. This not only improves the accuracy of a strong pretrained MT model on idiomatic sentences by up to 13% in absolute accuracy, but also holds potential benefits for non-idiomatic sentences.

12:15-12:30 (West 1)
**Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection**
*Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna Martindale and Marine Carpuat*
Neural sequence generation models are known to "hallucinate", by producing outputs that are unrelated to the source text. These hallucinations are potentially harmful, yet it remains unclear in what conditions they arise and how to mitigate their impact. In this work, we first identify internal model symptoms of hallucinations by analyzing the relative token contributions to the generation in contrastive hallucinated vs. non-hallucinated outputs generated via source perturbations. We then show that these symptoms are reliable indicators of natural hallucinations, by using them to design a lightweight hallucination detector which outperforms both model-free baselines and strong classifiers based on quality estimation or large pre-trained models on manually annotated English-Chinese and German-English translation test beds.

## Computational Social Science and Cultural Analytics

11:00-12:30 (West 2)

---

11:00-11:15 (West 2)
**Towards Interpretable Mental Health Analysis with Large Language Models**
*Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang and Sophia Ananiadou*
The latest large language models (LLMs) such as ChatGPT, exhibit strong capabilities in automated mental health analysis. However, existing relevant studies bear several limitations, including inadequate evaluations, lack of prompting strategies, and ignorance of exploring LLMs for explainability. To bridge these gaps, we comprehensively evaluate the mental health analysis and emotional reasoning ability of LLMs on 11 datasets across 5 tasks. We explore the effects of different prompting strategies with unsupervised and distantly supervised emotional information. Based on these prompts, we explore LLMs for interpretable mental health analysis by instructing them to generate explanations for each of their decisions. We convey strict human evaluations to assess the quality of the generated explanations, leading to a novel dataset with 163 human-assessed explanations. We benchmark existing automatic evaluation metrics on this dataset to guide future related works. According to the results, ChatGPT shows strong in-context learning ability but still has a significant gap with advanced task-specific methods. Careful prompt engineering with emotional cues and expert-written few-shot examples can also effectively improve performance on mental health analysis. In addition, ChatGPT generates explanations that approach human performance, showing its great potential in explainable mental health analysis.

11:15-11:30 (West 2)
**A Diachronic Analysis of Paradigm Shifts in NLP Research: When, How, and Why?**
*Aniket Pramanick, Yufang Hou, Saif M. Mohammad and Iryna Gurevych*
Understanding the fundamental concepts and trends in a scientific field is crucial for keeping abreast of its continuous advancement. In this study, we propose a systematic framework for analyzing the evolution of research topics in a scientific field using causal discovery and inference techniques. We define three variables to encompass diverse facets of the evolution of research topics within NLP and utilize a causal discovery algorithm to unveil the causal connections among these variables using observational data. Subsequently, we leverage this structure to measure the intensity of these relationships. By conducting extensive experiments on the ACL Anthology corpus, we demonstrate that our framework effectively uncovers evolutionary trends and the underlying causes for a wide range of NLP research topics. Specifically, we show that tasks and methods are primary drivers of research in NLP, with datasets following, while metrics have minimal impact.

11:30-11:45 (West 2)
**The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages**
*Chiyu Zhang, Khai Duy Doan, Qisheng Liao and Muhammad Abdul-Mageed*
Instruction tuned large language models (LLMs), such as ChatGPT, demonstrate remarkable performance in a wide range of tasks. Despite numerous recent studies that examine the performance of instruction-tuned LLMs on various NLP benchmarks, there remains a lack of comprehensive investigation into their ability to understand cross-lingual sociopragmatic meaning (SM), i.e., meaning embedded within social and interactive contexts. This deficiency arises partly from SM not being adequately represented in any of the existing benchmarks. To address this gap, we present SPARROW, an extensive multilingual benchmark specifically designed for SM understanding. SPARROW comprises 169 datasets covering 13 task types across six primary categories (e.g., anti-social language detection, emotion recognition). SPARROW datasets encompass 64 different languages originating from 12 language families representing 16 writing scripts. We evaluate the performance of various multilingual pretrained language models (e.g., mT5) and instruction-tuned LLMs (e.g., BLOOMZ, ChatGPT) on SPARROW through fine-tuning, zero-shot, and/or few-shot learning. Our comprehensive analysis reveals that existing open-source instruction tuned LLMs still struggle to understand SM across various languages, performing close to a random baseline in some cases. We also find that although ChatGPT outperforms many LLMs, it still falls behind task-specific finetuned models with a gap of 12.19 SPARROW score. Our benchmark is available at: https://github.com/UBC-NLP/SPARROW

11:45-12:00 (West 2)
**Rumor Detection on Social Media with Crowd Intelligence and ChatGPT-Assisted Networks**
*Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao and Jiaming Zhao*
In the era of widespread dissemination through social media, the task of rumor detection plays a pivotal role in establishing a trustworthy and reliable information environment. Nonetheless, existing research on rumor detection confronts several challenges: the limited expressive power of text encoding sequences, difficulties in domain knowledge coverage and effective information extraction with knowledge graph-based methods, and insufficient mining of semantic structural information. To address these issues, we propose a Crowd Intelligence and ChatGPT-Assisted Network(CICAN) for rumor classification. Specifically, we present a crowd intelligence-based semantic feature learning module to capture textual content's sequential and hierarchical features. Then, we design a knowledge-based semantic structural mining module that leverages ChatGPT for knowledge enhancement. Finally, we construct an entity-sentence heterogeneous graph and design Entity-Aware Heterogeneous Attention to effectively integrate diverse structural information meta-paths. Experimental results demonstrate that CICAN achieves performance improvement in rumor detection tasks, validating the effectiveness and rationality of using large language models as auxiliary tools.

12:00-12:15 (West 2)
**Cultural Concept Adaptation on Multimodal Reasoning**
*Zhi Li and Yin Zhang*
Developing cultural adaptation methods is important, which can improve the model performance on the low-resource ones and provide more equitable opportunities for everyone to benefit from advanced technology. Past methods primarily focused on multilingual and multimodal capabilities, and the improvement of multicultural competence is still an unexplored problem. This is largely due to the difficulty of data scarcity and expensive annotation. In this paper, we navigate this uncharted territory by leveraging high-resource cultures to facilitate comprehension of low-resource ones. We first introduce an annotation-free method for cultural-concept adaptation and construct a concept mapping set. To facilitate the model's comprehension of cultural-concept mappings, we propose a new multimodal data augmentation called CultureMixup. This approach employs a three-tier code-switching strategy on textual sentences. Additionally, it uses a cultural concept-based mixup method for the images. This combination effectively generates new data instances across culture, phrase, word, and image levels. For visually grounded reasoning across languages and cultures, experimental results on five languages show that our method consistently improves performance for four existing multilingual and multimodal models on both zero-shot and few-shot settings.

12:15-12:30 (West 2)
**Reformulating NLP tasks to Capture Longitudinal Manifestation of Language Disorders in People with Dementia.**
*Dimitris Gkoumas, Matthew Purver and Maria Liakata*
Dementia is associated with language disorders which impede communication. Here, we automatically learn linguistic disorder patterns by making use of a moderately-sized pre-trained language model and forcing it to focus on reformulated natural language processing (NLP) tasks and associated linguistic patterns. Our experiments show that NLP tasks that encapsulate contextual information and enhance the gradient signal with linguistic patterns benefit performance. We then use the probability estimates from the best model to construct digital linguistic markers measuring the overall quality in communication and the intensity of a variety of language disorders. We investigate how the digital markers characterize dementia speech from a longitudinal perspective. We find that our proposed communication marker is able to robustly and reliably characterize the language of people with dementia, outperforming existing linguistic approaches; and shows external validity via significant correlation with clinical markers of behaviour. Finally, our proposed linguistic disorder markers provide useful insights into gradual language impairment associated with disease progression.

## Dialogue and Interactive Systems 1

11:00-12:30 (West 3)

---

11:00-11:15 (West 3)
**Learning Retrieval Augmentation for Personalized Dialogue Generation**
*Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang and Lilian Tang*
Personalized dialogue generation, focusing on generating highly tailored responses by leveraging persona profiles and dialogue context, has gained significant attention in conversational AI applications. However, persona profiles, a prevalent setting in current personalized dialogue datasets, typically composed of merely four to five sentences, may not offer comprehensive descriptions of the persona about the agent, posing a challenge to generate truly personalized dialogues. To handle this problem, we propose **L**earning Retrieval **A**ugmentation for **P**ersonalized **D**ial**O**gue **G**eneration (**LAPDOG**), which studies the potential of leveraging external knowledge for persona dialogue generation. Specifically, the proposed LAPDOG model consists of a story retriever and a dialogue generator. The story retriever uses a given persona profile as queries to retrieve relevant information from the story document, which serves as a supplementary context to augment the persona profile. The dialogue generator utilizes both the dialogue history and the augmented persona profile to generate personalized responses. For optimization, we adopt a joint training framework that collaboratively learns the story retriever and dialogue generator, where the story retriever is optimized towards desired ultimate metrics (e.g., BLEU) to retrieve content for the dialogue generator to generate personalized responses. Experiments conducted on the CONVAI2 dataset with ROCStory as a supplementary data source show that the proposed LAPDOG method substantially outperforms the baselines, indicating the effectiveness of the proposed method. The LAPDOG model code is publicly available for further exploration.

11:15-11:30 (West 3)
**Prompt-Based Monte-Carlo Tree Search for Goal-oriented Dialogue Policy Planning**
*Xiao Yu, Maximillian Chen and Zhou Yu*
Planning for goal-oriented dialogue often requires simulating future dialogue interactions and estimating task progress. Many approaches thus consider training neural networks to perform look-ahead search algorithms such as A* search and Monte Carlo Tree Search (MCTS). However, this training often require abundant annotated data, which creates challenges when faced with noisy annotations or low-resource settings. We introduce GDP-Zero, an approach using Open-Loop MCTS to perform goal-oriented dialogue policy planning without any model training. GDP-Zero prompts a large language model to act as a policy prior, value function, user simulator, and system model during the tree search. We evaluate GDP-Zero on the goal-oriented task PersuasionForGood, and find that its responses are preferred over ChatGPT up to 59.32% of the time, and are rated more persuasive than ChatGPT during interactive evaluations.

11:30-11:45 (West 3)
**MADNet: Maximizing Addressee Deduction Expectation for Multi-Party Conversation Generation**
*Jia-Chen Gu, Chao-Hong Tan, Caiyuan Chu, Zhen-Hua Ling, Chongyang Tao, Quan Liu and Cong Liu*

Modeling multi-party conversations (MPCs) with graph neural networks has been proven effective at capturing complicated and graphical information flows. However, existing methods rely heavily on the necessary addressee labels and can only be applied to an ideal setting where each utterance must be tagged with an "@" or other equivalent addressee label. To study the scarcity of addressee labels which is a common issue in MPCs, we propose MADNet that maximizes addressee deduction expectation in heterogeneous graph neural networks for MPC generation. Given an MPC with a few addressee labels missing, existing methods fail to build a consecutively connected conversation graph, but only a few separate conversation fragments instead. To ensure message passing between these conversation fragments, four additional types of latent edges are designed to complete a fully-connected graph. Besides, to optimize the edge-type-dependent message passing for those utterances without addressee labels, an Expectation-Maximization-based method that iteratively generates silver addressee labels (E step), and optimizes the quality of generated responses (M step), is designed. Experimental results on two Ubuntu IRC channel benchmarks show that MADNet outperforms various baseline models on the task of MPC generation, especially under the more common and challenging setting where part of addressee labels are missing.

11:45-12:00 (West 3)
**SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization**
*Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap and Yejin Choi*
Data scarcity has been a long standing issue in the field of open-domain social dialogue. To quench this thirst, we present SODA: the first publicly available, million-scale high-quality social dialogue dataset. By contextualizing social commonsense knowledge from a knowledge graph, we are able to distill an exceptionally broad spectrum of social interactions from a large language model. Human evaluation shows that conversations in SODA are more consistent, specific, and (surprisingly) natural than those in prior human-authored datasets. Using SODA, we train COSMO: a generalizable conversation model that is significantly more natural and consistent on unseen datasets than best-performing conversation models (e.g., GODEL, BlenderBot-1, Koala, Vicuna). Experiments reveal COSMO is sometimes even preferred to the original human-written gold responses. Additionally, our results shed light on the distinction between knowledge-enriched conversations and natural social chitchats. We plan to make our data, model, and code public.

12:00-12:15 (West 3)
**Conversation Chronicles: Towards Diverse Temporal and Relational Dynamics in Multi-Session Conversations**
*Jihyoung Jang, Minseong Boo and Hyounghun Kim*
In the field of natural language processing, open-domain chatbots have emerged as an important research topic. However, a major limitation of existing open-domain chatbot research is its singular focus on short single-session dialogue, neglecting the potential need for understanding contextual information in multiple consecutive sessions that precede an ongoing dialogue. Among the elements that compose the context in multi-session conversation settings, the time intervals between sessions and the relationships between speakers would be particularly important. Despite their importance, current research efforts have not sufficiently addressed these dialogical components. In this paper, we introduce a new 1M multi-session dialogue dataset, called Conversation Chronicles, for implementing a long-term conversation setup in which time intervals and fine-grained speaker relationships are incorporated. Following recent works, we exploit a large language model to produce the data. The extensive human evaluation shows that dialogue episodes in Conversation Chronicles reflect those properties while maintaining coherent and consistent interactions across all the sessions. We also propose a dialogue model, called ReBot, which consists of chronological summarization and dialogue generation modules using only around 630M parameters. When trained on Conversation Chronicles, ReBot demonstrates long-term context understanding with a high human engagement score.

12:15-12:30 (West 3)
**Enhancing Code-Switching for Cross-lingual SLU: A Unified View of Semantic and Grammatical Coherence**
*Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen and Yuexian Zou*
Despite the success of spoken language understanding (SLU) in high-resource languages, achieving similar performance in low-resource settings, such as zero-shot scenarios, remains challenging due to limited labeled training data. To improve zero-shot cross-lingual SLU, recent studies have explored code-switched sentences containing tokens from multiple languages. However, vanilla code-switched sentences often lack semantic and grammatical coherence. We ascribe this lack to two issues: (1) randomly replacing code-switched tokens with equal probability and (2) disregarding token-level dependency within each language. To tackle these issues, in this paper, we propose a novel method termed SoGo, for zero-shot cross-lingual SLU. First, we use a saliency-based substitution approach to extract keywords as substitution options. Then, we introduce a novel token-level alignment strategy that considers the similarity between the context and the code-switched tokens, ensuring grammatical coherence in code-switched sentences. Extensive experiments and analyses demonstrate the superior performance of SoGo across nine languages on MultiATIS++.

## Demo session 1

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools**
*Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu and Markus Leippold*
In the face of climate change, are companies really taking substantial steps toward more sustainable operations? A comprehensive answer lies in the dense, information-rich landscape of corporate sustainability reports. However, the sheer volume and complexity of these reports make human analysis very costly. Therefore, only a few entities worldwide have the resources to analyze these reports at scale, which leads to a lack of transparency in sustainability reporting. Empowering stakeholders with LLM-based automatic analysis tools can be a promising way to democratize sustainability report analysis. However, developing such tools is challenging due to (1) the hallucination of LLMs and (2) the inefficiency of bringing domain experts into the AI development loop. In this paper, we introduce ChatReport, a novel LLM-based system to automate the analysis of corporate sustainability reports, addressing existing challenges by (1) making the answers traceable to reduce the harm of hallucination and (2) actively involving domain experts in the development loop. We make our methodology, annotated datasets, and generated analyses of 1015 reports publicly available. Video Introduction: https://www.youtube.com/watch?v=Q5AzaKzPE4M Github: https://github.com/EdisonNi-hku/chatreport Live web app: reports.chatclimate.ai

11:00-12:30 (East Foyer / Virtual)
**RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models**
*Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii and Jun Deguchi*
Retrieval-augmented large language models (R-LLMs) combine pre-trained large language models (LLMs) with information retrieval systems to improve the accuracy of factual question-answering. However, current libraries for building R-LLMs provide high-level abstractions

without sufficient transparency for evaluating and optimizing prompts within specific inference processes such as retrieval and generation. To address this gap, we present RaLLe, an open-source framework designed to facilitate the development, evaluation, and optimization of R-LLMs for knowledge-intensive tasks. With RaLLe, developers can easily develop and evaluate R-LLMs, improving hand-crafted prompts, assessing individual inference processes, and objectively measuring overall system performance quantitatively. By leveraging these features, developers can enhance the performance and accuracy of their R-LLMs in knowledge-intensive generation tasks.

11:00-12:30 (East Foyer / Virtual)
**VIST5: An Adaptive, Retrieval-Augmented Language Model for Visualization-oriented Dialog**
*Henrik Voigt, Nuno Carvalhais, Monique Meuschke, Markus Reichstein, Sina Zarrie and Kai Lawonn*
The advent of large language models has brought about new ways of interacting with data intuitively via natural language. In recent years, a variety of visualization systems have explored the use of natural language to create and modify visualizations through visualization-oriented dialog. However, the majority of these systems rely on tailored dialog agents to analyze domain-specific data and operate domain-specific visualization tools and libraries. This is a major challenge when trying to transfer functionalities between dialog interfaces of different visualization applications. To address this issue, we propose VIST5, a visualization-oriented dialog system that focuses on easy adaptability to an application domain as well as easy transferability of language-controllable visualization library functions between applications. Its architecture is based on a retrieval-augmented T5 language model that leverages few-shot learning capabilities to enable a rapid adaptation of the system.

11:00-12:30 (East Foyer / Virtual)
**H2O Open Ecosystem for State-of-the-art Large Language Models**
*Arno Candel, Jon McKinney, Philipp Singer, Pascal Pfeiffer, Maximilian Jeblick, Chun Ming Lee and Marcos Conde*
Large Language Models (LLMs) represent a revolution in AI. However, they also pose many significant risks, such as the presence of biased, private, copyrighted or harmful text. For this reason we need open, transparent and safe solutions. We introduce a complete open-source ecosystem for developing and testing LLMs. The goal of this project is to boost open alternatives to closed-source approaches. We release h2oGPT, a family of fine-tuned LLMs from 7 to 70 Billion parameters. We also introduce H2O LLM Studio, a framework and no-code GUI designed for efficient fine-tuning, evaluation, and deployment of LLMs using the most recent state-of-the-art techniques. Our code and models are licensed under fully permissive Apache 2.0 licenses. We believe open-source language models help to boost AI development and make it more accessible and trustworthy. Our demo is available at: https://gpt.h2o.ai/

11:00-12:30 (East Foyer / Virtual)
**Koala: An Index for Quantifying Overlaps with Pre-training Corpora**
*Thuy-Trang Vu, Xuanli He, Gholamreza Haffari and Ehsan Shareghi*
In very recent years more attention has been placed on probing the role of pre-training data in Large Language Models (LLMs) downstream behaviour. Despite the importance, there is no public tool that supports such analysis of pre-training corpora at large scale. To help research in this space, we launch Koala, a searchable index over large pre-training corpora using lossless compressed suffix arrays with highly efficient compression rate and search support. In its first release we index the public proportion of OPT 175B, GPT-3, GPT-Neo, GPT-Neo, LLaMA, BERT, ELECTRA, RoBERTA, XLNet pre-training corpora. Koala provides a framework to do forensic analysis on the current and future benchmarks as well as to assess the degree of memorization in the output from the LLMs. Koala is available for public use at https://koala-index.erc.monash.edu/.

11:00-12:30 (East Foyer / Virtual)
**Sudowoodo: A Chinese Lyric Imitation System with Source Lyrics**
*Yongzhu Chang, Rongsheng Zhang, Lin Jiang, Qihang Chen, Le Zhang and Jiashu Pu*
Lyrics generation is a well-known application in natural language generation research, with several previous studies focusing on generating accurate lyrics using precise control such as keywords, rhymes, etc. However, lyrics imitation, which involves writing new lyrics by imitating the style and content of the source lyrics, remains a challenging task due to the lack of a parallel corpus. In this paper, we introduce Sudowoodo, a Chinese lyrics imitation system that can generate new lyrics based on the text of source lyrics. To address the issue of lacking a parallel training corpus for lyrics imitation, we propose a novel framework to construct a parallel corpus based on a keyword-based lyrics model from source lyrics. Then the pairs *(new lyrics, source lyrics)* are used to train the lyrics imitation model. During the inference process, we utilize a post-processing module to filter and rank the generated lyrics, selecting the highest-quality ones. We incorporated audio information and aligned the lyrics with the audio to form the songs as a bonus. The human evaluation results show that our framework can perform better lyric imitation. Meanwhile, the *Sudowoodo* system and demo video of the system is available at Sudowoodo and https://youtu.be/u5BBT_j1L5M

11:00-12:30 (East Foyer / Virtual)
**ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format**
*Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic and Minlie Huang*
Task-oriented dialogue (TOD) systems function as digital assistants, guiding users through various tasks such as booking flights or finding restaurants. Existing toolkits for building TOD systems often fall short in delivering comprehensive arrays of data, model, and experimental environments with a user-friendly experience. We introduce ConvLab-3: a multifaceted dialogue system toolkit crafted to bridge this gap. Our unified data format simplifies the integration of diverse datasets and models, significantly reducing complexity and cost for studying generalization and transfer. Enhanced with robust reinforcement learning (RL) tools, featuring a streamlined training process, in-depth evaluation tools, and a selection of user simulators, ConvLab-3 supports the rapid development and evaluation of robust dialogue policies. Through an extensive study, we demonstrate the efficacy of transfer learning and RL and showcase that ConvLab-3 is not only a powerful tool for seasoned researchers but also an accessible platform for newcomers.

11:00-12:30 (East Foyer / Virtual)
**FLEEK: Factual Error Detection and Correction with Evidence Retrieved from External Knowledge**
*Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas and Yunyao Li*
Detecting factual errors of textual information, whether generated by large language models (LLM) or curated by humans, is crucial for making informed decisions. LLMs' inability to attribute their claims to external knowledge and their tendency to hallucinate makes it difficult to rely on their responses. Humans, too, are prone to factual errors in their writing. Since manual detection and correction of factual er- rors is labor-intensive, developing an automatic approach can greatly reduce human effort. We present a prototype tool that automatically extracts factual claims from text, gathers evidence from external knowledge sources, evaluates the factuality of each claim, and suggests revisions for identified errors using the collected evidence. Initial empirical evaluation on fact error detection (77-85% F1) shows the potential of our tool.

11:00-12:30 (East Foyer / Virtual)
**YATO: Yet Another deep learning based Text analysis Open toolkit**

*Zeqiang Wang, Yile Wang, Jiageng Wu, Zhiyang Teng and Jie Yang*
We introduce YATO, an open-source, easy-to-use toolkit for text analysis with deep learning. Different from existing heavily engineered toolkits and platforms, YATO is lightweight and user-friendly for researchers from cross-disciplinary areas. Designed in a hierarchical structure, YATO supports free combinations of three types of widely used features including 1) traditional neural networks (CNN, RNN, etc.); 2) pre-trained language models (BERT, RoBERTa, ELECTRA, etc.); and 3) user-customized neural features via a simple configurable file. Benefiting from the advantages of flexibility and ease of use, YATO can facilitate fast reproduction and refinement of state-of-the-art NLP models, and promote the cross-disciplinary applications of NLP techniques. The code, examples, and documentation are publicly available at https://github.com/jiesutd/YATO. A demo video is also available at https://www.youtube.com/playlist?list=PLJ0mhzMcRuDUlTkzBfAftOqiJRxYTTjXH.

11:00-12:30 (East Foyer / Virtual)
**INTELMO: Enhancing Models' Adoption of Interactive Interfaces**
*Chunxu Yang, Chien-Sheng Wu, Lidiya Murakhovs'ka, Philippe Laban and Xiang Anthony Chen*
This paper presents INTELMO, an easy-to-use library to help model developers adopt user-faced interactive interfaces and articles from real-time RSS sources for their language models. The library categorizes common NLP tasks and provides default style patterns, streamlining the process of creating interfaces with minimal code modifications while ensuring an intuitive user experience. Moreover, INTELMO employs a multi-granular hierarchical abstraction to provide developers with fine-grained and flexible control over user interfaces. INTELMO is under active development, with document available at https://intelmo.github.io.

## Poster session 1

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**#1 MemeCap: A Dataset for Captioning and Interpreting Memes**
*EunJeong Hwang and Vered Shwartz*
Memes are a widely popular tool for web users to express their thoughts using visual metaphors. Understanding memes requires recognizing and interpreting visual metaphors with respect to the text inside or around the meme, often while employing background knowledge and reasoning abilities. We present the task of meme captioning and release a new dataset, MemeCap. Our dataset contains 6.3K memes along with the title of the post containing the meme, the meme captions, the literal image caption, and the visual metaphors. Despite the recent success of vision and language (VL) models on tasks such as image captioning and visual question answering, our extensive experiments using state-of-the-art VL models show that they still struggle with visual metaphors, and perform substantially worse than humans.

11:00-12:30 (East Foyer / Virtual)
**#2 Incorporating Structured Representations into Pretrained Vision & Language Models Using Scene Graphs**
*Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell and Amir Globerson*
Vision and language models (VLMs) have demonstrated remarkable zero-shot (ZS) performance in a variety of tasks. However, recent works have shown that even the best VLMs struggle to capture aspects of compositional scene understanding, such as object attributes, relations, and action states. In contrast, obtaining structured annotations, such as scene graphs (SGs), that could improve these models is time-consuming and costly, and thus cannot be used on a large scale. Here we ask whether small SG datasets can provide sufficient information for enhancing structured understanding of pretrained VLMs. We show that it is indeed possible to improve VLMs when learning from SGs by integrating components that incorporate structured information into both visual and textual representations. For the visual side, we incorporate a special "SG Component" in the image transformer trained to predict SG information, while for the textual side, we utilize SGs to generate fine-grained captions that highlight different compositional aspects of the scene. Our method improves the performance of several popular VLMs on multiple VL datasets with only a mild degradation in ZS capabilities.

11:00-12:30 (East Foyer / Virtual)
**#3 From Wrong To Right: A Recursive Approach Towards Vision-Language Explanation**
*Jiaxin Ge, Sanjay Subramanian, Trevor Darrell and Boyi Li*
Addressing the challenge of adapting pre-trained vision-language models for generating insightful explanations for visual reasoning tasks with limited annotations, we present ReVisE: a Recursive Visual Explanation algorithm. Our method iteratively computes visual features (conditioned on the text input), an answer, and an explanation, to improve the explanation quality step by step until the answer converges. We find that this multi-step approach guides the model to correct its own answers and outperforms single-step explanation generation. Furthermore, explanations generated by ReVisE also serve as valuable annotations for few-shot self-training. Our approach outperforms previous methods while utilizing merely 5% of the human-annotated explanations across 10 metrics, demonstrating up to a 4.2 and 1.3 increase in BLEU-1 score on the VCR and VQA-X datasets, underscoring the efficacy and data-efficiency of our method.

11:00-12:30 (East Foyer / Virtual)
**#4 Variance Matters: Detecting Semantic Differences without Corpus/Word Alignment**
*Ryo Nagata, Hiroya Takamura, Naoki Otani and Yoshifumi Kawasaki*
In this paper, we propose methods for discovering semantic differences in words appearing in two corpora. The key idea is to measure the coverage of meanings of a word in a corpus through the norm of its mean word vector, which is equivalent to examining a kind of variance of the word vector distribution. The proposed methods do not require alignments between words and/or corpora for comparison that previous methods do. All they require are to compute variance (or norms of mean word vectors) for each word type. Nevertheless, they rival the best-performing system in the SemEval-2020 Task 1. In addition, they are (i) robust for the skew in corpus sizes; (ii) capable of detecting semantic differences in infrequent words; and (iii) effective in pinpointing word instances that have a meaning missing in one of the two corpora under comparison. We show these advantages for historical corpora and also for native/non-native English corpora.

11:00-12:30 (East Foyer / Virtual)
**#5 Improving Transformer-based Program Repair Model through False Behavior Diagnosis**
*Youngkyoung Kim, Misoo Kim and Eunseok Lee*
Research on automated program repairs using transformer-based models has recently gained considerable attention. The comprehension of the erroneous behavior of a model enables the identification of its inherent capacity and provides insights for improvement. However, the current landscape of research on program repair models lacks an investigation of their false behavior. Thus, we propose a methodology for diagnosing and treating the false behaviors of transformer-based program repair models. Specifically, we propose 1) a behavior vector that quantifies the behavior of the model when it generates an output, 2) a behavior discriminator (BeDisc) that identifies false behaviors, and 3) two methods for false behavior treatment. Through a large-scale experiment on 55,562 instances employing four datasets and three models,

the BeDisc exhibited a balanced accuracy of 86.6% for false behavior classification. The first treatment, namely, early abortion, successfully eliminated 60.4% of false behavior while preserving 97.4% repair accuracy. Furthermore, the second treatment, namely, masked bypassing, resulted in an average improvement of 40.5% in the top-1 repair accuracy. These experimental results demonstrated the importance of investigating false behaviors in program repair models.

11:00-12:30 (East Foyer / Virtual)
**#6 Coarse-to-Fine Contrastive Learning in Image-Text-Graph Space for Improved Vision-Language Compositionality**
*Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du and Yu Chen*
Contrastively trained vision-language models have achieved remarkable progress in vision and language representation learning. However, recent research has highlighted severe limitations of these models in their ability to perform compositional reasoning over objects, attributes, and relations. Scene graphs have emerged as an effective way to understand images compositionally. These are graph-structured semantic representations of images that contain objects, their attributes, and relations with other objects in a scene. In this work, we consider the scene graph parsed from text as a proxy for the image scene graph and propose a graph decomposition and augmentation framework along with a coarse-to-fine contrastive learning objective between images and text that aligns sentences of various complexities to the same image. We also introduce novel negative mining techniques in the scene graph space for improving attribute binding and relation understanding. Through extensive experiments, we demonstrate the effectiveness of our approach that significantly improves attribute binding, relation understanding, systematic generalization, and productivity on multiple recently proposed benchmarks (For example, improvements up to $\mathbf{18}\%$ for systematic generalization, $\mathbf{16.5}\%$ for relation understanding over a strong baseline), while achieving similar or better performance than CLIP on various general multimodal tasks.

11:00-12:30 (East Foyer / Virtual)
**#7 Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data**
*Canwen Xu, Daya Guo, Nan Duan and Julian McAuley*
Chat models, such as ChatGPT, have shown impressive capabilities and have been rapidly adopted across numerous domains. However, these models are only accessible through a restricted API, creating barriers for new research and progress in the field. We propose a pipeline that can automatically generate a high-quality multi-turn chat corpus by leveraging ChatGPT to engage in a conversation with itself. Subsequently, we employ parameter-efficient tuning to enhance LLaMA, an open-source large language model. The resulting model, named Baize, demonstrates good performance in multi-turn dialogues with guardrails that minimize potential risks. Additionally, we propose a new technique called Self-Distill with Feedback, to further improve the performance of the Baize models with feedback from ChatGPT.

11:00-12:30 (East Foyer / Virtual)
**#8 AutoTrial: Prompting Language Models for Clinical Trial Design**
*Zifeng Wang, Cao Xiao and Jimeng Sun*
Clinical trials are critical for drug development. Constructing the appropriate eligibility criteria (i.e., the inclusion/exclusion criteria for patient recruitment) is essential for the trial's success. Proper design of clinical trial protocols should consider similar precedent trials and their eligibility criteria to ensure sufficient patient coverage. In this paper, we present a method named AutoTrial to aid the design of clinical eligibility criteria using language models. It allows (1) controllable generation under instructions via a hybrid of discrete and neural prompting, (2) scalable knowledge incorporation via in-context learning, and (3) explicit reasoning chains to provide rationales for understanding the outputs. Experiments on over 70K clinical trials verify that AutoTrial generates high-quality criteria texts that are fluent and coherent and with high accuracy in capturing the relevant clinical concepts to the target trial. It is noteworthy that our method, with a much smaller parameter size, gains around 60% winning rate against the GPT-3.5 baselines via human evaluations.

11:00-12:30 (East Foyer / Virtual)
**#9 POE: Process of Elimination for Multiple Choice Reasoning**
*Chenkai Ma and Xinya Du*
Language models (LMs) are capable of conducting in-context learning for multiple choice reasoning tasks, but the options in these tasks are treated equally. As humans often first eliminate wrong options before picking the final correct answer, we argue a similar two-step strategy can make LMs better at these tasks. To this end, we present the Process of Elimination (POE), a two-step scoring method. In the first step, POE scores each option, and eliminates seemingly wrong options. In the second step, POE masks these wrong options, and makes the final prediction from the remaining options. Zero-shot experiments on 8 reasoning tasks illustrate the effectiveness of POE, and a following analysis finds our method to be especially performant on logical reasoning tasks. We further analyze the effect of masks, and show that POE applies to few-shot settings and large language models (LLMs) like ChatGPT.

11:00-12:30 (East Foyer / Virtual)
**#10 MProto: Multi-Prototype Network with Denoised Optimal Transport for Distantly Supervised Named Entity Recognition**
*Shuhui Wu, Yongliang Shen, Zeqi Tan, Wenqi Ren, Jietian Guo, Shiliang Pu and Weiming Lu*
Distantly supervised named entity recognition (DS-NER) aims to locate entity mentions and classify their types with only knowledge bases or gazetteers and unlabeled corpus. However, distant annotations are noisy and degrade the performance of NER models. In this paper, we propose a noise-robust prototype network named MProto for the DS-NER task. Different from previous prototype-based NER methods, MProto represents each entity type with multiple prototypes to characterize the intra-class variance among entity representations. To optimize the classifier, each token should be assigned an appropriate ground-truth prototype and we consider such token-prototype assignment as an optimal transport (OT) problem. Furthermore, to mitigate the noise from incomplete labeling, we propose a novel denoised optimal transport (DOT) algorithm. Specifically, we utilize the assignment result between *Other* class tokens and all prototypes to distinguish unlabeled entity tokens from true negatives. Experiments on several DS-NER benchmarks demonstrate that our MProto achieves state-of-the-art performance. The source code is now available on Github.

11:00-12:30 (East Foyer / Virtual)
**#11 Referring Image Segmentation via Joint Mask Contextual Embedding Learning and Progressive Alignment Network**
*Ziling Huang and Shin'ichi Satoh*
Referring image segmentation is a task that aims to predict pixel-wise masks corresponding to objects in an image described by natural language expressions. Previous methods for referring image segmentation employ a cascade framework to break down complex problems into multiple stages. However, its defects also obvious: existing methods within the cascade framework may encounter challenges in both maintaining a strong focus on the most relevant information during specific stages of the referring image segmentation process and rectifying errors propagated from early stages, which can ultimately result in sub-optimal performance. To address these limitations, we propose the Joint Mask Contextual Embedding Learning Network (JMCELN). JMCELN is designed to enhance the Cascade Framework by incorporating a Learnable Contextual Embedding and a Progressive Alignment Network (PAN). The Learnable Contextual Embedding module dynamically stores and utilizes reasoning information based on the current mask prediction results, enabling the network to adaptively capture and refine pertinent information for improved mask prediction accuracy. Furthermore, the Progressive Alignment Network (PAN) is introduced as an integral part of JMCELN. PAN leverages the output from the previous layer as a filter for the current output, effectively reducing inconsisten-

cies between predictions from different stages. By iteratively aligning the predictions, PAN guides the Learnable Contextual Embedding to incorporate more discriminative information for reasoning, leading to enhanced prediction quality and a reduction in error propagation. With these methods, we achieved state-of-the-art results on three commonly used benchmarks, especially in more intricate datasets. The code will be released.

11:00-12:30 (East Foyer / Virtual)
**#12 DiSTRICT: Dialogue State Tracking with Retriever Driven In-Context Tuning**
*Praveen Venkateswaran, Evelyn Duesterwald and Vatche Isahagian*
Dialogue State Tracking (DST), a key component of task-oriented conversation systems, represents user intentions by determining the values of pre-defined slots in an ongoing dialogue. Existing approaches use hand-crafted templates and additional slot information to fine-tune and prompt large pre-trained language models and elicit slot values from the dialogue context. Significant manual effort and domain knowledge is required to design effective prompts, limiting the generalizability of these approaches to new domains and tasks. In this work, we propose DiSTRICT, a generalizable in-context tuning approach for DST that retrieves highly relevant training examples for a given dialogue to fine-tune the model without any hand-crafted templates. Experiments with the MultiWOZ benchmark datasets show that DiSTRICT outperforms existing approaches in various zero-shot and few-shot settings using a much smaller model, thereby providing an important advantage for real-world deployments that often have limited resource availability.

11:00-12:30 (East Foyer / Virtual)
**#13 Generative Table Pre-training Empowers Models for Tabular Prediction**
*Tianping Zhang, Shaowen Wang, Shuicheng Yan, Li Jian and Qian Liu*
Recently, the topic of table pre-training has attracted considerable research interest. However, how to employ table pre-training to boost the performance of tabular prediction remains an open challenge. In this paper, we propose TapTap, the first attempt that leverages table pre-training to empower models for tabular prediction. After pre-training on a large corpus of real-world tabular data, TapTap can generate high-quality synthetic tables to support various applications on tabular data, including privacy protection, low resource regime, missing value imputation, and imbalanced classification. Extensive experiments on 12 datasets demonstrate that TapTap outperforms a total of 16 baselines in different scenarios. Meanwhile, it can be easily combined with various backbone models, including LightGBM, Multilayer Perceptron (MLP) and Transformer. Moreover, with the aid of table pre-training, models trained using synthetic data generated by TapTap can even compete with models using the original dataset on half of the experimental datasets, marking a milestone in the development of synthetic tabular data generation. The code and datasets are available at https://github.com/ZhangTP1996/TapTap.

11:00-12:30 (East Foyer / Virtual)
**#14 Spoiler Detection as Semantic Text Matching**
*Ryan Tran, Canwen Xu and Julian McAuley*
Engaging with discussion of TV shows online often requires individuals to refrain from consuming show-related content for extended periods to avoid spoilers. While existing research on spoiler detection shows promising results in safeguarding viewers from general spoilers, it fails to address the issue of users abstaining from show-related content during their watch. This is primarily because the definition of a spoiler varies depending on the viewer's progress in the show, and conventional spoiler detection methods lack the granularity to capture this complexity. To tackle this challenge, we propose the task of spoiler matching, which involves assigning an episode number to a spoiler given a specific TV show. We frame this task as semantic text matching and introduce a dataset comprised of comments and episode summaries to evaluate model performance. Given the length of each example, our dataset can also serve as a benchmark for long-range language models.

11:00-12:30 (East Foyer / Virtual)
**#15 Prompting Scientific Names for Zero-Shot Species Recognition**
*Shubham Parashar, Zhiqiu Lin, Yanan Li and Shu Kong*
Trained on web-scale image-text pairs, Vision-Language Models (VLMs) such as CLIP can recognize images of common objects in a zero-shot fashion. However, it is underexplored how to use CLIP for zero-shot recognition of highly specialized concepts, e.g., species of birds, plants, and animals, for which their scientific names are written in Latin or Greek. Indeed, CLIP performs poorly for zero-shot species recognition with prompts that use scientific names, e.g., "a photo of Lepus Timidus" (which is a scientific name in Latin). This is because these names are usually not included in CLIP's training set. To improve performance, we explore using large-language models (LLMs) to generate descriptions (e.g., of species color and shape) and additionally use them in prompts. However, this method improves only marginally. Instead, we are motivated to translate scientific names (e.g., Lepus Timidus) to common English names (e.g., `mountain hare`) and use such in the prompts. We find that common names are more likely to be included in CLIP's training set, and prompting them achieves 2$\sim$5 times higher accuracy on benchmarking datasets of fine-grained species recognition.

11:00-12:30 (East Foyer / Virtual)
**#16 R2H: Building Multimodal Navigation Helpers that Respond to Help Requests**
*Yue Fan, Jing Gu, Kaizhi Zheng and Xin Eric Wang*
Intelligent navigation-helper agents are critical as they can navigate users in unknown areas through environmental awareness and conversational ability, serving as potential accessibility tools for individuals with disabilities. In this work, we first introduce a novel benchmark, Respond to Help Requests (R2H), to promote the development of multi-modal navigation helpers capable of responding to requests for help, utilizing existing dialog-based embodied datasets. R2H mainly includes two tasks: (1) Respond to Dialog History (RDH), which assesses the helper agent's ability to generate informative responses based on a given dialog history, and (2) Respond during Interaction (RdI), which evaluates the effectiveness and efficiency of the response during consistent cooperation with a task performer. Furthermore, we explore two approaches to construct the navigation-helper agent, including fine-tuning a novel task-oriented multi-modal response generation model that can see and respond, named SeeRee, and employing a multi-modal large language model in a zero-shot manner. Analysis of the task and method was conducted based on both automatic benchmarking and human evaluations.

11:00-12:30 (East Foyer / Virtual)
**#17 Image Manipulation via Multi-Hop Instructions - A New Dataset and Weakly-Supervised Neuro-Symbolic Approach**
*Harman Singh, Poorva Garg, Mohit Gupta, Kevin Shah, Ashish Goswami, Satyam Modi, Arnab Kumar Mondal, Dinesh Khandelwal, Dinesh Garg and Parag Singla*
We are interested in image manipulation via natural language text – a task that is useful for multiple AI applications but requires complex reasoning over multi-modal spaces. We extend recently proposed Neuro Symbolic Concept Learning (NSCL), which has been quite effective for the task of Visual Question Answering (VQA), for the task of image manipulation. Our system referred to as NeuroSIM can perform complex multi-hop reasoning over multi-object scenes and only requires weak supervision in the form of annotated data for VQA. NeuroSIM parses an instruction into a symbolic program, based on a Domain Specific Language (DSL) comprising of object attributes and manipulation operations, that guides its execution. We create a new dataset for the task, and extensive experiments demonstrate that NeuroSIM is highly competitive with or beats SOTA baselines that make use of supervised data for manipulation.

11:00-12:30 (East Foyer / Virtual)
**#18 q2d: Turning Questions into Dialogs to Teach Models How to Search**
*Yonatan Bitton, Shlomi Cohen-Ganor, Ido Hakimi, Yoad Lewenberg, Roee Aharoni and Enav Weinreb*
One of the exciting capabilities of recent language models for dialog is their ability to independently search for relevant information to ground a given dialog response. However, obtaining training data to teach models how to issue search queries is time and resource consuming. In this work, we propose $q2d$: an automatic data generation pipeline that generates information-seeking dialogs from questions. We prompt a large language model (PaLM) to create conversational versions of question answering datasets, and use it to improve query generation models that communicate with external search APIs to ground dialog responses. Unlike previous approaches which relied on human written dialogs with search queries, our method allows to automatically generate query-based grounded dialogs with better control and scale. Our experiments demonstrate that: (1) For query generation on the QReCC dataset, models trained on our synthetically-generated data achieve 90%-97% of the performance of models trained on the human-generated data; (2) We can successfully generate data for training dialog models in new domains without any existing dialog data as demonstrated on the multi-hop MuSiQue and Bamboogle QA datasets. (3) We perform a thorough analysis of the generated dialogs showing that humans find them of high quality and struggle to distinguish them from human-written dialogs.

11:00-12:30 (East Foyer / Virtual)
**#19 The Benefits of Label-Description Training for Zero-Shot Text Classification**
*Lingyu Gao, Debanjan Ghosh and Kevin Gimpel*
Pretrained language models have improved zero-shot text classification by allowing the transfer of semantic knowledge from the training data in order to classify among specific label sets in downstream tasks. We propose a simple way to further improve zero-shot accuracies with minimal effort. We curate small finetuning datasets intended to describe the labels for a task. Unlike typical finetuning data, which has texts annotated with labels, our data simply describes the labels in language, e.g., using a few related terms, dictionary/encyclopedia entries, and short templates. Across a range of topic and sentiment datasets, our method is more accurate than zero-shot by 17-19% absolute. It is also more robust to choices required for zero-shot classification, such as patterns for prompting the model to classify and mappings from labels to tokens in the model's vocabulary. Furthermore, since our data merely describes the labels but does not use input texts, finetuning on it yields a model that performs strongly on multiple text domains for a given label set, even improving over few-shot out-of-domain classification in multiple settings.

11:00-12:30 (East Foyer / Virtual)
**#20 GEM: Gestalt Enhanced Markup Language Model for Web Understanding via Render Tree**
*Zirui Shao, Feiyu Gao, Zhongda Qi, Hangdi Xing, Jiajun Bu, Zhi Yu, Qi Zheng and Xiaozhong Liu*
Inexhaustible web content carries abundant perceptible information beyond text. Unfortunately, most prior efforts in pre-trained Language Models (LMs) ignore such cyber-richness, while few of them only employ plain HTMLs, and crucial information in the rendered web, such as visual, layout, and style, are excluded. Intuitively, those perceptible web information can provide essential intelligence to facilitate content understanding tasks. This study presents an innovative Gestalt Enhanced Markup (GEM) Language Model inspired by Gestalt psychological theory for hosting heterogeneous visual information from the render tree into the language model without requiring additional visual input. Comprehensive experiments on multiple downstream tasks, i.e., web question answering and web information extraction, validate GEM superiority.

11:00-12:30 (East Foyer / Virtual)
**#21 Pre-training Intent-Aware Encoders for Zero- and Few-Shot Intent Classification**
*Mujeen Sung, James Gung, Elman Mansimov, Nikolaos Pappas, Raphael Shu, Salvatore Romeo, Yi Zhang and Vittorio Castelli*
Intent classification (IC) plays an important role in task-oriented dialogue systems. However, IC models often generalize poorly when training without sufficient annotated examples for each user intent. We propose a novel pre-training method for text encoders that uses contrastive learning with intent psuedo-labels to produce embeddings that are well-suited for IC tasks, reducing the need for manual annotations. By applying this pre-training strategy, we also introduce Pre-trained Intent-aware Encoder (PIE), which is designed to align encodings of utterances with their intent names. Specifically, we first train a tagger to identify key phrases within utterances that are crucial for interpreting intents. We then use these extracted phrases to create examples for pre-training a text encoder in a contrastive manner. As a result, our PIE model achieves up to 5.4% and 4.0% higher accuracy than the previous state-of-the-art pre-trained text encoder for the N-way zero- and one-shot settings on four IC datasets.

11:00-12:30 (East Foyer / Virtual)
**#22 IC3: Image Captioning by Committee Consensus**
*David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross and John Canny*
If you ask a human to describe an image, they might do so in a thousand different ways. Traditionally, image captioning models are trained to generate a single "best' (most like a reference) image caption. Unfortunately, doing so encourages captions that are "informationally impoverished,' and focus on only a subset of the possible details, while ignoring other potentially useful information in the scene. In this work, we introduce a simple, yet novel, method: "Image Captioning by Committee Consensus' (IC3), designed to generate a single caption that captures high-level details from several annotator viewpoints. Humans rate captions produced by IC3 at least as helpful as baseline SOTA models more than two thirds of the time, and IC3 can improve the performance of SOTA automated recall systems by up to 84%, outperforming single human-generated reference captions, and indicating significant improvements over SOTA approaches for visual description. Code is available at [https://davidmchan.github.io/caption-by-committee/](https://davidmchan.github.io/caption-by-committee/)

11:00-12:30 (East Foyer / Virtual)
**#23 Towards Conceptualization of "Fair Explanation": Disparate Impacts of anti-Asian Hate Speech Explanations on Content Moderators**
*Tin Trung Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III and Marine Carpuat*
Recent research at the intersection of AI explainability and fairness has focused on how explanations can improve human-plus-AI task performance as assessed by fairness measures. We propose to characterize what constitutes an explanation that is itself "fair" – an explanation that does not adversely impact specific populations. We formulate a novel evaluation method of "fair explanations" using not just accuracy and label time, but also psychological impact of explanations on different user groups across many metrics (mental discomfort, stereotype activation, and perceived workload). We apply this method in the context of content moderation of potential hate speech, and its differential impact on Asian vs. non-Asian proxy moderators, across explanation approaches (saliency map and counterfactual explanation). We find that saliency maps generally perform better and show less evidence of disparate impact (group) and individual unfairness than counterfactual explanations. Content warning: This paper contains examples of hate speech and racially discriminatory language. The authors do not support such content. Please consider your risk of discomfort carefully before continuing reading!

11:00-12:30 (East Foyer / Virtual)
**#24 Contrastive Learning for Inference in Dialogue**
*Etsuko Ishii, Yan Xu, Bryan Wilie, Ziwei Ji, Holy Lovenia, Willy Chung and Pascale Fung*

Inference, especially those derived from inductive processes, is a crucial component in our conversation to complement the information implicitly or explicitly conveyed by a speaker. While recent large language models show remarkable advances in inference tasks, their performance in inductive reasoning, where not all information is present in the context, is far behind deductive reasoning. In this paper, we analyze the behavior of the models based on the task difficulty defined by the semantic information gap – which distinguishes inductive and deductive reasoning. Our analysis reveals that the information gap between dialogue contexts and desired inferences renders the inductive inference process more challenging. To mitigate this information gap, we investigate a contrastive learning approach by feeding negative samples. Our experiments suggest negative samples help models understand what is wrong and improve their inference generations.

11:00-12:30 (East Foyer / Virtual)
**#25 Post-hoc Utterance Refining Method by Entity Mining for Faithful Knowledge Grounded Conversations**
*Yoonna Jang, Suhyune Son, Jeongwoo Lee, Junyoung Son, Yuna Hur, Jungwoo Lim, Hyeonseok Moon, Kisu Yang and Heuiseok Lim*
Despite the striking advances in recent language generation performance, model-generated responses have suffered from the chronic problem of hallucinations that are either untrue or unfaithful to a given source. Especially in the task of knowledge grounded conversation, the models are required to generate informative responses, but hallucinated utterances lead to miscommunication. In particular, entity-level hallucination that causes critical misinformation and undesirable conversation is one of the major concerns. To address this issue, we propose a post-hoc refinement method called REM. It aims to enhance the quality and faithfulness of hallucinated utterances by refining them based on the source knowledge. If the generated utterance has a low source-faithfulness score with the given knowledge, REM mines the key entities in the knowledge and implicitly uses them for refining the utterances. We verify that our method reduces entity hallucination in the utterance. Also, we show the adaptability and efficacy of REM with extensive experiments and generative results. Our code is available at https://github.com/YOONNAJANG/REM.

11:00-12:30 (East Foyer / Virtual)
**#26 Content- and Topology-Aware Representation Learning for Scientific Multi-Literature**
*Kai Zhang, Kaisong Song, Yangyang Kang and Xiaozhong Liu*
Representation learning forms an essential building block in the development of natural language processing architectures. To date, mainstream approaches focus on learning textual information at the sentence- or document-level, unfortunately, overlooking the inter-document connections. This omission decreases the potency of downstream applications, particularly in multi-document settings. To address this issue, embeddings equipped with latent semantic and rich relatedness information are needed. In this paper, we propose SMRC$^2$, which extends representation learning to the multi-document level. Our model jointly learns latent semantic information from content and rich relatedness information from topological networks. Unlike previous studies, our work takes multi-document as input and integrates both semantic and relatedness information using a shared space via language model and graph structure. Our extensive experiments confirm the superiority and effectiveness of our approach. To encourage further research in scientific multi-literature representation learning, we will release our code and a new dataset from the biomedical domain.

11:00-12:30 (East Foyer / Virtual)
**#27 Can Language Models Understand Physical Concepts?**
*Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong and Qi Liu*
Language models (LMs) gradually become general-purpose interfaces in the interactive and embodied world, where the understanding of physical concepts is an essential prerequisite. However, it is unclear whether LMs can understand physical concepts in the human world. To investigate this, we design a benchmark VEC that covers the tasks of (i) Visual concepts, such as the shape and material of objects, and (ii) Embodied Concepts, learned from the interaction with the world such as the temperature of objects. Our zero (few)-shot prompting results show that the understanding of certain visual concepts emerges as scaling up LMs, but there are still basic concepts to which the scaling law does not apply. For example, OPT-175B performs close to humans with a zero-shot accuracy of $85\%$ on the material concept, yet behaves like random guessing on the mass concept. Instead, vision-augmented LMs such as CLIP and BLIP achieve a human-level understanding of embodied concepts. Analysis indicates that the rich semantics in visual representation can serve as a valuable source of embodied knowledge. Inspired by this, we propose a distillation method to transfer embodied knowledge from VLMs to LMs, achieving performance gain comparable with that by scaling up parameters of LMs $134\times$. Our dataset is available at https://github.com/TobiasLee/VEC.

11:00-12:30 (East Foyer / Virtual)
**#28 GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations**
*Muhammet Furkan Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Hwee Lim and Mike Zheng Shou*
The usage of exocentric and egocentric videos in Video Question Answering (VQA) is a new endeavor in human-robot interaction and collaboration studies. Particularly for egocentric videos, one may leverage eye-gaze information to understand human intentions during the task. In this paper, we build a novel task-oriented VQA dataset, called GazeVQA, for collaborative tasks where gaze information is captured during the task process. GazeVQA is designed with a novel QA format that covers thirteen different reasoning types to capture multiple aspects of task information and user intent. For each participant, GazeVQA consists of more than 1,100 textual questions and more than 500 labeled images that were annotated with the assistance of the Segment Anything Model. In total, 2,967 video clips, 12,491 labeled images, and 25,040 questions from 22 participants were included in the dataset. Additionally, inspired by the assisting models and common ground theory for industrial task collaboration, we propose a new AI model called AssistGaze that is designed to answer the questions with three different answer types, namely textual, image, and video. AssistGaze can effectively ground the perceptual input into semantic information while reducing ambiguities. We conduct comprehensive experiments to demonstrate the challenges of GazeVQA and the effectiveness of AssistGaze.

11:00-12:30 (East Foyer / Virtual)
**#29 Continual Named Entity Recognition without Catastrophic Forgetting**
*Duzhen Zhang, Wei Cong, Jiahua Dong, Yahan Yu, Xiuyi Chen, Yonggang Zhang and Zhen Fang*
Continual Named Entity Recognition (CNER) is a burgeoning area, which involves updating an existing model by incorporating new entity types sequentially. Nevertheless, continual learning approaches are often severely afflicted by catastrophic forgetting. This issue is intensified in CNER due to the consolidation of old entity types from previous steps into the non-entity type at each step, leading to what is known as the semantic shift problem of the non-entity type. In this paper, we introduce a pooled feature distillation loss that skillfully navigates the trade-off between retaining knowledge of old entity types and acquiring new ones, thereby more effectively mitigating the problem of catastrophic forgetting. Additionally, we develop a confidence-based pseudo-labeling for the non-entity type, i.e., predicting entity types using the old model to handle the semantic shift of the non-entity type. Following the pseudo-labeling process, we suggest an adaptive re-weighting type-balanced learning strategy to handle the issue of biased type distribution. We carried out comprehensive experiments on ten CNER settings using three different datasets. The results illustrate that our method significantly outperforms prior state-of-the-art approaches, registering an average improvement of 6.3% and 8.0% in Micro and Macro F1 scores, respectively.

11:00-12:30 (East Foyer / Virtual)
**#30 A Generation-based Deductive Method for Math Word Problems**
*Yuxuan Hu, Jing Zhang, Haoyang Li, Cuiping Li and Hong Chen*

Math word problems (MWP) involving advanced operators such as linear equation solver cannot be easily tackled by earlier MWP methods, because the existing generation methods suffer from repeated sub-expression generation and deductive methods are restricted to dealing with binary operations. This paper propose a new multivariate directed acyclic graph (mDAG) as an alternative to the generation methods' binary expression tree or the deductive methods' binary directed acyclic graph. Then to produce the topological ordering of mDAG, we propose a generation-based deductive (GeDe) model, which equips a generation model with a re-encoder to keep the deductive property but avoid the expensive enumeration of the deductive methods. GeDe performs well on math problems with many operators on the widely used benchmarks as well as solving multivariate operators on our own CMWPA benchmark. Our code is available at https://github.com/hyx1999/GeDe

11:00-12:30 (East Foyer / Virtual)
**#31 GNAT: A General Narrative Alignment Tool**
*Tanzir Pial and Steven Skiena*
Algorithmic sequence alignment identifies similar segments shared between pairs of documents, and is fundamental to many NLP tasks. But it is difficult to recognize similarities between distant versions of narratives such as translations and retellings, particularly for summaries and abridgements which are much shorter than the original novels. We develop a general approach to narrative alignment coupling the Smith-Waterman algorithm from bioinformatics with modern text similarity metrics. We show that the background of alignment scores fits a Gumbel distribution, enabling us to define rigorous p-values on the significance of any alignment. We apply and evaluate our general narrative alignment tool (GNAT) on four distinct problem domains differing greatly in both the relative and absolute length of documents, namely summary-to-book alignment, translated book alignment, short story alignment, and plagiarism detection—demonstrating the power and performance of our methods.

11:00-12:30 (East Foyer / Virtual)
**#32 Analyzing Film Adaptation through Narrative Alignment**
*Tanzir Pial, Shahreen Salim Aunti, Charuta Pethe, Allen Kim and Steven Skiena*
Novels are often adapted into feature films, but the differences between the two media usually require dropping sections of the source text from the movie script. Here we study this screen adaptation process by constructing narrative alignments using the Smith-Waterman local alignment algorithm coupled with SBERT embedding distance to quantify text similarity between scenes and book units. We use these alignments to perform an automated analysis of 40 adaptations, revealing insights into the screenwriting process concerning (i) faithfulness of adaptation, (ii) importance of dialog, (iii) preservation of narrative order, and (iv) gender representation issues reflective of the Bechdel test.

11:00-12:30 (East Foyer / Virtual)
**#33 DALE: Generative Data Augmentation for Low-Resource Legal NLP**
*Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi and Dinesh Manocha*
We present DALE, a novel and effective generative Data Augmentation framework for low-resource LEgal NLP. DALE addresses the challenges existing frameworks pose in generating effective data augmentations of legal documents - legal language, with its specialized vocabulary and complex semantics, morphology, and syntax, does not benefit from data augmentations that merely rephrase the source sentence. To address this, DALE, built on an Encoder-Decoder Language Model, is pre-trained on a novel unsupervised text denoising objective based on selective masking - our masking strategy exploits the domain-specific language characteristics of templatized legal documents to mask collocated spans of text. Denoising these spans help DALE acquire broad legal knowledge and develop the ability to generate coherent and diverse augmentations with novel contexts. Finally, DALE performs conditional generation to generate synthetic augmentations for low-resource Legal NLP tasks. We demonstrate the effectiveness of DALE on 13 datasets spanning 6 tasks and 4 low-resource settings. DALE outperforms all our baselines, including LLMs, qualitatively and quantitatively, with absolute improvements of 1%-50%.

11:00-12:30 (East Foyer / Virtual)
**#34 CompoundPiece: Evaluating and Improving Decompounding Performance of Language Models**
*Benjamin Minixhofer, Jonas Pfeiffer and Ivan Vulić*
While many languages possess processes of joining two or more words to create compound words, previous studies have been typically limited only to languages with excessively productive compound formation (e.g., German, Dutch) and there is no public dataset containing compound and non-compound words across a large number of languages. In this work, we systematically study decompounding, the task of splitting compound words into their constituents, at a wide scale. We first address the data gap by introducing a dataset of 255k compound and non-compound words across 56 diverse languages obtained from Wiktionary. We then use this dataset to evaluate an array of Large Language Models (LLMs) on the decompounding task. We find that LLMs perform poorly, especially on words which are tokenized unfavorably by subword tokenization. We thus introduce a novel methodology to train dedicated models for decompounding. The proposed two-stage procedure relies on a fully self-supervised objective in the first stage, while the second, supervised learning stage optionally fine-tunes the model on the annotated Wiktionary data. Our self-supervised models outperform the prior best unsupervised decompounding models by 13.9% accuracy on average. Our fine-tuned models outperform all prior (language-specific) decompounding tools. Furthermore, we use our models to leverage decompounding during the creation of a subword tokenizer, which we refer to as CompoundPiece. CompoundPiece tokenizes compound words more favorably on average, leading to improved performance on decompounding over an otherwise equivalent model using SentencePiece tokenization.

11:00-12:30 (East Foyer / Virtual)
**#35 ChatEdit: Towards Multi-turn Interactive Facial Image Editing via Dialogue**
*Xing Cui, Zekun Li, Pei Pei Li, Yibo Hu, Hailin Shi, Chunshui Cao and Zhaofeng He*
This paper explores interactive facial image editing through dialogue and presents the ChatEdit benchmark dataset for evaluating image editing and conversation abilities in this context. ChatEdit is constructed from the CelebA-HQ dataset, incorporating annotated multi-turn dialogues corresponding to user editing requests on the images. The dataset is challenging, as it requires the system to dynamically track and edit images based on user requests, while generating appropriate natural language responses. To address these challenges, we propose a framework comprising a dialogue module for tracking user requests as well as generating responses, and an image editing module for editing images accordingly. Unlike previous approaches, our framework directly tracks the user request of the current turn from the entire dialogue history and edits the initial image instead of manipulating the output from the previous turn, mitigating error accumulation and attribute forgetting issues. Extensive experiments on the ChatEdit dataset demonstrate the superiority of our framework over previous methods and also improvement rooms, encouraging future research. We will release the code and data publicly to facilitate advancements in complex interactive facial image editing.

11:00-12:30 (East Foyer / Virtual)
**#36 Towards LLM-driven Dialogue State Tracking**
*Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan and Xiao-Ming Wu*
Dialogue State Tracking (DST) is of paramount importance in ensuring accurate tracking of user goals and system actions within task-oriented dialogue systems. The emergence of large language models (LLMs) such as GPT3 and ChatGPT has sparked considerable interest in assessing their efficacy across diverse applications. In this study, we conduct an initial examination of ChatGPT's capabilities in DST. Our evaluation

uncovers the exceptional performance of ChatGPT in this task, offering valuable insights to researchers regarding its capabilities and providing useful directions for designing and enhancing dialogue systems. Despite its impressive performance, ChatGPT has significant limitations including its closed-source nature, request restrictions, raising data privacy concerns, and lacking local deployment capabilities. To address these concerns, we present LDST, an LLM-driven DST framework based on smaller, open-source foundation models. By utilizing a novel domain-slot instruction tuning method, LDST achieves performance on par with ChatGPT. Comprehensive evaluations across three distinct experimental settings, we find that LDST exhibits remarkable performance improvements in both zero-shot and few-shot setting compared to previous SOTA methods. The source code is provided for reproducibility.

11:00-12:30 (East Foyer / Virtual)
**#37 Turn-Level Active Learning for Dialogue State Tracking**
*Zihan Zhang, Meng Fang, Fanghua Ye, Ling Chen and Mohammad-Reza Namazi-Rad*
Dialogue state tracking (DST) plays an important role in task-oriented dialogue systems. However, collecting a large amount of turn-by-turn annotated dialogue data is costly and inefficient. In this paper, we propose a novel turn-level active learning framework for DST to actively select turns in dialogues to annotate. Given the limited labelling budget, experimental results demonstrate the effectiveness of selective annotation of dialogue turns. Additionally, our approach can effectively achieve comparable DST performance to traditional training approaches with significantly less annotated data, which provides a more efficient way to annotate new dialogue data.

11:00-12:30 (East Foyer / Virtual)
**#38 StructGPT: A General Framework for Large Language Model to Reason over Structured Data**
*Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao and Ji-Rong Wen*
In this paper, we aim to improve the reasoning ability of large language models (LLMs) over structured data in a unified way. Inspired by the studies on tool augmentation for LLMs, we develop an Iterative Reading-then-Reasoning (IRR) framework to solve question answering tasks based on structured data, called StructGPT. In this framework, we construct the specialized interfaces to collect relevant evidence from structured data (i.e., reading), and let LLMs concentrate on the reasoning task based on the collected information (i.e., reasoning). Specially, we propose an invoking-linearization-generation procedure to support LLMs in reasoning on the structured data with the help of the interfaces. By iterating this procedure with provided interfaces, our approach can gradually approach the target answers to a given query. Experiments conducted on three types of structured data show that StructGPT greatly improves the performance of LLMs, under the few-shot and zero-shot settings.

11:00-12:30 (East Foyer / Virtual)
**#39 Towards Low-Resource Automatic Program Repair with Meta-Learning and Pretrained Language Models**
*Weishi Wang, Yue Wang, Steven Hoi and Shafiq Joty*
Automatic program repair (APR) has gained increasing attention as an essential technique in software development to reduce manual debugging efforts and boost developers' productivity. Recent advances in deep learning (DL) based models have demonstrated promising results by learning from large-scale bug-fix examples in a data-driven manner. However, in practical scenarios, software bugs have an imbalanced distribution, and the fixing knowledge learned by APR models often only capture the patterns of frequent error types, making it inapplicable to handle the rare error types. To address this limitation, we investigate a novel task of low-resource APR, and propose Meta-APR, a new meta-learning framework integrated with code pretrained language models to generate fixes for low-resource bugs with limited training samples. Our Meta-APR learns better error-specific knowledge from high-resource bugs through efficient first-order meta-learning optimization, which allows for a faster adaptation to the target low-resource bugs. Besides, while we adopt CodeT5, a pretrained code-aware encoder-decoder Transformer, as the backbone model for Meta-APR, it is a model-agnostic framework that can be integrated with any neural models. Extensive experimental results on three benchmarks in various programming languages verify the superiority of our method over existing DL-based APR approaches.

11:00-12:30 (East Foyer / Virtual)
**#40 Semi-supervised multimodal coreference resolution in image narrations**
*Arushi Goel, Basura Fernando, Frank Keller and Hakan Bilen*
In this paper, we study multimodal coreference resolution, specifically where a longer descriptive text, i.e., a narration is paired with an image. This poses significant challenges due to fine-grained image-text alignment, inherent ambiguity present in narrative language, and unavailability of large annotated training sets. To tackle these challenges, we present a data efficient semi-supervised approach that utilizes image-narration pairs to resolve coreferences and narrative grounding in a multimodal context. Our approach incorporates losses for both labeled and unlabeled data within a cross-modal framework. Our evaluation shows that the proposed approach outperforms strong baselines both quantitatively and qualitatively, for the tasks of coreference resolution and narrative grounding.

11:00-12:30 (East Foyer / Virtual)
**#41 Beyond Detection: A Defend-and-Summarize Strategy for Robust and Interpretable Rumor Analysis on Social Media**
*Yi-Ting Chang, Yun-Zhu Song, Yi-Syuan Chen and Hong-Han Shuai*
As the impact of social media gradually escalates, people are more likely to be exposed to indistinguishable fake news. Therefore, numerous studies have attempted to detect rumors on social media by analyzing the textual content and propagation paths. However, fewer works on rumor detection tasks consider the malicious attacks commonly observed at response level. Moreover, existing detection models have poor interpretability. To address these issues, we propose a novel framework named **D**efend-**A**nd-**S**ummarize (DAS) based on the concept that responses sharing similar opinions should exhibit similar features. Specifically, DAS filters out the attack responses and summarizes the responsive posts of each conversation thread in both extractive and abstractive ways to provide multi-perspective prediction explanations. Furthermore, we enhance our detection architecture with the transformer and Bi-directional Graph Convolutional Networks. Experiments on three public datasets, *i.e.*, RumorEval2019, Twitter15, and Twitter16, demonstrate that our DAS defends against malicious attacks and provides prediction explanations, and the proposed detection model achieves state-of-the-art.

11:00-12:30 (East Foyer / Virtual)
**#42 VLIS: Unimodal Language Models Guide Multimodal Language Generation**
*Jiwan Chung and Youngjae Yu*
Multimodal language generation, which leverages the synergy of language and vision, is a rapidly expanding field. However, existing vision-language models face challenges in tasks that require complex linguistic understanding. To address this issue, we introduce Visual-Language models as Importance Sampling weights (VLIS), a novel framework that combines the visual conditioning capability of vision-language models with the language understanding of unimodal text-only language models without further training. It extracts pointwise mutual information of each image and text from a visual-language model and uses the value as an importance sampling weight to adjust the token likelihood from a text-only model. VLIS improves vision-language models on diverse tasks, including commonsense understanding (WHOOPS, OK-VQA, and ScienceQA) and complex text generation (Concadia, Image Paragraph Captioning, and ROCStories). Our results suggest that VLIS represents a promising new direction for multimodal language generation.

11:00-12:30 (East Foyer / Virtual)
**#43 Event Ontology Completion with Hierarchical Structure Evolution Networks**
*Pengfei Cao, Yupu Hao, Yubo Chen, Kang Liu, Jiexin Xu, Huaijun Li, Xiaojian Jiang and Jun Zhao*
Traditional event detection methods require predefined event schemas. However, manually defining event schemas is expensive and the coverage of schemas is limited. To this end, some works study the event type induction (ETI) task, which discovers new event types via clustering. However, the setting of ETI suffers from two limitations: event types are not linked into the existing hierarchy and have no semantic names. In this paper, we propose a new research task named Event Ontology Completion (EOC), which aims to simultaneously achieve event clustering, hierarchy expansion and type naming. Furthermore, we develop a Hierarchical Structure Evolution Network (HalTon) for this new task. Specifically, we first devise a Neighborhood Contrastive Clustering module to cluster unlabeled event instances. Then, we propose a Hierarchy-Aware Linking module to incorporate the hierarchical information for event expansion. Finally, we generate meaningful names for new types via an In-Context Learning-based Naming module. Extensive experiments indicate that our method achieves the best performance, outperforming the baselines by 8.23%, 8.79% and 8.10% of ARI score on three datasets.

11:00-12:30 (East Foyer / Virtual)
**#44 LLM-powered Data Augmentation for Enhanced Cross-lingual Performance**
*Chenxi Whitehouse, Monojit Choudhury and Alham Fikri Aji*
This paper explores the potential of leveraging Large Language Models (LLMs) for data augmentation in multilingual commonsense reasoning datasets where the available training data is extremely limited. To achieve this, we utilise several LLMs, namely Dolly-v2, StableVicuna, ChatGPT, and GPT-4, to augment three datasets: XCOPA, XWinograd, and XStoryCloze. Subsequently, we evaluate the effectiveness of fine-tuning smaller multilingual models, mBERT and XLMR, using the synthesised data. We compare the performance of training with data generated in English and target languages, as well as translated English-generated data, revealing the overall advantages of incorporating data generated by LLMs, e.g. a notable 13.4 accuracy score improvement for the best case. Furthermore, we conduct a human evaluation by asking native speakers to assess the naturalness and logical coherence of the generated examples across different languages. The results of the evaluation indicate that LLMs such as ChatGPT and GPT-4 excel at producing natural and coherent text in most languages, however, they struggle to generate meaningful text in certain languages like Tamil. We also observe that ChatGPT falls short in generating plausible alternatives compared to the original dataset, whereas examples from GPT-4 exhibit competitive logical consistency.

11:00-12:30 (East Foyer / Virtual)
**#45 Multi-level Adaptive Contrastive Learning for Knowledge Internalization in Dialogue Generation**
*Chenxu Yang, Zheng Lin, Lanrui Wang, Chong Tian, Liang Pang, Jiangnan Li, Qirong Ho, Yanan Cao and Weiping Wang*
Knowledge-grounded dialogue generation aims to mitigate the issue of text degeneration by incorporating external knowledge to supplement the context. However, the model often fails to internalize this information into responses in a human-like manner. Instead, it simply inserts segments of the provided knowledge into generic responses. As a result, the generated responses tend to be tedious, incoherent, and in lack of interactivity which means the degeneration problem is still unsolved. In this work, we first find that such copying-style degeneration is primarily due to the weak likelihood objective, which allows the model to "cheat" the objective by merely duplicating knowledge segments in a superficial pattern matching based on overlap. To overcome this challenge, we then propose a Multi-level Adaptive Contrastive Learning (MACL) framework that dynamically samples negative examples and subsequently penalizes degeneration behaviors at both the token-level and sequence-level. Extensive experiments on the WoW dataset demonstrate the effectiveness of our approach across various pre-trained models and decoding strategies.

11:00-12:30 (East Foyer / Virtual)
**#46 End-to-end Task-oriented Dialogue: A Survey of Tasks, Methods, and Future Directions**
*Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che and Min Li*
End-to-end task-oriented dialogue (EToD) can directly generate responses in an end-to-end fashion without modular training, which attracts escalating popularity. The advancement of deep neural networks, especially the successful use of large pre-trained models, has further led to significant progress in EToD research in recent years. In this paper, we present a thorough review and provide a unified perspective to summarize existing approaches as well as recent trends to advance the development of EToD research. The contributions of this paper can be summarized: (1) First survey: to our knowledge, we take the first step to present a thorough survey of this research field; (2) New taxonomy: we first introduce a unified perspective for EToD, including (i) Modularly EToD and (ii) Fully EToD; (3) New Frontiers: we discuss some potential frontier areas as well as the corresponding challenges, hoping to spur breakthrough research in EToD field; (4) Abundant resources: we build a public website, where EToD researchers could directly access the recent progress. We hope this work can serve as a thorough reference for the EToD research community.

11:00-12:30 (East Foyer / Virtual)
**#47 KRLS: Improving End-to-End Response Generation in Task Oriented Dialog with Reinforced Keywords Learning**
*Xiao Yu, Qingyang Wu, Kun Qian and Zhou Yu*
In task-oriented dialogs (TOD), reinforcement learning (RL) algorithms train a model to directly optimize response for task-related metrics. However, RL often needs to perform exploration, which can be time-consuming due to the slow auto-regressive sequence generation process. We investigate an approach to create a more efficient RL-based algorithm to improve TOD performance in an offline setting. First, we use a faster generation procedure that samples from independent next-word distributions after training the language model (LM) with supervised learning. We then introduce a fine-grained reward function to help the model focus on learning key information in a dialog, by measuring the importance and semantic closeness of each generated token. Experiments on the MultiWoZ dataset show our new training algorithm, Keywords Reinforcement Learning with Next-word Sampling (KRLS), achieves state-of-the-art performance on the end-to-end response generation task, with a 15% training time reduction compared to a standard RL algorithm using auto-regressive generation.

11:00-12:30 (East Foyer / Virtual)
**#48 COFFEE: Counterfactual Fairness for Personalized Text Generation in Explainable Recommendation**
*Nan Wang, Qifan Wang, Yi-Chia Wang, Maziar Sanjabi, Jingzhou Liu, Hamed Firooz, Hongning Wang and Shaoliang Nie*
As language models become increasingly integrated into our digital lives, Personalized Text Generation (PTG) has emerged as a pivotal component with a wide range of applications. However, the bias inherent in user written text, often used for PTG model training, can inadvertently associate different levels of linguistic quality with users' protected attributes. The model can inherit the bias and perpetuate inequality in generating text w.r.t. users' protected attributes, leading to unfair treatment when serving users. In this work, we investigate fairness of PTG in the context of personalized explanation generation for recommendations. We first discuss the biases in generated explanations and their fairness implications. To promote fairness, we introduce a general framework to achieve measure-specific counterfactual fairness in explanation generation. Extensive experiments and human evaluations demonstrate the effectiveness of our method.

11:00-12:30 (East Foyer / Virtual)
**#49 Multi-Source Probing for Open-Domain Conversational Understanding**
*Yuanxi Li, Hao Zhou, Jie Zhou and Minlie Huang*

Dialogue comprehension and generation are vital to the success of open-domain dialogue systems. Although pre-trained generative conversation models have made significant progress in generating fluent responses, people have difficulty judging whether they understand and efficiently model the contextual information of the conversation. In this study, we propose a Multi-Source Probing (MSP) method to probe the dialogue comprehension abilities of open-domain dialogue models. MSP aggregates features from multiple sources to accomplish diverse task goals and conducts downstream tasks in a generative manner that is consistent with dialogue model pre-training to leverage model capabilities. We conduct probing experiments on seven tasks that require various dialogue comprehension skills, based on the internal representations encoded by dialogue models. Experimental results show that open-domain dialogue models can encode semantic information in the intermediate hidden states, which facilitates dialogue comprehension tasks. Models of different scales and structures possess different conversational understanding capabilities. Our findings encourage a comprehensive evaluation and design of open-domain dialogue models.

11:00-12:30 (East Foyer / Virtual)
**#50 Enhancing Textbooks with Visuals from the Web for Improved Learning**
*Janvijay Singh, Vilém Zouhar and Mrinmaya Sachan*
Textbooks are one of the main mediums for delivering high-quality education to students. In particular, explanatory and illustrative visuals play a key role in retention, comprehension and general transfer of knowledge. However, many textbooks lack these interesting visuals to support student learning. In this paper, we investigate the effectiveness of vision-language models to automatically enhance textbooks with images from the web. We collect a dataset of e-textbooks in the math, science, social science and business domains. We then set up a text-image matching task that involves retrieving and appropriately assigning web images to textbooks, which we frame as a matching optimization problem. Through a crowd-sourced evaluation, we verify that (1) while the original textbook images are rated higher, automatically assigned ones are not far behind, and (2) the precise formulation of the optimization problem matters. We release the dataset of textbooks with an associated image bank to inspire further research in this intersectional area of computer vision and NLP for education.

11:00-12:30 (East Foyer / Virtual)
**#51 Dual-Feedback Knowledge Retrieval for Task-Oriented Dialogue Systems**
*Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan and Qifan Wang*
Efficient knowledge retrieval plays a pivotal role in ensuring the success of end-to-end task-oriented dialogue systems by facilitating the selection of relevant information necessary to fulfill user requests. However, current approaches generally integrate knowledge retrieval and response generation, which poses scalability challenges when dealing with extensive knowledge bases. Taking inspiration from open-domain question answering, we propose a retriever-generator architecture that harnesses a retriever to retrieve pertinent knowledge and a generator to generate system responses. Due to the lack of retriever training labels, we propose relying on feedback from the generator as pseudo-labels to train the retriever. To achieve this, we introduce a dual-feedback mechanism that generates both positive and negative feedback based on the output of the generator. Our method demonstrates superior performance in task-oriented dialogue tasks, as evidenced by experimental results on three benchmark datasets.

11:00-12:30 (East Foyer / Virtual)
**#52 Evaluating Object Hallucination in Large Vision-Language Models**
*Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao and Ji-Rong Wen*
Inspired by the superior language abilities of large language models (LLM), large vision-language models (LVLM) have been recently proposed by integrating powerful LLMs for improving the performance on complex multimodal tasks. Despite the promising progress on LVLMs, we find that they suffer from object hallucinations, i.e., they tend to generate objects inconsistent with the target images in the descriptions. To investigate it, this work presents the first systematic study on object hallucination of LVLMs. We conduct the evaluation experiments on several representative LVLMs, and show that they mostly suffer from severe object hallucination issues. We further discuss that the visual instructions may influence the hallucination, and find that: objects that frequently appear in the visual instructions or co-occur with the image objects are obviously prone to be hallucinated by LVLMs. Besides, we further design a polling-based query method called POPE for better evaluation of object hallucination. Experiment results show that our POPE can evaluate object hallucination in a more stable and flexible way.

11:00-12:30 (East Foyer / Virtual)
**#53 DueT: Image-Text Contrastive Transfer Learning with Dual-adapter Tuning**
*Taku Hasegawa, Kyosuke Nishida, Koki Maeda and Kuniko Saito*
This paper presents DueT, a novel transfer learning method for vision and language models built by contrastive learning. In DueT, adapters are inserted into the image and text encoders, which have been initialized using models pre-trained on uni-modal corpora and then frozen. By training only these adapters, DueT enables efficient learning with a reduced number of trainable parameters. Moreover, unlike traditional adapters, those in DueT are equipped with a gating mechanism, enabling effective transfer and connection of knowledge acquired from pre-trained uni-modal encoders while preventing catastrophic forgetting. We report that DueT outperformed simple fine-tuning, the conventional method fixing only the image encoder and training only the text encoder, and the LoRA-based adapter method in accuracy and parameter efficiency for 0-shot image and text retrieval in both English and Japanese domains.

11:00-12:30 (East Foyer / Virtual)
**#54 BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations**
*Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia and Rui Yan*
Recent advancements in biological research leverage the integration of molecules, proteins, and natural language to enhance drug discovery. However, current models exhibit several limitations, such as the generation of invalid molecular SMILES, underutilization of contextual information, and equal treatment of structured and unstructured knowledge. To address these issues, we propose BioT5, a comprehensive pre-training framework that enriches cross-modal integration in biology with chemical knowledge and natural language associations. BioT5 utilizes SELFIES for 100% robust molecular representations and extracts knowledge from the surrounding context of bio-entities in unstructured biological literature. Furthermore, BioT5 distinguishes between structured and unstructured knowledge, leading to more effective utilization of information. After fine-tuning, BioT5 shows superior performance across a wide range of tasks, demonstrating its strong capability of capturing underlying relations and properties of bio-entities. Our code is available at https://github.com/QizhiPei/BioT5.

11:00-12:30 (East Foyer / Virtual)
**#55 Target-oriented Proactive Dialogue Systems with Personalization: Problem Formulation and Dataset Curation**
*Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong and Wenjie Li*
Target-oriented dialogue systems, designed to proactively steer conversations toward predefined targets or accomplish specific system-side goals, are an exciting area in conversational AI. In this work, by formulating a <dialogue act, topic> pair as the conversation target, we explore a novel problem of personalized target-oriented dialogue by considering personalization during the target accomplishment process. However, there remains an emergent need for high-quality datasets, and building one from scratch requires tremendous human effort. To address this, we propose an automatic dataset curation framework using a role-playing approach. Based on this framework, we construct a large-scale personalized target-oriented dialogue dataset, TopDial, which comprises about 18K multi-turn dialogues. The experimental results show that this dataset is of high quality and could contribute to exploring personalized target-oriented dialogue.

11:00-12:30 (East Foyer / Virtual)
**#56 Multi-Source Multi-Type Knowledge Exploration and Exploitation for Dialogue Generation**
*Xuanfan Ni, Hongliang Dai, Zhaochun Ren and Piji Li*
Open-domain multi-turn dialogue generation encounters the significant challenge of lacking various types of knowledge from diverse sources. Existing models typically focus on identifying specific types of dialogue knowledge and utilize corresponding datasets for training. However, this approach often leads to limited generalization capabilities and increased computational resource requirements. Recently, large language models (LLMs) have shown impressive performance on natural language processing tasks. To harness the knowledge storage of LLMs, we propose a framework named KnowEE that explores multi-source multi-type knowledge from LLMs by leveraging diverse datasets and then exploits the obtained knowledge for response generation. Our framework comprises two phases: First, we leverage five external datasets encompassing various types of knowledge to extract the most relevant samples to the dialogue context which are served as prompts to generate corresponding type of knowledge; Second, we inject the acquired knowledge into the ongoing dialogue context in fine-grained and coarse-grained manners, which is then fed into LLMs to generate the final dialogue response. Both automatic and manual evaluation results validate the effectiveness of our framework in exploring and exploiting multi-source multi-type knowledge to generate coherent, informative, and fluent responses.

11:00-12:30 (East Foyer / Virtual)
**#57 EDIS: Entity-Driven Image Search over Multimodal Web Content**
*Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen and William Yang Wang*
Making image retrieval methods practical for real-world search applications requires significant progress in dataset scales, entity comprehension, and multimodal information fusion. In this work, we introduce Entity-Driven Image Search (EDIS), a challenging dataset for cross-modal image search in the news domain. EDIS consists of 1 million web images from actual search engine results and curated datasets, with each image paired with a textual description. Unlike datasets that assume a small set of single-modality candidates, EDIS reflects real-world web image search scenarios by including a million multimodal image-text pairs as candidates. EDIS encourages the development of retrieval models that simultaneously address cross-modal information fusion and matching. To achieve accurate ranking results, a model must: 1) understand named entities and events from text queries, 2) ground entities onto images or text descriptions, and 3) effectively fuse textual and visual representations. Our experimental results show that EDIS challenges state-of-the-art methods with dense entities and the large-scale candidate set. The ablation study also proves that fusing textual features with visual features is critical in improving retrieval results.

11:00-12:30 (East Foyer / Virtual)
**#58 Fine-grained Conversational Decoding via Isotropic and Proximal Search**
*Yuxuan Yao, Han Wu, Qiling Xu and Linqi Song*
General-purpose text decoding approaches are usually adopted for dialogue response generation. Although the quality of the generated responses can be improved with dialogue-specific encoding methods, conversational decoding methods are still under-explored. Inspired by SimDRC that a good dialogue feature space should follow the rules of locality and isotropy, we present a fine-grained conversational decoding method, termed isotropic and proximal search (IPS). Our method is designed to generate the semantic-concentrated response, while still maintaining informativeness and discrimination against the context. Experiments show that our approach significantly outperforms existing decoding strategies in the dialogue field across both automatic and human evaluation metrics. More in-depth analyses further confirm the effectiveness of our approach.

11:00-12:30 (East Foyer / Virtual)
**#59 RobustGEC: Robust Grammatical Error Correction Against Subtle Context Perturbation**
*Yue Zhang, Leyang Cui, Enbo Zhao, Wei Bi and Shuming Shi*
Grammatical Error Correction (GEC) systems play a vital role in assisting people with their daily writing tasks. However, users may sometimes come across a GEC system that initially performs well but fails to correct errors when the inputs are slightly modified. To ensure an ideal user experience, a reliable GEC system should have the ability to provide consistent and accurate suggestions when encountering irrelevant context perturbations, which we refer to as context robustness. In this paper, we introduce RobustGEC, a benchmark designed to evaluate the context robustness of GEC systems. RobustGEC comprises 5,000 GEC cases, each with one original error-correct sentence pair and five variants carefully devised by human annotators. Utilizing RobustGEC, we reveal that state-of-the-art GEC systems still lack sufficient robustness against context perturbations. Moreover, we propose a simple yet effective method for remitting this issue.

11:00-12:30 (East Foyer / Virtual)
**#60 Reinforced Target-driven Conversational Promotion**
*Huy Quang Dao, Lizi Liao, Dung D. Le and Yuxiang Nie*
The ability to proactively engage with users towards pitching products is highly desired for conversational assistants. However, existing conversational recommendation methods overemphasize on acquiring user preferences while ignore the strategic planning for nudging users towards accepting a designated item. Hence, these methods fail to promote specified items with engaging responses. In this work, we propose a Reinforced Target-driven Conversational Promotion (RTCP) framework for conversational promotion. RTCP integrates short-term and long-term planning via a balanced gating mechanism. Inside which, the dialogue actions are predicted via a knowledge-integrated multi-head attention and guided via reinforcement learning rewards. RTCP then employs action-guided prefix tuning to generate relevant responses. Experimental results demonstrate that our model outperforms state-of-the-art models on both automatic metrics and human evaluation. Moreover, RTCP has a strong capability in quickly adapting to unseen scenarios just by updating prefix parameters without re-training the whole model.

11:00-12:30 (East Foyer / Virtual)
**#61 Polar Ducks and Where to Find Them: Enhancing Entity Linking with Duck Typing and Polar Box Embeddings**
*Mattia Atzeni, Mikhail Plekhanov, Frederic A Dreyer, Nora Kassner, Simone Merello, Louis Martin and Nicola Cancedda*
Entity linking methods based on dense retrieval are widely adopted in large-scale applications for their efficiency, but they can fall short of generative models, as they are sensitive to the structure of the embedding space. To address this issue, this paper introduces DUCK, an approach to infusing structural information in the space of entity representations, using prior knowledge of entity types. Inspired by duck typing in programming languages, we define the type of an entity based on its relations with other entities in a knowledge graph. Then, porting the concept of box embeddings to spherical polar coordinates, we represent relations as boxes on the hypersphere. We optimize the model to place entities inside the boxes corresponding to their relations, thereby clustering together entities of similar type. Our experiments show that our method sets new state-of-the-art results on standard entity-disambiguation benchmarks. It improves the performance of the model by up to 7.9 F1 points, outperforms other type-aware approaches, and matches the results of generative models with 18 times more parameters.

11:00-12:30 (East Foyer / Virtual)
**#62 ScanDL: A Diffusion Model for Generating Synthetic Scanpaths on Texts**
*Lena Sophia Bolliger, David Robert Reich, Patrick Haller, Deborah Noemie Jakobi, Paul Prasse and Lena Ann Jäger*

Eye movements in reading play a crucial role in psycholinguistic research studying the cognitive mechanisms underlying human language processing. More recently, the tight coupling between eye movements and cognition has also been leveraged for language-related machine learning tasks such as the interpretability, enhancement, and pre-training of language models, as well as the inference of reader- and text-specific properties. However, scarcity of eye movement data and its unavailability at application time poses a major challenge for this line of research. Initially, this problem was tackled by resorting to cognitive models for synthesizing eye movement data. However, for the sole purpose of generating human-like scanpaths, purely data-driven machine-learning-based methods have proven to be more suitable. Following recent advances in adapting diffusion processes to discrete data, we propose ScanDL, a novel discrete sequence-to-sequence diffusion model that generates synthetic scanpaths on texts. By leveraging pre-trained word representations and jointly embedding both the stimulus text and the fixation sequence, our model captures multi-modal interactions between the two inputs. We evaluate ScanDL within- and across-dataset and demonstrate that it significantly outperforms state-of-the-art scanpath generation methods. Finally, we provide an extensive psycholinguistic analysis that underlines the model's ability to exhibit human-like reading behavior. Our implementation is made available at https://github.com/DiLi-Lab/ScanDL.

11:00-12:30 (East Foyer / Virtual)
**#63 Did You Mean...? Confidence-based Trade-offs in Semantic Parsing**
*Elias Stengel-Eskin and Benjamin Van Durme*
We illustrate how a calibrated model can help balance common trade-offs in task-oriented parsing. In a simulated annotator-in-the-loop experiment, we show that well-calibrated confidence scores allow us to balance cost with annotator load, improving accuracy with a small number of interactions. We then examine how confidence scores can help optimize the trade-off between usability and safety. We show that confidence-based thresholding can substantially reduce the number of incorrect low-confidence programs executed; however, this comes at a cost to usability. We propose the DidYouMean system which better balances usability and safety by rephrasing low-confidence inputs.

11:00-12:30 (East Foyer / Virtual)
**#64 How do languages influence each other? Studying cross-lingual data sharing during LM fine-tuning**
*Rochelle Choenni, Dan Garrette and Ekaterina Shutova*
Multilingual language models (MLMs) are jointly trained on data from many different languages such that representation of individual languages can benefit from other languages' data. Impressive performance in zero-shot cross-lingual transfer shows that these models are able to exploit this property. Yet, it remains unclear to what extent, and under which conditions, languages rely on each other's data. To answer this question, we use TracIn (Pruthi et al., 2020), a training data attribution (TDA) method, to retrieve training samples from multilingual data that are most influential for test predictions in a given language. This allows us to analyse cross-lingual sharing mechanisms of MLMs from a new perspective. While previous work studied cross-lingual sharing at the model parameter level, we present the first approach to study it at the data level. We find that MLMs rely on data from multiple languages during fine-tuning and this reliance increases as fine-tuning progresses. We further find that training samples from other languages can both reinforce and complement the knowledge acquired from data of the test language itself.

11:00-12:30 (East Foyer / Virtual)
**#65 Text Embeddings Reveal (Almost) As Much As Text**
*John Xavier Morris, Volodymyr Kuleshov, Vitaly Shmatikov and Alexander M Rush*
How much private information do text embeddings reveal about the original text? We investigate the problem of embedding *inversion*, reconstructing the full text represented in dense text embeddings. We frame the problem as controlled generation: generating text that, when reembedded, is close to a fixed point in latent space. We find that although a naive model conditioned on the embedding performs poorly, a multi-step method that iteratively corrects and re-embeds text is able to recover 92% of 32-token text inputs exactly. We train our model to decode text embeddings from two state-of-the-art embedding models, and also show that our model can recover important personal information (full names) from a dataset of clinical notes.

11:00-12:30 (East Foyer / Virtual)
**#66 Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement**
*Rosamond Elizabeth Thalken, Edward Stiglitz, David Mimno and Matthew Wilkens*
Generative language models (LMs) are increasingly used for document class-prediction tasks and promise enormous improvements in cost and efficiency. Existing research often examines simple classification tasks, but the capability of LMs to classify on complex or specialized tasks is less well understood. We consider a highly complex task that is challenging even for humans: the classification of legal reasoning according to jurisprudential philosophy. Using a novel dataset of historical United States Supreme Court opinions annotated by a team of domain experts, we systematically test the performance of a variety of LMs. We find that generative models perform poorly when given instructions (i.e. prompts) equal to the instructions presented to human annotators through our codebook. Our strongest results derive from fine-tuning models on the annotated dataset; the best performing model is an in-domain model, LEGAL-BERT. We apply predictions from this fine-tuned model to study historical trends in jurisprudence, an exercise that both aligns with prominent qualitative historical accounts and points to areas of possible refinement in those accounts. Our findings generally sound a note of caution in the use of generative LMs on complex tasks without fine-tuning and point to the continued relevance of human annotation-intensive classification methods.

11:00-12:30 (East Foyer / Virtual)
**#67 An Expression Tree Decoding Strategy for Mathematical Equation Generation**
*Wenqi Zhang, Yongliang Shen, Qingpeng Nong, Zeqi Tan, Yanna Ma and Weiming Lu*
Generating mathematical equations from natural language requires an accurate understanding of the relations among math expressions. Existing approaches can be broadly categorized into token-level and expression-level generation. The former treats equations as a mathematical language, sequentially generating math tokens. Expression-level methods generate each expression one by one. However, each expression represents a solving step, and there naturally exist parallel or dependent relations between these steps, which are ignored by current sequential methods. Therefore, we integrate tree structure into the expression-level generation and advocate an expression tree decoding strategy. To generate a tree with expression as its node, we employ a layer-wise parallel decoding strategy: we decode multiple independent expressions (leaf nodes) in parallel at each layer and repeat parallel decoding layer by layer to sequentially generate these parent node expressions that depend on others. Besides, a bipartite matching algorithm is adopted to align multiple predictions with annotations for each layer. Experiments show our method outperforms other baselines, especially for these equations with complex structures.

11:00-12:30 (East Foyer / Virtual)
**#68 BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases**
*Yiming Zhang, Sravani Uttara Nanduri, Liwei Jiang, Tongshuang Wu and Maarten Sap*
Toxicity annotators and content moderators often default to mental shortcuts when making decisions. This can lead to subtle toxicity being missed, and seemingly toxic but harmless content being over-detected. We introduce BiasX, a framework that enhances content moderation setups with free-text explanations of statements' implied social biases, and explore its effectiveness through a large-scale crowdsourced user study. We show that indeed, participants substantially benefit from explanations for correctly identifying subtly (non-)toxic content. The

quality of explanations is critical: imperfect machine-generated explanations (+2.4% on hard toxic examples) help less compared to expert-written human explanations (+7.2%). Our results showcase the promise of using free-text explanations to encourage more thoughtful toxicity moderation.

11:00-12:30 (East Foyer / Virtual)
**#69 APoLLo : Unified Adapter and Prompt Learning for Vision Language Models**
*Sanjoy Chowdhury, Sayan Nag and Dinesh Manocha*
The choice of input text prompt plays a critical role in the performance of Vision-Language Pretrained (VLP) models such as CLIP. We present APoLLo, a unified multi-modal approach that combines Adapter and Prompt learning for Vision-Language models. Our method is designed to substantially improve the generalization capabilities of VLP models when they are fine-tuned in a few-shot setting. We introduce trainable cross-attention-based adapter layers in conjunction with vision and language encoders to strengthen the alignment between the two modalities. We enforce consistency between the respective encoder branches (receiving augmented inputs) to prevent overfitting in downstream tasks. Our method is evaluated on three representative tasks: generalization to novel classes, cross-dataset evaluation, and unseen domain shifts. In practice, APoLLo achieves a relative gain up to 6.03% over MaPLe (SOTA) on novel classes for 10 diverse image recognition datasets.

11:00-12:30 (East Foyer / Virtual)
**#70 Human Learning by Model Feedback: The Dynamics of Iterative Prompting with Midjourney**
*Shachar Don-Yehiya, Leshem Choshen and Omri Abend*
Generating images with a Text-to-Image model often requires multiple trials, where human users iteratively update their prompt based on feedback, namely the output image. Taking inspiration from cognitive work on reference games and dialogue alignment, this paper analyzes the dynamics of the user prompts along such iterations. We compile a dataset of iterative interactions of human users with Midjourney. Our analysis then reveals that prompts predictably converge toward specific traits along these iterations. We further study whether this convergence is due to human users, realizing they missed important details, or due to adaptation to the model's "preferences", producing better images for a specific language style. We show initial evidence that both possibilities are at play. The possibility that users adapt to the model's preference raises concerns about reusing user data for further training. The prompts may be biased towards the preferences of a specific model, rather than align with human intentions and natural manner of expression.

11:00-12:30 (East Foyer / Virtual)
**#71 PALS: Personalized Active Learning for Subjective Tasks in NLP**
*Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon and Przemyslaw Kazienko*
For subjective NLP problems, such as classification of hate speech, aggression, or emotions, personalized solutions can be exploited. Then, the learned models infer about the perception of the content independently for each reader. To acquire training data, texts are commonly randomly assigned to users for annotation, which is expensive and highly inefficient. Therefore, for the first time, we suggest applying an active learning paradigm in a personalized context to better learn individual preferences. It aims to alleviate the labeling effort by selecting more relevant training samples. In this paper, we present novel Personalized Active Learning techniques for Subjective NLP tasks (PALS) to either reduce the cost of the annotation process or to boost the learning effect. Our five new measures allow us to determine the relevance of a text in the context of learning users personal preferences. We validated them on three datasets: Wiki discussion texts individually labeled with aggression and toxicity, and on Unhealthy Conversations dataset. Our PALS techniques outperform random selection even by more than 30%. They can also be used to reduce the number of necessary annotations while maintaining a given quality level. Personalized annotation assignments based on our controversy measure decrease the amount of data needed to just 25%-40% of the initial size.

11:00-12:30 (East Foyer / Virtual)
**#72 Symbolic Planning and Code Generation for Grounded Dialogue**
*Justin T Chiu, Wenting Zhao, Derek Chen, Saujas Vaduguru, Alexander M Rush and Daniel Fried*
Large language models (LLMs) excel at processing and generating text and code. However, LLMs have had limited applicability in grounded task-oriented dialogue as they are difficult to steer toward task objectives and fail to handle novel grounding. We present a modular and interpretable grounded dialogue system that addresses these shortcomings by composing LLMs with a symbolic planner and grounded code execution. Our system, consists of a reader and planner: the reader leverages an LLM to convert partner utterances into executable code, calling functions that perform grounding. The translated code's output is stored to track dialogue state, while a symbolic planner determines the next appropriate response. We evaluate our system's performance on the demanding OneCommon dialogue task, involving collaborative reference resolution on abstract images of scattered dots. Our system substantially outperforms the previous state-of-the-art, including improving task success in human evaluations from 56% to 69% in the most challenging setting.

11:00-12:30 (East Foyer / Virtual)
**#73 Global Voices, Local Biases: Socio-Cultural Prejudices across Languages**
*Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu and Antonios Anastasopoulos*
Human biases are ubiquitous but not uniform: disparities exist across linguistic, cultural, and societal borders. As large amounts of recent literature suggest, language models (LMs) trained on human data can reflect and often amplify the effects of these social biases. However, the vast majority of existing studies on bias are heavily skewed towards Western and European languages. In this work, we scale the Word Embedding Association Test (WEAT) to 24 languages, enabling broader studies and yielding interesting findings about LM bias. We additionally enhance this data with culturally relevant information for each language, capturing local contexts on a global scale. Further, to encompass more widely prevalent societal biases, we examine new bias dimensions across toxicity, ableism, and more. Moreover, we delve deeper into the Indian linguistic landscape, conducting a comprehensive regional bias analysis across six prevalent Indian languages. Finally, we highlight the significance of these social biases and the new dimensions through an extensive comparison of embedding methods, reinforcing the need to address them in pursuit of more equitable language models.

11:00-12:30 (East Foyer / Virtual)
**#74 XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models**
*Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer and Madian Khabsa*
Large multilingual language models typically rely on a single vocabulary shared across 100+ languages. As these models have increased in parameter count and depth, vocabulary size has remained largely unchanged. This *vocabulary bottleneck* limits the representational capabilities of multilingual models like XLM-R. In this paper, we introduce a new approach for scaling to very large multilingual vocabularies by de-emphasizing token sharing between languages with little lexical overlap and assigning vocabulary capacity to achieve sufficient coverage for each individual language. Tokenizations using our vocabulary are typically more semantically meaningful and shorter compared to XLM-R. Leveraging this improved vocabulary, we train XLM-V, a multilingual language model with a one million token vocabulary. XLM-V outperforms XLM-R on every task we tested on ranging from natural language inference (XNLI), question answering (MLQA, XQuAD, TyDiQA), to named entity recognition (WikiAnn). XLM-V is particularly effective on low-resource language tasks and outperforms XLM-R by 11.2% and 5.8% absolute on MasakhaNER and Americas NLI, respectively.

11:00-12:30 (East Foyer / Virtual)
**#75 Transfer-Free Data-Efficient Multilingual Slot Labeling**
*Evgeniia Razumovskaia, Ivan Vulić and Anna Korhonen*
Slot labeling (SL) is a core component of task-oriented dialogue (TOD) systems, where slots and corresponding values are usually language-, task- and domain-specific. Therefore, extending the system to any new language-domain-task configuration requires (re)running an expensive and resource-intensive data annotation process. To mitigate the inherent data scarcity issue, current research on multilingual ToD assumes that sufficient English-language annotated data are always available for particular tasks and domains, and thus operates in a standard cross-lingual transfer setup. In this work, we depart from this often unrealistic assumption. We examine challenging scenarios where such transfer-enabling English annotated data cannot be guaranteed, and focus on bootstrapping multilingual data-efficient slot labelers in transfer-free scenarios directly in the target languages without any English-ready data. We propose a two-stage slot labeling approach (termed TWOSL) which transforms standard multilingual sentence encoders into effective slot labelers. In Stage 1, relying on SL-adapted contrastive learning with only a handful of SL-annotated examples, we turn sentence encoders into task-specific span encoders. In Stage 2, we recast SL from a token classification into a simpler, less data-intensive span classification task. Our results on two standard multilingual TOD datasets and across diverse languages confirm the effectiveness and robustness of TWOSL. It is especially effective for the most challenging transfer-free few-shot setups, paving the way for quick and data-efficient bootstrapping of multilingual slot labelers for TOD.

11:00-12:30 (East Foyer / Virtual)
**#76 A Systematic Study of Performance Disparities in Multilingual Task-Oriented Dialogue Systems**
*Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen and Ivan Vulić*
Achieving robust language technologies that can perform well across the world's many languages is a central goal of multilingual NLP. In this work, we take stock of and empirically analyse task performance disparities that exist between multilingual task-oriented dialogue (ToD) systems. We first define new quantitative measures of absolute and relative equivalence in system performance, capturing disparities across languages and within individual languages. Through a series of controlled experiments, we demonstrate that performance disparities depend on a number of factors: the nature of the ToD task at hand, the underlying pretrained language model, the target language, and the amount of ToD annotated data. We empirically prove the existence of the adaptation and intrinsic biases in current ToD systems: e.g., ToD systems trained for Arabic or Turkish using annotated ToD data fully parallel to English ToD data still exhibit diminished ToD task performance. Beyond providing a series of insights into the performance disparities of ToD systems in different languages, our analyses offer practical tips on how to approach ToD data collection and system development for new languages.

11:00-12:30 (East Foyer / Virtual)
**#77 DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules**
*Yanchen Liu, William Barr Held and Diyi Yang*
Existing large language models (LLMs) that mainly focus on Standard American English (SAE) often lead to significantly worse performance when being applied to other English dialects. While existing mitigations tackle discrepancies for individual target dialects, they assume access to high-accuracy dialect identification systems. The boundaries between dialects are inherently flexible, making it difficult to categorize language into discrete predefined categories. In this paper, we propose DADA (Dialect Adaptation via Dynamic Aggregation), a modular approach to imbue SAE-trained models with multi-dialectal robustness by composing adapters which handle specific linguistic features. The compositional architecture of DADA allows for both targeted adaptation to specific dialect variants and simultaneous adaptation to various dialects. We show that DADA is effective for both single task and instruction finetuned language models, offering an extensible and interpretable framework for adapting existing LLMs to different English dialects.

11:00-12:30 (East Foyer / Virtual)
**#78 QA-NatVer: Question Answering for Natural Logic-based Fact Verification**
*Rami Aly, Marek Strong and Andreas Vlachos*
Fact verification systems assess a claim's veracity based on evidence. An important consideration in designing them is faithfulness, i.e. generating explanations that accurately reflect the reasoning of the model. Recent works have focused on natural logic, which operates directly on natural language by capturing the semantic relation of spans between an aligned claim with its evidence via set-theoretic operators. However, these approaches rely on substantial resources for training, which are only available for high-resource languages. To this end, we propose to use question answering to predict natural logic operators, taking advantage of the generalization capabilities of instruction-tuned language models. Thus, we obviate the need for annotated training data while still relying on a deterministic inference system. In a few-shot setting on FEVER, our approach outperforms the best baseline by 4.3 accuracy points, including a state-of-the-art pre-trained seq2seq natural logic system, as well as a state-of-the-art prompt-based classifier. Our system demonstrates its robustness and portability, achieving competitive performance on a counterfactual dataset and surpassing all approaches without further annotation on a Danish verification dataset. A human evaluation indicates that our approach produces more plausible proofs with fewer erroneous natural logic operators than previous natural logic-based systems.

11:00-12:30 (East Foyer / Virtual)
**#79 Can Language Models Laugh at YouTube Short-form Videos?**
*Dayoon Ko, Sangho Lee and Gunhee Kim*
As short-form funny videos on social networks are gaining popularity, it becomes demanding for AI models to understand them for better communication with humans. Unfortunately, previous video humor datasets target specific domains such as speeches or sitcoms, and mostly focus on verbal cues. We curate a user-generated dataset of 10K multimodal funny videos from YouTube, called ExFunTube. Using a video filtering pipeline with GPT-3.5, we verify both verbal and visual elements contributing to humor. After filtering, we annotate each video with timestamps and text explanations for funny moments. Our ExFunTube is unique over existing datasets in that our videos cover a wide range of domains with various types of humor that necessitate a multimodal understanding of the content. Also, we develop a zero-shot video-to-text prompting to maximize video humor understanding of large language models (LLMs). With three different evaluation methods using automatic scores, rationale quality experiments, and human evaluations, we show that our prompting significantly improves LLMs' ability for humor explanation.

11:00-12:30 (East Foyer / Virtual)
**#80 Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?**
*Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter and Ming-Wei Chang*
Pre-trained vision and language models have demonstrated state-of-the-art capabilities over existing tasks involving images and texts, including visual question answering. However, it remains unclear whether these models possess the capability to answer questions that are not only querying visual content but knowledge-intensive and information-seeking. In this study, we introduce InfoSeek, a visual question answering dataset tailored for information-seeking questions that cannot be answered with only common sense knowledge. Using InfoSeek, we analyze various pre-trained visual question answering models and gain insights into their characteristics. Our findings reveal that state-of-the-art pre-trained multi-modal models (e.g., PaLI-X, BLIP2, InstructBLIP) face challenges in answering visual information-seeking questions, but fine-tuning on the InfoSeek dataset elicits models to use fine-grained knowledge that was learned during pre-training. Furthermore, we show

that accurate visual entity recognition can be used to improve performance on InfoSeek by retrieving relevant documents, showing a significant space for improvement.

11:00-12:30 (East Foyer / Virtual)
**#81 SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables**
*Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov and Min-Yen Kan*
Current scientific fact-checking benchmarks exhibit several shortcomings, such as biases arising from crowd-sourced claims and an over-reliance on text-based evidence. We present SCITAB, a challenging evaluation dataset consisting of 1.2K expert-verified scientific claims that 1) originate from authentic scientific publications and 2) require compositional reasoning for verification. The claims are paired with evidence-containing scientific tables annotated with labels. Through extensive evaluations, we demonstrate that SCITAB poses a significant challenge to state-of-the-art models, including table-based pretraining models and large language models. All models except GPT-4 achieved performance barely above random guessing. Popular prompting techniques, such as Chain-of-Thought, do not achieve much performance gains on SCITAB. Our analysis uncovers several unique challenges posed by SCITAB, including table grounding, claim ambiguity, and compositional reasoning. Our codes and data are publicly available at https://github.com/XinyuanLu00/SciTab.

11:00-12:30 (East Foyer / Virtual)
**#82 Causal Reasoning through Two Cognition Layers for Improving Generalization in Visual Question Answering**
*Trang Nguyen and Naoaki Okazaki*
Generalization in Visual Question Answering (VQA) requires models to answer questions about images with contexts beyond the training distribution. Existing attempts primarily refine unimodal aspects, overlooking enhancements in multimodal aspects. Besides, diverse interpretations of the input lead to various modes of answer generation, highlighting the role of causal reasoning between interpreting and answering steps in VQA. Through this lens, we propose Cognitive pathways VQA (CopVQA) improving the multimodal predictions by emphasizing causal reasoning factors. CopVQA first operates a pool of pathways that capture diverse causal reasoning flows through interpreting and answering stages. Mirroring human cognition, we decompose the responsibility of each stage into distinct experts and a cognition-enabled component (CC). The two CCs strategically execute one expert for each stage at a time. Finally, we prioritize answer predictions governed by pathways involving both CCs while disregarding answers produced by either CC, thereby emphasizing causal reasoning and supporting generalization. Our experiments on real-life and medical data consistently verify that CopVQA improves VQA performance and generalization across baselines and domains. Notably, CopVQA achieves a new state-of-the-art (SOTA) on the PathVQA dataset and comparable accuracy to the current SOTA on VQA-CPv2, VQAv2, and VQA- RAD, with one-fourth of the model size.

11:00-12:30 (East Foyer / Virtual)
**#83 Log-FGAER: Logic-Guided Fine-Grained Address Entity Recognition from Multi-Turn Spoken Dialogue**
*Xue Han, Yitong Wang, Qian Hu, Pengwei Hu, Chao Deng and Junlan Feng*
Fine-grained address entity recognition (FGAER) from multi-turn spoken dialogues is particularly challenging. The major reason lies in that a full address is often formed through a conversation process. Different parts of an address are distributed through multiple turns of a dialogue with spoken noises. It is nontrivial to extract by turn and combine them. This challenge has not been well emphasized by main-stream entity extraction algorithms. To address this issue, we propose in this paper a logic-guided fine-grained address recognition method (Log-FGAER), where we formulate the address hierarchy relationship as the logic rule and softly apply it in a probabilistic manner to improve the accuracy of FGAER. In addition, we provide an ontology-based data augmentation methodology that employs ChatGPT to augment a spoken dialogue dataset with labeled address entities. Experiments are conducted using datasets generated by the proposed data augmentation technique and derived from real-world scenarios. The results of the experiment demonstrate the efficacy of our proposal.

11:00-12:30 (East Foyer / Virtual)
**#84 Document-level Relationship Extraction by Bidirectional Constraints of Beta Rules**
*Yichun Liu, Zizhong Zhu, Xiaowang Zhang, Zhiyong Feng, Daoqi Chen and Yaxin Li*
Document-level Relation Extraction (DocRE) aims to extract relations among entity pairs in documents. Some works introduce logic constraints into DocRE, addressing the issues of opacity and weak logic in original DocRE models. However, they only focus on forward logic constraints and the rules mined in these works often suffer from pseudo rules with high standard-confidence but low support. In this paper, we proposes Bidirectional Constraints of Beta Rules(BCBR), a novel logic constraint framework. BCBR first introduces a new rule miner which model rules by beta contribtion. Then forward and reverse logic constraints are constructed based on beta rules. Finally, BCBR reconstruct rule consistency loss by bidirectional constraints to regulate the output of the DocRE model. Experiments show that BCBR outperforms original DocRE models in terms of relation extraction performance ($\sim$2.7 F1 score) and logical consistency($\sim$3.1 logic score). Furthermore, BCBR consistently outperforms two other logic constraint frameworks.

11:00-12:30 (East Foyer / Virtual)
**#85 LACMA: Language-Aligning Contrastive Learning with Meta-Actions for Embodied Instruction Following**
*Cheng-Fu Yang, Yen-Chun Chen, Jianwei Yang, Xiyang Dai, Lu Yuan, Yu-Chiang Frank Wang and Kai-Wei Chang*
End-to-end Transformers have demonstrated an impressive success rate for Embodied Instruction Following when the environment has been seen in training. However, they tend to struggle when deployed in an unseen environment. This lack of generalizability is due to the agent's insensitivity to subtle changes in natural language instructions. To mitigate this issue, we propose explicitly aligning the agent's hidden states with the instructions via contrastive learning. Nevertheless, the semantic gap between high-level language instructions and the agent's low-level action space remains an obstacle. Therefore, we further introduce a novel concept of meta-actions to bridge the gap. Meta-actions are ubiquitous action patterns that can be parsed from the original action sequence. These patterns represent higher-level semantics that are intuitively aligned closer to the instructions. When meta-actions are applied as additional training signals, the agent generalizes better to unseen environments. Compared to a strong multi-modal Transformer baseline, we achieve a significant 4.5% absolute gain in success rate in unseen environments of ALFRED Embodied Instruction Following. Additional analysis shows that the contrastive objective and meta-actions are complementary in achieving the best results, and the resulting agent better aligns its states with corresponding instructions, making it more suitable for real-world embodied agents.

11:00-12:30 (East Foyer / Virtual)
**#86 Language Model is Suitable for Correction of Handwritten Mathematical Expressions Recognition**
*Zui Chen, Jiaqi Han, Chaofan Yang and Yi Zhou*
Handwritten mathematical expression recognition (HMER) is a multidisciplinary task that generates LaTeX sequences from images. Existing approaches, employing tree decoders within attention-based encoder-decoder architectures, aim to capture the hierarchical tree structure, but are limited by CFGs and pre-generated triplet data, hindering expandability and neglecting visual ambiguity challenges. This article investigates the distinctive language characteristics of LaTeX mathematical expressions, revealing two key observations: 1) the presence of explicit structural symbols, and 2) the treatment of symbols, particularly letters, as minimal units with context-dependent semantics, representing variables or constants. Rooted in these properties, we propose that language models have the potential to synchronously and complementarily provide both structural and semantic information, making them suitable for correction of HMER. To validate our proposition, we propose an

architecture called Recognize and Language Fusion Network (RLFN), which integrates recognition and language features to output corrected sequences while jointly optimizing with a string decoder recognition model. Experiments show that RLFN outperforms existing state-of-the-art methods on the CROHME 2014/2016/2019 datasets.

11:00-12:30 (East Foyer / Virtual)
**#87 Analyzing Cognitive Plausibility of Subword Tokenization**
*Lisa Beinborn and Yuval Pinter*
Subword tokenization has become the de-facto standard for tokenization although comparative evaluations of their quality across languages are scarce. Existing evaluation studies focus on the effect of a tokenization algorithm on the performance in downstream tasks, or on engineering criteria such as the compression rate. We present a new evaluation paradigm that focuses on the cognitive plausibility of subword tokenization. We analyze the correlation of the tokenizer output with the reading time and accuracy of human responses on a lexical decision task. We compare three tokenization algorithms across several languages and vocabulary sizes. Our results indicate that the Unigram algorithm yields less cognitively plausible tokenization behavior and a worse coverage of derivational morphemes, in contrast with prior work.

11:00-12:30 (East Foyer / Virtual)
**#88 An Iteratively Parallel Generation Method with the Pre-Filling Strategy for Document-level Event Extraction**
*Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang and Weinan E*
In document-level event extraction (DEE) tasks, a document typically contains many event records with multiple event roles. Therefore, accurately extracting all event records is a big challenge since the number of event records is not given. Previous works present the entity-based directed acyclic graph (EDAG) generation methods to autoregressively generate event roles, which requires a given generation order. Meanwhile, parallel methods are proposed to generate all event roles simultaneously, but suffer from the inadequate training which manifests zero accuracies on some event roles. In this paper, we propose an Iteratively Parallel Generation method with the Pre-Filling strategy (IPGPF). Event roles in an event record are generated in parallel to avoid order selection, and the event records are iteratively generated to utilize historical results. Experiments on two public datasets show our IPGPF improves $11.7$ F1 than previous parallel models and up to $5.1$ F1 than auto-regressive models under the control variable settings. Moreover, our enhanced IPGPF outperforms other entity-enhanced models and achieves new state-of-the-art performance.

11:00-12:30 (East Foyer / Virtual)
**#89 Empirical Study of Zero-Shot NER with ChatGPT**
*Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu and Hongwei Wang*
Large language models (LLMs) exhibited powerful capability in various natural language processing tasks. This work focuses on exploring LLM performance on zero-shot information extraction, with a focus on the ChatGPT and named entity recognition (NER) task. Inspired by the remarkable reasoning capability of LLM on symbolic and arithmetic reasoning, we adapt the prevalent reasoning methods to NER and propose reasoning strategies tailored for NER. First, we explore a decomposed question-answering paradigm by breaking down the NER task into simpler subproblems by labels. Second, we propose syntactic augmentation to stimulate the model's intermediate thinking in two ways: syntactic prompting, which encourages the model to analyze the syntactic structure itself, and tool augmentation, which provides the model with the syntactic information generated by a parsing tool. Besides, we adapt self-consistency to NER by proposing a two-stage majority voting strategy, which first votes for the most consistent mentions, then the most consistent types. The proposed methods achieve remarkable improvements for zero-shot NER across seven benchmarks, including Chinese and English datasets, and on both domain-specific and general-domain scenarios. In addition, we present a comprehensive analysis of the error types with suggestions for optimization directions. We also verify the effectiveness of the proposed methods on the few-shot setting and other LLMs.

11:00-12:30 (East Foyer / Virtual)
**#90 SAMRank: Unsupervised Keyphrase Extraction using Self-Attention Map in BERT and GPT-2**
*Byungha Kang and Youhyun Shin*
We propose a novel unsupervised keyphrase extraction approach, called SAMRank, which uses only a self-attention map in a pre-trained language model (PLM) to determine the importance of phrases. Most recent approaches for unsupervised keyphrase extraction mainly utilize contextualized embeddings to capture semantic relevance between words, sentences, and documents. However, due to the anisotropic nature of contextual embeddings, these approaches may not be optimal for semantic similarity measurements. SAMRank as proposed here computes the importance of phrases solely leveraging a self-attention map in a PLM, in this case BERT and GPT-2, eliminating the need to measure embedding similarities. To assess the level of importance, SAMRank combines both global and proportional attention scores through calculations using a self-attention map. We evaluate the SAMRank on three keyphrase extraction datasets: Inspec, SemEval2010, and SemEval2017. The experimental results show that SAMRank outperforms most embedding-based models on both long and short documents and demonstrating that it is possible to use only a self-attention map for keyphrase extraction without relying on embeddings. Source code is available at https://github.com/kangnlp/SAMRank.

11:00-12:30 (East Foyer / Virtual)
**#91 Revisiting Sparse Retrieval for Few-shot Entity Linking**
*Yulin Chen, Zhenran Xu, Baotian Hu and Min Zhang*
Entity linking aims to link ambiguous mentions to their corresponding entities in a knowledge base. One of the key challenges comes from insufficient labeled data for specific domains. Although dense retrievers have achieved excellent performance on several benchmarks, their performance decreases significantly when only a limited amount of in-domain labeled data is available. In such few-shot setting, we revisit the sparse retrieval method, and propose an ELECTRA-based keyword extractor to denoise the mention context and construct a better query expression. For training the extractor, we propose a distant supervision method to automatically generate training data based on overlapping tokens between mention contexts and entity descriptions. Experimental results on the ZESHEL dataset demonstrate that the proposed method outperforms state-of-the-art models by a significant margin across all test domains, showing the effectiveness of keyword-enhanced sparse retrieval.

11:00-12:30 (East Foyer / Virtual)
**#92 Weakly-Supervised Learning of Visual Relations in Multimodal Pretraining**
*Emanuele Bugliarello, Aida Nematzadeh and Lisa Anne Hendricks*
Recent work in vision-and-language pretraining has investigated supervised signals from object detection data to learn better, fine-grained multimodal representations. In this work, we take a step further and explore how we can tap into supervision from small-scale visual relation data. In particular, we propose two pretraining approaches to contextualise visual entities in a multimodal setup. With verbalised scene graphs, we transform visual relation triplets into structured captions, and treat them as additional image descriptions. With masked relation prediction, we further encourage relating entities from image regions with visually masked contexts. When applied to strong baselines pretrained on large amounts of Web data, zero-shot evaluations on both coarse-grained and fine-grained tasks show the efficacy of our methods in learning multimodal representations from weakly-supervised relations data.

11:00-12:30 (East Foyer / Virtual)
**#93 Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models**
*Laura Cabello, Emanuele Bugliarello, Stephanie Brandl and Desmond Elliott*
Pretrained machine learning models are known to perpetuate and even amplify existing biases in data, which can result in unfair outcomes that ultimately impact user experience. Therefore, it is crucial to understand the mechanisms behind those prejudicial biases to ensure that model performance does not result in discriminatory behaviour toward certain groups or populations. In this work, we define gender bias as our case study. We quantify bias amplification in pretraining and after fine-tuning on three families of vision-and-language models. We investigate the connection, if any, between the two learning stages, and evaluate how bias amplification reflects on model performance. Overall, we find that bias amplification in pretraining and after fine-tuning are independent. We then examine the effect of continued pretraining on gender-neutral data, finding that this reduces group disparities, i.e., promotes fairness, on VQAv2 and retrieval tasks without significantly compromising task performance.

11:00-12:30 (East Foyer / Virtual)
**#94 ACTOR: Active Learning with Annotator-specific Classification Heads to Embrace Human Label Variation**
*Xinpeng Wang and Barbara Plank*
Label aggregation such as majority voting is commonly used to resolve annotator disagreement in dataset creation. However, this may disregard minority values and opinions. Recent studies indicate that learning from individual annotations outperforms learning from aggregated labels, though they require a considerable amount of annotation. Active learning, as an annotation cost-saving strategy, has not been fully explored in the context of learning from disagreement. We show that in the active learning setting, a multi-head model performs significantly better than a single-head model in terms of uncertainty estimation. By designing and evaluating acquisition functions with annotator-specific heads on two datasets, we show that group-level entropy works generally well on both datasets. Importantly, it achieves performance in terms of both prediction and uncertainty estimation comparable to full-scale training from disagreement, while saving 70% of the annotation budget.

11:00-12:30 (East Foyer / Virtual)
**#95 Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews**
*Hye Sun Yun, Iain James Marshall, Thomas Trikalinos and Byron C Wallace*
Medical systematic reviews play a vital role in healthcare decision making and policy. However, their production is time-consuming, limiting the availability of high-quality and up-to-date evidence summaries. Recent advancements in LLMs offer the potential to automatically generate literature reviews on demand, addressing this issue. However, LLMs sometimes generate inaccurate (and potentially misleading) texts by hallucination or omission. In healthcare, this can make LLMs unusable at best and dangerous at worst. We conducted 16 interviews with international systematic review experts to characterize the perceived utility and risks of LLMs in the specific context of medical evidence reviews. Experts indicated that LLMs can assist in the writing process by drafting summaries, generating templates, distilling information, and crosschecking information. They also raised concerns regarding confidently composed but inaccurate LLM outputs and other potential downstream harms, including decreased accountability and proliferation of low-quality reviews. Informed by this qualitative analysis, we identify criteria for rigorous evaluation of biomedical LLMs aligned with domain expert views.

11:00-12:30 (East Foyer / Virtual)
**#96 Impressions: Visual Semiotics and Aesthetic Impact Understanding**
*Julia Kruk, Caleb Ziems and Diyi Yang*
Is aesthetic impact different from beauty? Is visual salience a reflection of its capacity for effective communication? We present Impressions, a novel dataset through which to investigate the semiotics of images, and how specific visual features and design choices can elicit specific emotions, thoughts and beliefs. We posit that the impactfulness of an image extends beyond formal definitions of aesthetics, to its success as a communicative act, where style contributes as much to meaning formation as the subject matter. We also acknowledge that existing Image Captioning datasets are not designed to empower state-of-the-art architectures to model potential human impressions or interpretations of images. To fill this need, we design an annotation task heavily inspired by image analysis techniques in the Visual Arts to collect 1,440 image-caption pairs and 4,320 unique annotations exploring impact, pragmatic image description, impressions and aesthetic design choices. We show that existing multimodal image captioning and conditional generation models struggle to simulate plausible human responses to images. However, this dataset significantly improves their ability to model impressions and aesthetic evaluations of images through fine-tuning and few-shot adaptation.

11:00-12:30 (East Foyer / Virtual)
**#97 CLEVR-Implicit: A Diagnostic Dataset for Implicit Reasoning in Referring Expression Comprehension**
*Jingwei Zhang, Xin Wu and Yi Cai*
Recently, pre-trained vision-language (VL) models have achieved remarkable success in various cross-modal tasks, including referring expression comprehension (REC). These models are pre-trained on the large-scale image-text pairs to learn the alignment between words in textual descriptions and objects in the corresponding images and then fine-tuned on downstream tasks. However, the performance of VL models is hindered when dealing with implicit text, which describes objects through comparisons between two or more objects rather than explicitly mentioning them. This is because the models struggle to align the implicit text with the objects in the images. To address the challenge, we introduce CLEVR-Implicit, a dataset consisting of synthetic images and corresponding two types of implicit text for the REC task. Additionally, to enhance the performance of VL models on implicit text, we propose a method called Transforming Implicit text into Explicit text (TIE), which enables VL models to reason with the implicit text. TIE consists of two modules: (1) the prompt design module builds prompts for implicit text by adding masked tokens, and (2) the cloze procedure module fine-tunes the prompts by utilizing masked language modeling (MLM) to predict the explicit words with the implicit prompts. Experimental results on our dataset demonstrate a significant improvement of 37.94% in the performance of VL models on implicit text after employing our TIE method.

11:00-12:30 (East Foyer / Virtual)
**#98 Vision-Enhanced Semantic Entity Recognition in Document Images via Visually-Asymmetric Consistency Learning**
*Hao Wang, Xiahua Chen, Rui Wang and Chenhui Chu*
Extracting meaningful entities belonging to predefined categories from Visually-rich Form-like Documents (VFDs) is a challenging task. Visual and layout features such as font, background, color, and bounding box location and size provide important cues for identifying entities of the same type. However, existing models commonly train a visual encoder with weak cross-modal supervision signals, resulting in a limited capacity to capture these non-textual features and suboptimal performance. In this paper, we propose a novel Visually-Asymmetric coNsistenCy Learning (VANCL) approach that addresses the above limitation by enhancing the model's ability to capture fine-grained visual and layout features through the incorporation of color priors. Experimental results on benchmark datasets show that our approach substantially outperforms the strong LayoutLM series baseline, demonstrating the effectiveness of our approach. Additionally, we investigate the effects of different color schemes on our approach, providing insights for optimizing model performance. We believe our work will inspire future research on multimodal information extraction.

11:00-12:30 (East Foyer / Virtual)

**#99 LIMIT: Language Identification, Misidentification, and Translation using Hierarchical Models in 350+ Languages**
*Milind Agarwal, Md Mahfuz Ibn Alam and Antonios Anastasopoulos*
Knowing the language of an input text/audio is a necessary first step for using almost every NLP tool such as taggers, parsers, or translation systems. Language identification is a well-studied problem, sometimes even considered solved; in reality, due to lack of data and computational challenges, current systems cannot accurately identify most of the world's 7000 languages. To tackle this bottleneck, we first compile a corpus, MCS-350, of 50K multilingual and parallel children's stories in 350+ languages. MCS-350 can serve as a benchmark for language identification of short texts and for 1400+ new translation directions in low-resource Indian and African languages. Second, we propose a novel misprediction-resolution hierarchical model, LIMIT, for language identification that reduces error by 55% (from 0.71 to 0.32) on our compiled children's stories dataset and by 40% (from 0.23 to 0.14) on the FLORES-200 benchmark. Our method can expand language identification coverage into low-resource languages by relying solely on systemic misprediction patterns, bypassing the need to retrain large models from scratch.

11:00-12:30 (East Foyer / Virtual)
**#100 ViPE: Visualise Pretty-much Everything**
*Hassan Shahmohammadi, Adhiraj Ghosh and Hendrik Lensch*
Figurative and non-literal expressions are profoundly integrated in human communication. Visualising such expressions allow us to convey our creative thoughts, and evoke nuanced emotions. Recent text-to-image models like Stable Diffusion, on the other hand, struggle to depict non-literal expressions. Recent works primarily deal with this issue by compiling humanly annotated datasets on a small scale, which not only demands specialized expertise but also proves highly inefficient. To address this issue, we introduce ViPE: Visualise Pretty-much Everything. ViPE offers a series of lightweight and robust language models that have been trained on a large-scale set of lyrics with noisy visual descriptions that represent their implicit meaning. The synthetic visual descriptions are generated by GPT3.5 relying on neither human annotations nor images. ViPE effectively expresses any arbitrary piece of text into a visualisable description, enabling meaningful and high-quality image generation. We provide compelling evidence that ViPE is more robust than GPT3.5 in synthesising visual elaborations. ViPE also exhibits an understanding of figurative expressions comparable to human experts, providing a powerful and open-source backbone to many downstream applications such as music video and caption generation.

11:00-12:30 (East Foyer / Virtual)
**#101 Visually-Situated Natural Language Understanding with Contrastive Reading Model and Frozen Large Language Models**
*Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee and Seunghyun Park*
Recent advances in Large Language Models (LLMs) have stimulated a surge of research aimed at extending their applications to the visual domain. While these models exhibit promise in generating abstract image captions and facilitating natural conversations, their performance on text-rich images still requires improvement. In this paper, we introduce Contrastive Reading Model (Cream), a novel neural architecture designed to enhance the language-image understanding capability of LLMs by capturing intricate details that are often overlooked in existing methods. Cream combines vision and auxiliary encoders, fortified by a contrastive feature alignment technique, to achieve a more effective comprehension of language information in visually situated contexts within the images. Our approach bridges the gap between vision and language understanding, paving the way for the development of more sophisticated Document Intelligence Assistants. Through rigorous evaluations across diverse visually-situated language understanding tasks that demand reasoning capabilities, we demonstrate the compelling performance of Cream, positioning it as a prominent model in the field of visual document understanding. We provide our codebase and newly-generated datasets at https://github.com/naver-ai/cream.

11:00-12:30 (East Foyer / Virtual)
**#102 Unifying Cross-Lingual Transfer across Scenarios of Resource Scarcity**
*Alan Ansell, Marinela Parović, Ivan Vulić, Anna Korhonen and Edoardo Ponti*
The scarcity of data in many of the world's languages necessitates the transfer of knowledge from other, resource-rich languages. However, the level of scarcity varies significantly across multiple dimensions, including: i) the amount of task-specific data available in the source and target languages; ii) the amount of monolingual and parallel data available for both languages; and iii) the extent to which they are supported by pretrained multilingual and translation models. Prior work has largely treated these dimensions and the various techniques for dealing with them separately; in this paper, we offer a more integrated view by exploring how to deploy the arsenal of cross-lingual transfer tools across a range of scenarios, especially the most challenging, low-resource ones. To this end, we run experiments on the AmericasNLI and NusaX benchmarks over 20 languages, simulating a range of few-shot settings. The best configuration in our experiments employed parameter-efficient language and task adaptation of massively multilingual Transformers, trained simultaneously on source language data and both machine-translated and natural data for multiple target languages. In addition, we show that pre-trained translation models can be easily adapted to unseen languages, thus extending the range of our hybrid technique and translation-based transfer more broadly. Beyond new insights into the mechanisms of cross-lingual transfer, we hope our work will provide practitioners with a toolbox to integrate multiple techniques for different real-world scenarios. Our code is available at https://github.com/parovicm/unified-xlt.

11:00-12:30 (East Foyer / Virtual)
**#103 NeuSTIP: A Neuro-Symbolic Model for Link and Time Prediction in Temporal Knowledge Graphs**
*Ishaan Singh, Navdeep Kaur, Garima Gaur and Mausam*
Neuro-symbolic (NS) models for knowledge graph completion (KGC) combine the benefits of symbolic models (interpretable inference) with those of distributed representations (parameter sharing, high accuracy). While several NS models exist for KGs with static facts, there is limited work on temporal KGC (TKGC) for KGs where a fact is associated with a time interval. In response, we propose a novel NS model for TKGC called NeuSTIP, which performs link prediction and time interval prediction in a TKG. NeuSTIP learns temporal rules with Allen predicates, which ensure temporal consistency between neighboring predicates in the rule body. We further design a unique scoring function that evaluates the confidence of the candidate answers while performing link and time interval predictions by utilizing the learned rules. Our empirical evaluation on two time interval based TKGC datasets shows that our model shows competitive performance on link prediction and establishes a new state of the art on time prediction.

11:00-12:30 (East Foyer / Virtual)
**#104 Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models**
*Jirui Qi, Raquel Fernández and Arianna Bisazza*
Multilingual large-scale Pretrained Language Models (PLMs) have been shown to store considerable amounts of factual knowledge, but large variations are observed across languages. With the ultimate goal of ensuring that users with different language backgrounds obtain consistent feedback from the same model, we study the cross-lingual consistency (CLC) of factual knowledge in various multilingual PLMs. To this end, we propose a Ranking-based Consistency (RankC) metric to evaluate knowledge consistency across languages independently from accuracy. Using this metric, we conduct an in-depth analysis of the determining factors for CLC, both at model level and at language-pair level. Among other results, we find that increasing model size leads to higher factual probing accuracy in most languages, but does not improve cross-lingual consistency. Finally, we conduct a case study on CLC when new factual associations are inserted in the PLMs via model editing. Results on a small sample of facts inserted in English reveal a clear pattern whereby the new piece of knowledge transfers only to lan-

guages with which English has a high RankC score. All code and data are released at https://github.com/Betswish/Cross-Lingual-Consistency.

11:00-12:30 (East Foyer / Virtual)
**#105 Set Learning for Generative Information Extraction**
*Jiangnan Li, Yice Zhang, Bin Liang, Kam-Fai Wong and Ruifeng Xu*
Recent efforts have endeavored to employ the sequence-to-sequence (Seq2Seq) model in Information Extraction (IE) due to its potential to tackle multiple IE tasks in a unified manner. Under this formalization, multiple structured objects are concatenated as the target sequence in a predefined order. However, structured objects, by their nature, constitute an unordered set. Consequently, this formalization introduces a potential order bias, which can impair model learning. Targeting this issue, this paper proposes a set learning approach that considers multiple permutations of structured objects to optimize set probability approximately. Notably, our approach does not require any modifications to model structures, making it easily integrated into existing generative IE frameworks. Experiments show that our method consistently improves existing frameworks on vast tasks and datasets.

11:00-12:30 (East Foyer / Virtual)
**#106 Confidence-based Ensembling of Perspective-aware Models**
*Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti and Cristina Bosco*
Research in the field of NLP has recently focused on the variability that people show in selecting labels when performing an annotation task. Exploiting disagreements in annotations has been shown to offer advantages for accurate modelling and fair evaluation. In this paper, we propose a strongly perspectivist model for supervised classification of natural language utterances. Our approach combines the predictions of several perspective-aware models using key information of their individual confidence to capture the subjectivity encoded in the annotation of linguistic phenomena. We validate our method through experiments on two case studies, irony and hate speech detection, in in-domain and cross-domain settings. The results show that confidence-based ensembling of perspective-aware models seems beneficial for classification performance in all scenarios. In addition, we demonstrate the effectiveness of our method with automatically extracted perspectives from annotations when the annotators' metadata are not available.

11:00-12:30 (East Foyer / Virtual)
**#107 Multitask Multimodal Prompted Training for Interactive Embodied Task Completion**
*Georgios Pantazopoulos, Malvina Nikandrou, Amit Parekh, Bhathiya Hemanthage, Arash Eshghi, Ioannis Konstas, Verena Rieser, Oliver Lemon and Alessandro Suglia*
Interactive and embodied tasks pose at least two fundamental challenges to existing Vision & Language (VL) models, including 1) grounding language in trajectories of actions and observations, and 2) referential disambiguation. To tackle these challenges, we propose an Embodied MultiModal Agent (EMMA): a unified encoder-decoder model that reasons over images and trajectories, and casts action prediction as multimodal text generation. By unifying all tasks as text generation, EMMA learns a language of actions which facilitates transfer across tasks. Different to previous modular approaches with independently trained components, we use a single multitask model where each task contributes to goal completion. EMMA performs on par with similar models on several VL benchmarks and sets a new state-of-the-art performance (36.81% success rate) on the Dialog-guided Task Completion (DTC), a benchmark to evaluate dialog-guided agents in the Alexa Arena.

11:00-12:30 (East Foyer / Virtual)
**#108 Improving Unsupervised Relation Extraction by Augmenting Diverse Sentence Pairs**
*Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li and Qi Li*
Unsupervised relation extraction (URE) aims to extract relations between named entities from raw text without requiring manual annotations or pre-existing knowledge bases. In recent studies of URE, researchers put a notable emphasis on contrastive learning strategies for acquiring relation representations. However, these studies often overlook two important aspects: the inclusion of diverse positive pairs for contrastive learning and the exploration of appropriate loss functions. In this paper, we propose AugURE with both within-sentence pairs augmentation and augmentation through cross-sentence pairs extraction to increase the diversity of positive pairs and strengthen the discriminative power of contrastive learning. We also identify the limitation of noise-contrastive estimation (NCE) loss for relation representation learning and propose to apply margin loss for sentence pairs. Experiments on NYT-FB and TACRED datasets demonstrate that the proposed relation representation learning and a simple K-Means clustering achieves state-of-the-art performance.

11:00-12:30 (East Foyer / Virtual)
**#109 GROOViST: A Metric for Grounding Objects in Visual Storytelling**
*Aditya Kaushik Surikuchi, Sandro Pezzelle and Raquel Fernández*
A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

11:00-12:30 (East Foyer / Virtual)
**#110 Multimodal Embodied Plan Prediction Augmented with Synthetic Embodied Dialogue**
*Aishwarya Padmakumar, Mert Inan, Spandana Gella, Patrick L. Lange and Dilek Hakkani-Tur*
Embodied task completion is a challenge where an agent in a simulated environment must predict environment actions to complete tasks based on natural language instructions and ego-centric visual observations. We propose a variant of this problem where the agent predicts actions at a higher level of abstraction called a plan, which helps make agent actions more interpretable and can be obtained from the appropriate prompting of large language models. We show that multimodal transformer models can outperform language-only models for this problem but fall significantly short of oracle plans. Since collecting human-human dialogues for embodied environments is expensive and time-consuming, we propose a method to synthetically generate such dialogues, which we then use as training data for plan prediction. We demonstrate that multimodal transformer models can attain strong zero-shot performance from our synthetic data, outperforming language-only models trained on human-human data.

11:00-12:30 (East Foyer / Virtual)
**#111 S2abEL: A Dataset for Entity Linking from Scientific Tables**
*Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik and Doug Downey*
Entity linking (EL) is the task of linking a textual mention to its corresponding entry in a knowledge base, and is critical for many knowledge-intensive NLP applications. When applied to tables in scientific papers, EL is a step toward large-scale scientific knowledge bases that could enable advanced scientific question answering and analytics. We present the first dataset for EL in scientific tables. EL for scientific tables is especially challenging because scientific knowledge bases can be very incomplete, and disambiguating table mentions typically requires

understanding the paper's text in addition to the table. Our dataset, Scientific Table Entity Linking (S2abEL), focuses on EL in machine learning results tables and includes hand-labeled cell types, attributed sources, and entity links from the PaperswithCode taxonomy for 8,429 cells from 732 tables. We introduce a neural baseline method designed for EL on scientific tables containing many out-of-knowledge-base mentions, and show that it significantly outperforms a state-of-the-art generic table EL method. The best baselines fall below human performance, and our analysis highlights avenues for improvement.

11:00-12:30 (East Foyer / Virtual)
**#112 Revisiting the Optimality of Word Lengths**
*Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald and Ryan Cotterell*
Zipf (1935) posited that wordforms are optimized to minimize utterances' communicative costs. Under the assumption that cost is given by an utterance's length, he supported this claim by showing that words' lengths are inversely correlated with their frequencies. Communicative cost, however, can be operationalized in different ways. Piantadosi et al. (2011) claim that cost should be measured as the distance between an utterance's information rate and channel capacity, which we dub the channel capacity hypothesis (CCH) here. Following this logic, they then proposed that a word's length should be proportional to the expected value of its surprisal (negative log-probability in context). In this work, we show that Piantadosi et al.'s derivation does not minimize CCH's cost, but rather a lower bound, which we term CCH-lower. We propose a novel derivation, suggesting an improved way to minimize CCH's cost. Under this method, we find that a language's word lengths should instead be proportional to the surprisal's expectation plus its variance-to-mean ratio. Experimentally, we compare these three communicative cost functions: Zipf's, CCH-lower , and CCH. Across 13 languages and several experimental settings, we find that length is better predicted by frequency than either of the other hypotheses. In fact, when surprisal's expectation, or expectation plus variance-to-mean ratio, is estimated using better language models, it leads to worse word length predictions. We take these results as evidence that Zipf's longstanding hypothesis holds.

11:00-12:30 (East Foyer / Virtual)
**#113 ZGUL: Zero-shot Generalization to Unseen Languages using Multi-source Ensembling of Language Adapters**
*Vipul Kumar Rathore, Rajdeep Dhingra, Parag Singla and Mausam*
We tackle the problem of zero-shot cross-lingual transfer in NLP tasks via the use of language adapters (LAs). Most of the earlier works have explored training with adapter of a single source (often English), and testing either using the target LA or LA of another related language. Training target LA requires unlabeled data, which may not be readily available for low resource *unseen* languages: those that are neither seen by the underlying multilingual language model (e.g., mBERT), nor do we have any (labeled or unlabeled) data for them. We posit that for more effective cross-lingual transfer, instead of just one source LA, we need to leverage LAs of multiple (linguistically or geographically related) source languages, both at train and test-time - which we investigate via our novel neural architecture, ZGUL. Extensive experimentation across four language groups, covering 15 unseen target languages, demonstrates improvements of up to 3.2 average F1 points over standard fine-tuning and other strong baselines on POS tagging and NER tasks. We also extend ZGUL to settings where either (1) some unlabeled data or (2) few-shot training examples are available for the target language. We find that ZGUL continues to outperform baselines in these settings too.

11:00-12:30 (East Foyer / Virtual)
**#114 Code-Switching Metrics Using Intonation Units**
*Rebecca Pattichis, Dora LaCasse, Sonya Mitrovich Trawick and Rena Torres Cacoullos*
Code-switching (CS) metrics in NLP that are based on word-level units are misaligned with true bilingual CS behavior. Crucially, CS is not equally likely between any two words, but follows syntactic and prosodic rules. We adapt two metrics, multilinguality and CS probability, and apply them to transcribed bilingual speech, for the first time putting forward Intonation Units (IUs) – prosodic speech segments – as basic tokens for NLP tasks. In addition, we calculate these two metrics separately for distinct mixing types: alternating-language multi-word strings and single-word incorporations from one language into another. Results indicate that individual differences according to the two CS metrics are independent. However, there is a shared tendency among bilinguals for multi-word CS to occur across, rather than within, IU boundaries. That is, bilinguals tend to prosodically separate their two languages. This constraint is blurred when metric calculations do not distinguish multi-word and single-word items. These results call for a reconsideration of units of analysis in future development of CS datasets for NLP tasks.

11:00-12:30 (East Foyer / Virtual)
**#115 Joint Entity and Relation Extraction with Span Pruning and Hypergraph Neural Networks**
*Zhaohui Yan, Songlin Yang, Wei Liu and Kewei Tu*
Entity and Relation Extraction (ERE) is an important task in information extraction. Recent marker-based pipeline models achieve state-of-the-art performance, but still suffer from the error propagation issue. Also, most of current ERE models do not take into account higher-order interactions between multiple entities and relations, while higher-order modeling could be beneficial.In this work, we propose HyperGraph neural network for ERE (HGERE), which is built upon the PL-marker (a state-of-the-art marker-based pipleline model). To alleviate error propagation, we use a high-recall pruner mechanism to transfer the burden of entity identification and labeling from the NER module to the joint module of our model. For higher-order modeling, we build a hypergraph, where nodes are entities (provided by the span pruner) and relations thereof, and hyperedges encode interactions between two different relations or between a relation and its associated subject and object entities. We then run a hypergraph neural network for higher-order inference by applying message passing over the built hypergraph. Experiments on three widely used benchmarks (ACE2004, ACE2005 and SciERC) for ERE task show significant improvements over the previous state-of-the-art PL-marker.

11:00-12:30 (East Foyer / Virtual)
**#116 Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction**
*Martin Josifoski, Marija Sakota, Maxime Peyrard and Robert West*
Large language models (LLMs) have great potential for synthetic data generation. This work shows that useful data can be synthetically generated even for tasks that cannot be solved directly by LLMs: for problems with structured outputs, it is possible to prompt an LLM to perform the task in the reverse direction, by generating plausible input text for a target output structure. Leveraging this asymmetry in task difficulty makes it possible to produce large-scale, high-quality data for complex tasks. We demonstrate the effectiveness of this approach on closed information extraction, where collecting ground-truth data is challenging, and no satisfactory dataset exists to date. We synthetically generate a dataset of 1.8M data points, establish its superior quality compared to existing datasets in a human evaluation, and use it to finetune small models (220M and 770M parameters), termed SynthIE, that outperform the prior state of the art (with equal model size) by a substantial margin of 57 absolute points in micro-F1 and 79 points in macro-F1. Code, data, and models are available at anonymous.

11:00-12:30 (East Foyer / Virtual)
**#117 The BLA Benchmark: Investigating Basic Language Abilities of Pre-Trained Multimodal Models**
*Xinyi Chen, Raquel Fernández and Sandro Pezzelle*
Despite the impressive performance achieved by pre-trained language-and-vision models in downstream tasks, it remains an open question

whether this reflects a proper understanding of image-text interaction. In this work, we explore to what extent they handle basic linguistic constructions—active-passive voice, coordination, and relative clauses—that even preschool children can typically master. We present BLA, a novel, automatically constructed benchmark to evaluate multimodal models on these Basic Language Abilities. We show that different types of Transformer-based systems, such as CLIP, ViLBERT, and BLIP2, generally struggle with BLA in a zero-shot setting, in line with previous findings. Our experiments, in particular, show that most of the tested models only marginally benefit when fine-tuned or prompted with construction-specific samples. Yet, the generative BLIP2 shows promising trends, especially in an in-context learning setting. This opens the door to using BLA not only as an evaluation benchmark but also to improve models' basic language abilities.

11:00-12:30 (East Foyer / Virtual)
**#118 Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors**
*Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine C Khoong, Marine Carpuat and Niloufar Salehi*
A major challenge in the practical use of Machine Translation (MT) is that users lack information on translation quality to make informed decisions about how to rely on outputs. Progress in quality estimation research provides techniques to automatically assess MT quality, but these techniques have primarily been evaluated in vitro by comparison against human judgments outside of a specific context of use. This paper evaluates quality estimation feedback in vivo with a human study in realistic high-stakes medical settings. Using Emergency Department discharge instructions, we study how interventions based on quality estimation versus backtranslation assist physicians in deciding whether to show MT outputs to a patient. We find that quality estimation improves appropriate reliance on MT, but backtranslation helps physicians detect more clinically harmful errors that QE alone often misses.

11:00-12:30 (East Foyer / Virtual)
**#119 Explaining with Contrastive Phrasal Highlighting: A Case Study in Assisting Humans to Detect Translation Differences**
*Eleftheria Briakou, Navita Goyal and Marine Carpuat*
Explainable NLP techniques primarily explain by answering "Which tokens in the input are responsible for this prediction?". We argue that for NLP models that make predictions by comparing two input texts, it is more useful to explain by answering "What differences between the two inputs explain this prediction?". We introduce a technique to generate contrastive phrasal highlights that explain the predictions of a semantic divergence model via phrase alignment guided erasure. We show that the resulting highlights match human rationales of cross-lingual semantic differences better than popular post-hoc saliency techniques and that they successfully help people detect fine-grained meaning differences in human translations and critical machine translation errors.

11:00-12:30 (East Foyer / Virtual)
**#120 Understanding the Role of Input Token Characters in Language Models: How Does Information Loss Affect Performance?**
*Ahmed Alajrami, Katerina Margatina and Nikolaos Aletras*
Understanding how and what pre-trained language models (PLMs) learn about language is an open challenge in natural language processing. Previous work has focused on identifying whether they capture semantic and syntactic information, and how the data or the pre-training objective affects their performance. However, to the best of our knowledge, no previous work has specifically examined how information loss in input token characters affects the performance of PLMs. In this study, we address this gap by pre-training language models using small subsets of characters from individual tokens. Surprisingly, we find that pre-training even under extreme settings, i.e. using only one character of each token, the performance retention in standard NLU benchmarks and probing tasks compared to full-token models is high. For instance, a model pre-trained only on single first characters from tokens achieves performance retention of approximately 90% and 77% of the full-token model in SuperGLUE and GLUE tasks, respectively.

11:00-12:30 (East Foyer / Virtual)
**#121 What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on QA Systems**
*Navita Goyal, Eleftheria Briakou, Amanda Stephanie Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare R. Voss, Marine Carpuat and Hal Daumé III*
NLP systems have shown impressive performance at answering questions by retrieving relevant context. However, with the increasingly large models, it is impossible and often undesirable to constrain models' knowledge or reasoning to only the retrieved context. This leads to a mismatch between the information that *the models* access to derive the answer and the information that is available to *the user* to assess the model predicted answer. In this work, we study how users interact with QA systems in the absence of sufficient information to assess their predictions. Further, we ask whether adding the requisite background helps mitigate users' over-reliance on predictions. Our study reveals that users rely on model predictions even in the absence of sufficient information needed to assess the model's correctness. Providing the relevant background, however, helps users better catch model errors, reducing over-reliance on incorrect predictions. On the flip side, background information also increases users' confidence in their accurate as well as inaccurate judgments. Our work highlights that supporting users' verification of QA predictions is an important, yet challenging, problem.

11:00-12:30 (East Foyer / Virtual)
**#122 HiddenTables and PyQTax: A Cooperative Game and Dataset For TableQA to Ensure Scale and Data Privacy Across a Myriad of Taxonomies**
*William Watson, Nicole Cho, Tucker Balch and Manuela Veloso*
A myriad of different Large Language Models (LLMs) face a common challenge in contextually analyzing table question-answering tasks. These challenges are engendered from (1) finite context windows for large tables, (2) multi-faceted discrepancies amongst tokenization patterns against cell boundaries, and (3) various limitations stemming from data confidentiality in the process of using external models such as gpt-35-turbo. We propose a cooperative game dubbed "HiddenTables" as a potential resolution to this challenge. In essence, "HiddenTables" is played between the code-generating LLM "Solver" and the "Oracle" which evaluates the ability of the LLM agents to solve TableQA tasks. This game is based on natural language schemas and importantly, ensures the security of the underlying data. We provide evidential experiments on a diverse set of tables that demonstrate an LLM's collective inability to generalize and perform on complex queries, handle compositional dependencies, and align natural language to programmatic commands when concrete table schemas are provided. Unlike encoder-based models, we have pushed the boundaries of "HiddenTables" to not be limited by the number of rows - therefore we exhibit improved efficiency in prompt and completion tokens. Our infrastructure has spawned a new dataset "PyQTax" that spans across 116,671 question-table-answer triplets and provides additional fine-grained breakdowns and labels for varying question taxonomies. Therefore, in tandem with our academic contributions regarding LLMs' deficiency in TableQA tasks, "HiddenTables" is a tactile manifestation of how LLMs can interact with massive datasets while ensuring data security and minimizing generation costs.

11:00-12:30 (East Foyer / Virtual)
**#123 Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers**
*Daniela Teodorescu, Tiffany Cheng, Alona Fyshe and Saif M. Mohammad*
Research in psychopathology has shown that, at an aggregate level, the patterns of emotional change over time—emotion dynamics—are indicators of one's mental health. One's patterns of emotion change have traditionally been determined through self-reports of emotions;

however, there are known issues with accuracy, bias, and convenience. Recent approaches to determining emotion dynamics from one's everyday utterances, addresses many of these concerns, but it is not yet known whether these measures of utterance emotion dynamics (UED) correlate with mental health diagnoses. Here, for the first time, we study the relationship between tweet emotion dynamics and mental health disorders. We find that each of the UED metrics studied varied by the user's self-disclosed diagnosis. For example: average valence was significantly higher (i.e., more positive text) in the control group compared to users with ADHD, MDD, and PTSD. Valence variability was significantly lower in the control group compared to ADHD, depression, bipolar disorder, MDD, PTSD, and OCD but not PPD. Rise and recovery rates of valence also exhibited significant differences from the control. This work provides important early evidence for how linguistic cues pertaining to emotion dynamics can play a crucial role as biosocial markers for mental illnesses and aid in the understanding, diagnosis, and management of mental health disorders.

11:00-12:30 (East Foyer / Virtual)
**#124 Evaluating and Modeling Attribution for Cross-Lingual Question Answering**
*Benjamin Muller, John Frederick Wieting, Jonathan H. Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig and Xinyi Wang*
Trustworthy answer content is abundant in many high-resource languages and is instantly accessible through question answering systems — yet this content can be hard to access for those that do not speak these languages. The leap forward in cross-lingual modeling quality offered by generative language models offers much promise, yet their raw generations often fall short in factuality. To improve trustworthiness in these systems, a promising direction is to attribute the answer to a retrieved source, possibly in a content-rich language different from the query. Our work is the first to study attribution for cross-lingual question answering. First, we collect data in 5 languages to assess the attribution level of a state-of-the-art cross-lingual QA system. To our surprise, we find that a substantial portion of the answers is not attributable to any retrieved passages (up to 50% of answers exactly matching a gold reference) despite the system being able to attend directly to the retrieved text. Second, to address this poor attribution level, we experiment with a wide range of attribution detection techniques. We find that Natural Language Inference models and PaLM 2 fine-tuned on a very small amount of attribution data can accurately detect attribution. With these models, we improve the attribution level of a cross-lingual QA system. Overall, we show that current academic generative cross-lingual QA systems have substantial shortcomings in attribution and we build tooling to mitigate these issues.

11:00-12:30 (East Foyer / Virtual)
**#125 Analyzing Modular Approaches for Visual Question Decomposition**
*Apoorv Khandelwal, Ellie Pavlick and Chen Sun*
Modular neural networks without additional training have recently been shown to surpass end-to-end neural networks on challenging vision–language tasks. The latest such methods simultaneously introduce LLM-based code generation to build programs and a number of skill-specific, task-oriented modules to execute them. In this paper, we focus on ViperGPT and ask where its additional performance comes from and how much is due to the (state-of-art, end-to-end) BLIP-2 model it subsumes vs. additional symbolic components. To do so, we conduct a controlled study (comparing end-to-end, modular, and prompting-based methods across several VQA benchmarks). We find that ViperGPT's reported gains over BLIP-2 can be attributed to its selection of task-specific modules, and when we run ViperGPT using a more task-agnostic selection of modules, these gains go away. ViperGPT retains much of its performance if we make prominent alterations to its selection of modules: e.g. removing or retaining only BLIP-2. We also compare ViperGPT against a prompting-based decomposition strategy and find that, on some benchmarks, modular approaches significantly benefit by representing subtasks with natural language, instead of code. Our code is fully available at https://github.com/brown-palm/visual-question-decomposition.

11:00-12:30 (East Foyer / Virtual)
**#126 Emergence of Abstract State Representations in Embodied Sequence Modeling**
*Tian Yun, Zilai Zeng, Kunal Handa, Ashish V Thapliyal, Bo Pang, Ellie Pavlick and Chen Sun*
Decision making via sequence modeling aims to mimic the success of language models, where actions taken by an embodied agent are modeled as tokens to predict. Despite their promising performance, it remains unclear if embodied sequence modeling leads to the emergence of internal representations that represent the environmental state information. A model that lacks abstract state representations would be liable to make decisions based on surface statistics which fail to generalize. We take the BabyAI environment, a grid world in which language-conditioned navigation tasks are performed, and build a sequence modeling Transformer, which takes a language instruction, a sequence of actions, and environmental observations as its inputs. In order to investigate the emergence of abstract state representations, we design a "blindfolded" navigation task, where only the initial environmental layout, the language instruction, and the action sequence to complete the task are available for training. Our probing results show that intermediate environmental layouts can be reasonably reconstructed from the internal activations of a trained model, and that language instructions play a role in the reconstruction accuracy. Our results suggest that many key features of state representations can emerge via embodied sequence modeling, supporting an optimistic outlook for applications of sequence modeling objectives to more complex embodied decision-making domains.

11:00-12:30 (East Foyer / Virtual)
**#127 Task-Agnostic Low-Rank Adapters for Unseen English Dialects**
*Zedian Xiao, William Barr Held, Yanchen Liu and Diyi Yang*
Large Language Models (LLMs) are trained on corpora disproportionally weighted in favor of Standard American English. As a result, speakers of other dialects experience significantly more failures when interacting with these technologies. In practice, these speakers often accommodate their speech to be better understood. Our work shares the belief that language technologies should be designed to accommodate the diversity in English dialects and not the other way around. However, prior work on dialect struggle with generalizing to evolving and emerging dialects in a scalable manner. To fill this gap, our method, HyperLoRA, leverages expert linguistic knowledge to enable resource-efficient adaptation via hypernetworks. By disentangling dialect-specific and cross-dialectal information, HyperLoRA improves generalization to unseen dialects in a task-agnostic fashion. Not only is HyperLoRA more scalable in the number of parameters, but it also achieves the best or most competitive performance across 5 dialects in a zero-shot setting. In this way, our approach facilitates access to language technology for billions of English dialect speakers who are traditionally underrepresented.

11:00-12:30 (East Foyer / Virtual)
**#128 Abstractive Open Information Extraction**
*Kevin Song Pei, Ishan Jindal and Kevin Chang*
Open Information Extraction (OpenIE) is a traditional NLP task that extracts structured information from unstructured text to be used for other downstream applications. Traditionally, OpenIE focuses on extracting the surface forms of relations as they appear in the raw text, which we term extractive OpenIE. One of the main drawbacks of this approach is that implicit semantic relations (inferred relations) can not be extracted, compromising the performance of downstream applications. In this paper, we broaden the scope of OpenIE relations from merely the surface form of relations to include inferred relations, which we term abstractive OpenIE. This new task calls for the development of a new abstractive OpenIE training dataset and a baseline neural model that can extract those inferred relations. We also demonstrate the necessity for a new semantics-based metric for evaluating abstractive OpenIE extractions. Via a case study on Complex QA, we demonstrate the effectiveness of abstractive OpenIE.

11:00-12:30 (East Foyer / Virtual)
**#129 Grounding Visual Illusions in Language: Do Vision-Language Models Perceive Illusions Like Humans?**
*Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan and Joyce Chai*
Vision-Language Models (VLMs) are trained on vast amounts of data captured by humans emulating our understanding of the world. However, known as visual illusions, human's perception of reality isn't always faithful to the physical world. This raises a key question: do VLMs have the similar kind of illusions as humans do, or do they faithfully learn to represent reality? To investigate this question, we build a dataset containing five types of visual illusions and formulate four tasks to examine visual illusions in state-of-the-art VLMs. Our findings have shown that although the overall alignment is low, larger models are closer to human perception and more susceptible to visual illusions. Our dataset and initial findings will promote a better understanding of visual illusions in humans and machines and provide a stepping stone for future computational models that can better align humans and machines in perceiving and communicating about the shared visual world. The code and data are available at [github.com/vl-illusion/dataset](https://github.com/vl-illusion/dataset).

11:00-12:30 (East Foyer / Virtual)
**#130 UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning**
*Ahmed Masry, Parsa Kavehzadeh, Do Xuan Long, Enamul Hoque and Shafiq Joty*
Charts are widely used for data analysis, providing visual representations and insights into complex data. To facilitate chart-based data analysis using natural language, several downstream tasks have been introduced recently such as chart question answering and chart summarization. However, existing methods for these tasks often rely on pretraining on language or vision-language tasks, neglecting the explicit modeling of chart structures (e.g., how chart elements are related to each other). To address this, we first build a large corpus of charts covering diverse topics and visual styles. We then present UniChart, a pretrained model for chart comprehension and reasoning. UniChart encodes the relevant text, data, and visual elements of charts and then uses a chart-grounded text decoder for text generation. We propose several chart-specific pretraining tasks that include: (i) low-level tasks to extract the visual elements (e.g., bars, lines) and data from charts, and (ii) high-level tasks to acquire chart understanding and reasoning skills. Our experiments demonstrate that pretraining UniChart on a large corpus with chart-specific objectives, followed by fine-tuning, yields state-of-the-art performance on four downstream tasks. Moreover, our model exhibits superior generalizability to unseen chart corpus, surpassing previous approaches that lack chart-specific objectives and utilize limited chart resources.

11:00-12:30 (East Foyer / Virtual)
**#131 Better Quality Pre-training Data and T5 Models for African Languages**
*Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani and Jimmy Lin*
In this study, we highlight the importance of enhancing the quality of pretraining data in multilingual language models. Existing web crawls have demonstrated quality issues, particularly in the context of low-resource languages. Consequently, we introduce a new multilingual pretraining corpus for 16 African languages, designed by carefully auditing existing pretraining corpora to understand and rectify prevalent quality issues. To compile this dataset, we undertake a rigorous examination of current data sources for thirteen languages within one of the most extensive multilingual web crawls, mC4, and extract cleaner data through meticulous auditing and improved web crawling strategies. Subsequently, we pretrain a new T5-based model on this dataset and evaluate its performance on multiple downstream tasks. Our model demonstrates better downstream effectiveness over existing pretrained models across four NLP tasks, underscoring the critical role data quality plays in pretraining language models in low-resource scenarios. Specifically, on cross-lingual QA evaluation, our new model is more than twice as effective as multilingual T5. All code, data and models are publicly available at https://github.com/castorini/AfriTeVa-keji.

11:00-12:30 (East Foyer / Virtual)
**#132 Semi-automatic Data Enhancement for Document-Level Relation Extraction with Distant Supervision from Large Language Models**
*Junpeng Li, Zixia Jia and Zilong Zheng*
Document-level Relation Extraction (DocRE), which aims to extract relations from a long context, is a critical challenge in achieving fine-grained structural comprehension and generating interpretable document representations. Inspired by recent advances in in-context learning capabilities emergent from large language models (LLMs), such as ChatGPT, we aim to design an automated annotation method for DocRE with minimum human effort. Unfortunately, vanilla in-context learning is infeasible for DocRE due to the plenty of predefined fine-grained relation types and the uncontrolled generations of LLMs. To tackle this issue, we propose a method integrating an LLM and a natural language inference (NLI) module to generate relation triples, thereby augmenting document-level relation datasets. We demonstrate the effectiveness of our approach by introducing an enhanced dataset known as DocGNRE, which excels in re-annotating numerous long-tail relation types. We are confident that our method holds the potential for broader applications in domain-specific relation type definitions and offers tangible benefits in advancing generalized language semantic comprehension.

11:00-12:30 (East Foyer / Virtual)
**#133 Unified Low-Resource Sequence Labeling by Sample-Aware Dynamic Sparse Finetuning**
*Sarkar Snigdha Sarathi Das, Haoran Ranran Zhang, Peng Shi, Wenpeng Yin and Rui Zhang*
Unified Sequence Labeling that articulates different sequence labeling problems such as Named Entity Recognition, Relation Extraction, Semantic Role Labeling, etc. in a generalized sequence-to-sequence format opens up the opportunity to make the maximum utilization of large language model knowledge toward structured prediction. Unfortunately, this requires formatting them into specialized augmented format unknown to the base pretrained language model (PLMs) necessitating finetuning to the target format. This significantly bounds its usefulness in data-limited settings where finetuning large models cannot properly generalize to the target format. To address this challenge and leverage PLM knowledge effectively, we propose FISH-DIP, a sample-aware dynamic sparse finetuning strategy that selectively focuses on a fraction of parameters, informed by feedback from highly regressing examples, during the fine-tuning process. By leveraging the dynamism of sparsity, our approach mitigates the impact of well-learned samples and prioritizes underperforming instances for improvement in generalization. Across five tasks of sequence labeling, we demonstrate that FISH-DIP can smoothly optimize the model in low resource settings offering upto 40% performance improvements over full fine-tuning depending on target evaluation settings. Also, compared to in-context learning and other parameter-efficient fine-tuning approaches, FISH-DIP performs comparably or better, notably in extreme low-resource settings. The source code of FISH-DIP will be available at [this URL](https://github.com/psunlpgroup/FISH-DIP)

11:00-12:30 (East Foyer / Virtual)
**#134 ACQUIRED: A Dataset for Answering Counterfactual Questions In Real-Life Videos**
*Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Reddy Chandra, Marjorie Freedman, Ralph M. Weischedel and Nanyun Peng*
Multimodal counterfactual reasoning is a vital yet challenging ability for AI systems. It involves predicting the outcomes of hypothetical circumstances based on vision and language inputs, which enables AI models to learn from failures and explore hypothetical scenarios. Despite its importance, there are only a few datasets targeting the counterfactual reasoning abilities of multimodal models. Among them, they only cover reasoning over synthetic environments or specific types of events (e.g. traffic collisions), making them hard to reliably benchmark the model generalization ability in diverse real-world scenarios and reasoning dimensions. To overcome these limitations, we develop a video

question answering dataset, ACQUIRED: it consists of 3.9K annotated videos, encompassing a wide range of event types and incorporating both first and third-person viewpoints, which ensures a focus on real-world diversity. In addition, each video is annotated with questions that span three distinct dimensions of reasoning, including physical, social, and temporal, which can comprehensively evaluate the model counterfactual abilities along multiple aspects. We benchmark our dataset against several state-of-the-art language-only and multimodal models and experimental results demonstrate a significant performance gap (>13%) between models and humans. The findings suggest that multimodal counterfactual reasoning remains an open challenge and ACQUIRED is a comprehensive and reliable benchmark for inspiring future research in this direction.

11:00-12:30 (East Foyer / Virtual)
**#135 When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks**
*Eve Fleisig, Rediet Abebe and Dan Klein*
Though majority vote among annotators is typically used for ground truth labels in machine learning, annotator disagreement in tasks such as hate speech detection may reflect systematic differences in opinion across groups, not noise. Thus, a crucial problem in hate speech detection is determining if a statement is offensive to the demographic group that it targets, when that group may be a small fraction of the annotator pool. We construct a model that predicts individual annotator ratings on potentially offensive text and combines this information with the predicted target group of the text to predict the ratings of target group members. We show gains across a range of metrics, including raising performance over the baseline by 22% at predicting individual annotators' ratings and by 33% at predicting variance among annotators, which provides a metric for model uncertainty downstream. We find that annotators' ratings can be predicted using their demographic information as well as opinions on online content, and that non-invasive questions on annotators' online experiences minimize the need to collect demographic information when predicting annotators' opinions.

11:00-12:30 (East Foyer / Virtual)
**#136 Let's Think Frame by Frame with VIP: A Video Infilling and Prediction Dataset for Evaluating Video Chain-of-Thought**
*Vaishnavi Himakunthala, Andy Ouyang, Daniel Philip Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon and William Yang Wang*
Despite exciting recent results showing vision-language systems' capacity to reason about images using natural language, their capacity for video reasoning remains underexplored. We motivate framing video reasoning as the sequential understanding of a small number of keyframes, thereby leveraging the power and robustness of vision-language while alleviating the computational complexities of processing videos. To evaluate this novel application, we introduce VIP, an inference-time challenge dataset designed to explore models' reasoning capabilities through video chain-of-thought. Inspired by visually descriptive scene plays, we propose two formats for keyframe description: unstructured dense captions and structured scene descriptions that identify the focus, action, mood, objects, and setting (FAMOuS) of the keyframe. To evaluate video reasoning, we propose two tasks: Video Infilling and Video Prediction, which test abilities to generate multiple intermediate keyframes and predict future keyframes, respectively. We benchmark GPT-4, GPT-3, and VICUNA on VIP, demonstrate the performance gap in these complex video reasoning tasks, and encourage future work to prioritize language models for efficient and generalized video reasoning.

11:00-12:30 (East Foyer / Virtual)
**#137 Struct-XLM: A Structure Discovery Multilingual Language Model for Enhancing Cross-lingual Transfer through Reinforcement Learning**
*Linjuan Wu and Weiming Lu*
Cross-lingual transfer learning heavily relies on well-aligned cross-lingual representations. The syntactic structure is recognized as beneficial for cross-lingual transfer, but limited researches utilize it for aligning representation in multilingual pre-trained language models (PLMs). Additionally, existing methods require syntactic labels that are difficult to obtain and of poor quality for low-resource languages. To address this gap, we propose Struct-XLM, a novel multilingual language model that leverages reinforcement learning (RL) to autonomously discover universal syntactic structures for improving the cross-lingual representation alignment of PLM. Struct-XLM integrates a policy network (PNet) and a translation ranking task. The PNet is designed to discover structural information and integrate it into the last layer of the PLM through the structural multi-head attention module to obtain structural representation. The translation ranking task obtains a delayed reward based on the structural representation to optimize the PNet while improving the alignment of cross-lingual representation. Experiments show the effectiveness of the proposed approach for enhancing cross-lingual transfer of multilingual PLM on the XTREME benchmark.

11:00-12:30 (East Foyer / Virtual)
**#138 GlobalBench: A Benchmark for Global Progress in Natural Language Processing**
*Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos and Graham Neubig*
Despite the major advances in NLP, significant disparities in NLP system performance across languages still exist. Arguably, these are due to uneven resource allocation and sub-optimal incentives to work on less resourced languages. To track and further incentivize the global development of equitable language technology, we introduce GlobalBench. Prior multilingual benchmarks are static and have focused on a limited number of tasks and languages. In contrast, GlobalBench is an ever-expanding collection that aims to dynamically track progress on all NLP datasets in all languages. Rather than solely measuring accuracy, GlobalBench also tracks the estimated per-speaker utility and equity of technology across all languages, providing a multi-faceted view of how language technology is serving people of the world. Furthermore, GlobalBench is designed to identify the most under-served languages, and rewards research efforts directed towards those languages. At present, the most under-served languages are the ones with a relatively high population, but nonetheless overlooked by composite multilingual benchmarks (like Punjabi, Portuguese, and Wu Chinese). Currently, GlobalBench covers 966 datasets in 190 languages, and has 1,128 system submissions spanning 62 languages.

11:00-12:30 (East Foyer / Virtual)
**#139 Towards Building More Robust NER datasets: An Empirical Study on NER Dataset Bias from a Dataset Difficulty View**
*Ruotian Ma, Xiaolei Wang, Xin Zhou, Qi Zhang and Xuanjing Huang*
Recently, many studies have illustrated the robustness problem of Named Entity Recognition (NER) systems: the NER models often rely on superficial entity patterns for predictions, without considering evidence from the context. Consequently, even state-of-the-art NER models generalize poorly to out-of-domain scenarios when out-of-distribution (OOD) entity patterns are introduced. Previous research attributes the robustness problem to the existence of NER dataset bias, where simpler and regular entity patterns induce shortcut learning. In this work, we bring new insights into this problem by comprehensively investigating the NER dataset bias from a dataset difficulty view. We quantify the entity-context difficulty distribution in existing datasets and explain their relationship with model robustness. Based on our findings, we explore three potential ways to de-bias the NER datasets by altering entity-context distribution, and we validate the feasibility with intensive experiments. Finally, we show that the de-biased datasets can transfer to different models and even benefit existing model-based robustness-improving methods, indicating that building more robust datasets is fundamental for building more robust NER systems.

11:00-12:30 (East Foyer / Virtual)

**#140 ALDi: Quantifying the Arabic Level of Dialectness of Text**
*Amr Keleg, Sharon Goldwater and Walid Magdy*
Transcribed speech and user-generated text in Arabic typically contain a mixture of Modern Standard Arabic (MSA), the standardized language taught in schools, and Dialectal Arabic (DA), used in daily communications. To handle this variation, previous work in Arabic NLP has focused on Dialect Identification (DI) on the sentence or the token level. However, DI treats the task as binary, whereas we argue that Arabic speakers perceive a spectrum of dialectness, which we operationalize at the sentence level as the Arabic Level of Dialectness (ALDi), a continuous linguistic variable. We introduce the AOC-ALDi dataset (derived from the AOC dataset), containing 127,835 sentences (17% from news articles and 83% from user comments on those articles) which are manually labeled with their level of dialectness. We provide a detailed analysis of AOC-ALDi and show that a model trained on it can effectively identify levels of dialectness on a range of other corpora (including dialects and genres not included in AOC-ALDi), providing a more nuanced picture than traditional DI systems. Through case studies, we illustrate how ALDi can reveal Arabic speakers' stylistic choices in different situations, a useful property for sociolinguistic analyses.

11:00-12:30 (East Foyer / Virtual)
**#141 Learning to Rank Context for Named Entity Recognition Using a Synthetic Dataset**
*Arthur Amalvy, Vincent Labatut and Richard Dufour*
While recent pre-trained transformer-based models can perform named entity recognition (NER) with great accuracy, their limited range remains an issue when applied to long documents such as whole novels. To alleviate this issue, a solution is to retrieve relevant context at the document level. Unfortunately, the lack of supervision for such a task means one has to settle for unsupervised approaches. Instead, we propose to generate a synthetic context retrieval training dataset using Alpaca, an instruction-tuned large language model (LLM). Using this dataset, we train a neural context retriever based on a BERT model that is able to find relevant context for NER. We show that our method outperforms several retrieval baselines for the NER task on an English literary dataset composed of the first chapter of 40 books.

11:00-12:30 (East Foyer / Virtual)
**#142 From Dissonance to Insights: Dissecting Disagreements in Rationale Construction for Case Outcome Classification**
*Shanshan Xu, Santosh T.Y.S.S, Oana Ichim, Isabella Risini, Barbara Plank and Matthias Grabmair*
In legal NLP, Case Outcome Classification (COC) must not only be accurate but also trustworthy and explainable. Existing work in explainable COC has been limited to annotations by a single expert. However, it is well-known that lawyers may disagree in their assessment of case facts. We hence collect a novel dataset RaVE: Rationale Variation in ECHR, which is obtained from two experts in the domain of international human rights law, for whom we observe weak agreement. We study their disagreements and build a two-level task-independent taxonomy, supplemented with COC-specific subcategories. To our knowledge, this is the first work in the legal NLP that focuses on human label variation. We quantitatively assess different taxonomy categories and find that disagreements mainly stem from underspecification of the legal context, which poses challenges given the typically limited granularity and noise in COC metadata. We further assess the explainablility of state-of-the-art COC models on RaVE and observe limited agreement between models and experts. Overall, our case study reveals hitherto underappreciated complexities in creating benchmark datasets in legal NLP that revolve around identifying aspects of a case's facts supposedly relevant for its outcome.

11:00-12:30 (East Foyer / Virtual)
**#143 Language Model Quality Correlates with Psychometric Predictive Power in Multiple Languages**
*Ethan Wilcox, Clara Meister, Ryan Cotterell and Tiago Pimentel*
Surprisal theory (Hale, 2001; Levy, 2008) posits that a word's reading time is proportional to its surprisal (i.e., to its negative log probability given the proceeding context). Since we are unable to access a word's ground-truth probability, surprisal theory has been empirically tested using surprisal estimates from language models (LMs). Under the premise that surprisal theory holds, we would expect that higher quality language models provide more powerful predictors of human reading behavior—a conjecture we dub the quality–power (QP) hypothesis. Unfortunately, empirical support for the QP hypothesis is mixed. Some studies in English have found correlations between LM quality and predictive power, but other studies using Japanese data, as well as using larger English LMs, find no such correlations. In this work, we conduct a systematic crosslinguistic assessment of the QP hypothesis. We train LMs from scratch on small- and medium-sized datasets from 13 languages (across five language families) and assess their ability to predict eye tracking data. We find correlations between LM quality and power in eleven of these thirteen languages, suggesting that, within the range of model classes and sizes tested, better language models are indeed better predictors of human language processing behaviors.

11:00-12:30 (East Foyer / Virtual)
**#144 Clustering Pseudo Language Family in Multilingual Translation Models with Fisher Information Matrix**
*Xinyu Ma, Xuebo Liu and Min Zhang*
In multilingual translation research, the comprehension and utilization of language families are of paramount importance. Nevertheless, clustering languages based solely on their ancestral families can yield suboptimal results due to variations in the datasets employed during the model's training phase. To mitigate this challenge, we introduce an innovative method that leverages the fisher information matrix (FIM) to cluster language families, anchored on the multilingual translation model's characteristics. We hypothesize that language pairs with similar effects on model parameters exhibit a considerable degree of linguistic congruence and should thus be grouped cohesively. This concept has led us to define pseudo language families. We provide an in-depth discussion regarding the inception and application of these pseudo language families. Empirical evaluations reveal that employing these pseudo language families enhances performance over conventional language families in adapting a multilingual translation model to unfamiliar language pairs. The proposed methodology may also be extended to scenarios requiring language similarity measurements. The source code and associated scripts can be accessed at https://github.com/ecolihit/PseudoFamily.

11:00-12:30 (East Foyer / Virtual)
**#145 HyperNetwork-based Decoupling to Improve Model Generalization for Few-Shot Relation Extraction**
*Liang Zhang, Chulun Zhou, Fandong Meng, Jinsong Su, Yidong Chen and Jie Zhou*
Few-shot relation extraction (FSRE) aims to train a model that can deal with new relations using only a few labeled examples. Most existing studies employ Prototypical Networks for FSRE, which usually overfits the relation classes in the training set and cannot generalize well to unseen relations. By investigating the class separation of an FSRE model, we find that model upper layers are prone to learn relation-specific knowledge. Therefore, in this paper, we propose a HyperNetwork-based Decoupling approach to improve the generalization of FSRE models. Specifically, our model consists of an encoder, a network generator (for producing relation classifiers) and the produced-then-finetuned classifiers for every N-way-K-shot episode. Meanwhile, we design a two-step training framework along with a class-agnostic aligner, in which the generated classifiers focus on acquiring relation-specific knowledge and the encoder is encouraged to learn more general relation knowledge. In this way, the roles of upper and lower layers in an FSRE model are explicitly decoupled, thus enhancing its generalizing capability during testing. Experiments on two public datasets demonstrate the effectiveness of our method.

11:00-12:30 (East Foyer / Virtual)
**#146 When Reviewers Lock Horns: Finding Disagreements in Scientific Peer Reviews**

*Sandeep Kumar, Tirthankar Ghosal and Asif Ekbal*
To this date, the efficacy of the scientific publishing enterprise fundamentally rests on the strength of the peer review process. The journal editor or the conference chair primarily relies on the expert reviewers' assessment, *identify points of agreement and disagreement* and try to reach a consensus to make a fair and informed decision on whether to accept or reject a paper. However, with the escalating number of submissions requiring review, especially in top-tier Artificial Intelligence (AI) conferences, the editor/chair, among many other works, invests a significant, sometimes stressful effort to mitigate reviewer disagreements. Here in this work, we introduce a novel task of automatically identifying contradictions among reviewers on a given article. To this end, we introduce *ContraSciView*, a comprehensive review-pair contradiction dataset on around 8.5k papers (with around 28k review pairs containing nearly 50k review pair comments) from the open review-based ICLR and NeurIPS conferences. We further propose a baseline model that detects contradictory statements from the review pairs. To the best of our knowledge, we make the first attempt to identify disagreements among peer reviewers automatically. We make our dataset and code public for further investigations.

11:00-12:30 (East Foyer / Virtual)
**#147 Adaptive End-to-End Metric Learning for Zero-Shot Cross-Domain Slot Filling**
*Yuanjun Shi, Linzhi Wu and Minglai Shao*
Recently slot filling has witnessed great development thanks to deep learning and the availability of large-scale annotated data. However, it poses a critical challenge to handle a novel domain whose samples are never seen during training. The recognition performance might be greatly degraded due to severe domain shifts. Most prior works deal with this problem in a two-pass pipeline manner based on metric learning. In practice, these dominant pipeline models may be limited in computational efficiency and generalization capacity because of non-parallel inference and context-free discrete label embeddings. To this end, we re-examine the typical metric-based methods, and propose a new adaptive end-to-end metric learning scheme for the challenging zero-shot slot filling. Considering simplicity, efficiency and generalizability, we present a cascade-style joint learning framework coupled with context-aware soft label representations and slot-level contrastive representation learning to mitigate the data and label shift problems effectively. Extensive experiments on public benchmarks demonstrate the superiority of the proposed approach over a series of competitive baselines.

11:00-12:30 (East Foyer / Virtual)
**#148 Detecting Spoilers in Movie Reviews with External Movie Knowledge and User Networks**
*Heng Wang, Wenqian Zhang, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Qinghua Zheng and Minnan Luo*
Online movie review platforms are providing crowdsourced feedback for the film industry and the general public, while spoiler reviews greatly compromise user experience. Although preliminary research efforts were made to automatically identify spoilers, they merely focus on the review content itself, while robust spoiler detection requires putting the review into the context of facts and knowledge regarding movies, user behavior on film review platforms, and more. In light of these challenges, we first curate a large-scale network-based spoiler detection dataset LCS and a comprehensive and up-to-date movie knowledge base UKM. We then propose MVSD, a novel spoiler detection model that takes into account the external knowledge about movies and user activities on movie review platforms. Specifically, MVSD constructs three interconnecting heterogeneous information networks to model diverse data sources and their multi-view attributes, while we design and employ a novel heterogeneous graph neural network architecture for spoiler detection as node-level classification. Extensive experiments demonstrate that MVSD advances the state-of-the-art on two spoiler detection datasets, while the introduction of external knowledge and user interactions help ground robust spoiler detection.

11:00-12:30 (East Foyer / Virtual)
**#149 NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models**
*Yongchao Chen, Rujul Gandhi, Yang Zhang and Chuchu Fan*
Temporal Logic (TL) can be used to rigorously specify complex high-level specification for systems in many engineering applications. The translation between natural language (NL) and TL has been under-explored due to the lack of dataset and generalizable model across different application domains. In this paper, we propose an accurate and generalizable transformation framework of English instructions from NL to TL, exploring the use of Large Language Models (LLMs) at multiple stages. Our contributions are twofold. First, we develop a framework to create a dataset of NL-TL pairs combining LLMs and human annotation. We publish a dataset with 23K NL-TL pairs. Then, we finetune T5 models on the lifted versions (i.e., the specific Atomic Propositions (AP) are hidden) of the NL and TL. The enhanced generalizability originates from two aspects: 1) Usage of lifted NL-TL characterizes common logical structures, without constraints of specific domains. 2) Application of LLMs in dataset creation largely enhances corpus richness. We test the generalization of trained models on five varied domains. To achieve full NL-TL transformation, we either combine the lifted model with AP recognition task or do the further finetuning on each specific domain. During the further finetuning, our model achieves higher accuracy (> 95%) using only <10% training data, compared with the baseline sequence to sequence (Seq2Seq) model.

11:00-12:30 (East Foyer / Virtual)
**#150 HyperRank: Hyperbolic Ranking Model for Unsupervised Keyphrase Extraction**
*Mingyang Song, Huafeng Liu and Liping Jing*
Given the exponential growth in the number of documents on the web in recent years, there is an increasing demand for accurate models to extract keyphrases from such documents. Keyphrase extraction is the task of automatically identifying representative keyphrases from the source document. Typically, candidate keyphrases exhibit latent hierarchical structures embedded with intricate syntactic and semantic information. Moreover, the relationships between candidate keyphrases and the document also form hierarchical structures. Therefore, it is essential to consider these latent hierarchical structures when extracting keyphrases. However, many recent unsupervised keyphrase extraction models overlook this aspect, resulting in incorrect keyphrase extraction. In this paper, we address this issue by proposing a new hyperbolic ranking model (HyperRank). HyperRank is designed to jointly model global and local context information for estimating the importance of each candidate keyphrase within the hyperbolic space, enabling accurate keyphrase extraction. Experimental results demonstrate that HyperRank significantly outperforms recent state-of-the-art baselines.

11:00-12:30 (East Foyer / Virtual)
**#151 A Picture is Worth a Thousand Words: Language Models Plan from Pixels**
*Anthony Zhe Liu, Lajanugen Logeswaran, Sungryull Sohn and Honglak Lee*
Planning is an important capability of artificial agents that perform long-horizon tasks in real-world environments. In this work, we explore the use of pre-trained language models (PLMs) to reason about plan sequences from text instructions in embodied visual environments. Prior PLM based approaches for planning either assume observations are available in the form of text by a captioning model, reason about plans from the instruction alone, or incorporate information about the visual environment in limited ways (such as a pre-trained affordance function). In contrast, we show that the PLM can accurately plan even when observations are directly encoded as input prompts for the PLM. We show this simple approach outperforms prior approaches in experiments on the ALFWorld and VirtualHome benchmarks.

11:00-12:30 (East Foyer / Virtual)
**#152 Reader: Model-based language-instructed reinforcement learning**

*Nicola Dainese, Pekka Marttinen and Alexander Ilin*
We explore how we can build accurate world models, which are partially specified by language, and how we can plan with them in the face of novelty and uncertainty. We propose the first model-based reinforcement learning approach to tackle the environment Read To Fight Monsters (Zhong et al., 2019), a grounded policy learning problem. In RTFM an agent has to reason over a set of rules and a goal, both described in a language manual, and the observations, while taking into account the uncertainty arising from the stochasticity of the environment, in order to generalize successfully its policy to test episodes. We demonstrate the superior performance and sample efficiency of our model-based approach to the existing model-free SOTA agents in eight variants of RTFM. Furthermore, we show how the agent's plans can be inspected, which represents progress towards more interpretable agents.

11:00-12:30 (East Foyer / Virtual)
**#153 GenEx: A Commonsense-aware Unified Generative Framework for Explainable Cyberbullying Detection**
*Krishanu Maity, Raghav Jain, Prince Jha, Sriparna Saha and Pushpak Bhattacharyya*
With the rise of social media and online communication, the issue of cyberbullying has gained significant prominence. While extensive research is being conducted to develop more effective models for detecting cyberbullying in monolingual languages, a significant gap exists in understanding code-mixed languages and the need for explainability in this context. To address this gap, we have introduced a novel benchmark dataset named BullyExplain for explainable cyberbullying detection in code-mixed language. In this dataset, each post is meticulously annotated with four labels: bully, sentiment, target, and rationales, indicating the specific phrases responsible for identifying the post as a bully. Our current research presents an innovative unified generative framework, GenEx, which reimagines the multitask problem as a text-to-text generation task. Our proposed approach demonstrates its superiority across various evaluation metrics when applied to the BullyExplain dataset, surpassing other baseline models and current state-of-the-art approaches.

11:00-12:30 (East Foyer / Virtual)
**#154 Fine-grained Medical Vision-Language Representation Learning for Radiology Report Generation**
*Siyuan Wang, Bo Peng, Yichao Liu and Qi Peng*
Given the input radiology images, the objective of radiology report generation is to produce accurate and comprehensive medical reports, which typically include multiple descriptive clinical sentences associated with different phenotypes. Most existing works have relied on a pre-trained vision encoder to extract the visual representations of the images. In this study, we propose a phenotype-driven medical vision-language representation learning framework to efficiently bridge the gap between visual and textual modalities for improved text-oriented generation. In contrast to conventional methods which learn medical vision-language representations by contrasting images with entire reports, our approach learns more fine-grained representations by contrasting images with each sentence within the reports. The learned fine-grained representations can be used to improve radiology report generation. The experiments on two widely-used datasets MIMIC-CXR and IU X-ray demonstrate that our method can achieve promising performances and substantially outperform the conventional vision-language representation learning methods.

11:00-12:30 (East Foyer / Virtual)
**#155 A Cross-Linguistic Pressure for Uniform Information Density in Word Order**
*Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell and Roger Levy*
While natural languages differ widely in both canonical word order and word order flexibility, their word orders still follow shared cross-linguistic statistical patterns, often attributed to functional pressures. In the effort to identify these pressures, prior work has compared real and counterfactual word orders. Yet one functional pressure has been overlooked in such investigations: the uniform information density (UID) hypothesis, which holds that information should be spread evenly throughout an utterance. Here, we ask whether a pressure for UID may have influenced word order patterns cross-linguistically. To this end, we use computational models to test whether real orders lead to greater information uniformity than counterfactual orders. In our empirical study of 10 typologically diverse languages, we find that: (i) among SVO languages, real word orders consistently have greater uniformity than reverse word orders, and (ii) only linguistically implausible counterfactual orders consistently exceed the uniformity of real orders. These findings are compatible with a pressure for information uniformity in the development and usage of natural languages.

11:00-12:30 (East Foyer / Virtual)
**#156 T2-NER: A Two-Stage Span-based Framework For Unified Named Entity Recognition with Templates**
*Minghao Hu, Peixin Huang, Xiang Zhao, Zhen Tan and Weidong Xiao*
Named Entity Recognition (NER) has so far evolved from the traditional flat NER to the overlapped and discontinuous NER. They have mostly been solved separately, with only several exceptions that concurrently tackle three tasks with a single model. Current best-performing method formalizes the unified NER as word-word relation classification, which barely focuses on mention content learning and fails to detect entity mentions comprising a single word. In this paper, we propose a two-stage span-based framework with templates, namely T2-NER, to resolve the unified NER task. The first stage is to extract entity spans, where flat and overlapped entities can be recognized. The second stage is to classify over all entity span pairs, where discontinuous entities can be recognized. Finally, multi-task learning is used to jointly train two stages. To improve the efficiency of span-based model, we design grouped templates and typed templates for two stages to realize batch computations. We also apply an adjacent packing strategy and a latter packing strategy to model discriminative boundary information and learn better span (pair) representation. Moreover, we introduce the syntax information to enhance our span representation. We perform extensive experiments on eight benchmark datasets for flat, overlapped, and discontinuous NER, where our model beats all the current competitive baselines, obtaining the best performances of unified NER.

11:00-12:30 (East Foyer / Virtual)
**#157 Testing the Predictions of Surprisal Theory in 11 Languages**
*Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell and Roger Levy*
A fundamental result in psycholinguistics is that less predictable words take a longer time to process. One theoretical explanation for this finding is Surprisal Theory (Hale, 2001; Levy, 2008), which quantifies a word's predictability as its surprisal, i.e. its negative log probability given a context. While evidence supporting the predictions of Surprisal Theory have been replicated widely, most have focused on a very narrow slice of data: native English speakers reading English texts. Indeed, no comprehensive multilingual analysis exists. We address this gap in the current literature by investigating the relationship between surprisal and reading times in eleven different languages, distributed across five language families. Deriving estimates from language models trained on monolingual and multilingual corpora, we test three predictions associated with surprisal theory: (i) whether surprisal is predictive of reading times; (ii) whether expected surprisal, i.e. contextual entropy, is predictive of reading times; (iii) and whether the linking function between surprisal and reading times is linear. We find that all three predictions are borne out crosslinguistically. By focusing on a more diverse set of languages, we argue that these results offer the most robust link to-date between information theory and incremental language processing across languages.

11:00-12:30 (East Foyer / Virtual)
**#158 U-CORE: A Unified Deep Cluster-wise Contrastive Framework for Open Relation Extraction**
*Hongkui Tu, Jie Zhou, Shenpo Dong, Yunxin Huang, Meihan Wu, Haili Li, Jingnan Wang and Xiaodong Wang*

Within Open Relation Extraction (ORE) tasks, the Zero-shot ORE method is to generalize undefined relations from predefined relations, while the Unsupervised ORE method is to extract undefined relations without the need for annotations. However, despite the possibility of overlap between predefined and undefined relations in the training data, a unified framework for both Zero-shot and Unsupervised ORE has yet to be established. To address this gap, we propose U-CORE: A Unified Deep Cluster-wise Contrastive Framework for both Zero-shot and Unsupervised ORE by leveraging techniques from Contrastive Learning (CL) and Clustering. U-CORE overcomes the limitations of CL-based Zero-shot ORE methods by employing Cluster-wise CL that preserves both local smoothness as well as global semantics. Additionally, we employ a deep-cluster-based updater that optimizes the cluster center, thus enhancing the accuracy and efficiency of the model. To increase the stability of the model, we adopt Adaptive Self-paced Learning that effectively addresses the data-shifting problems. Experimental results on three well-known datasets demonstrate that U-CORE significantly improves upon existing methods by showing an average improvement of 7.35% ARI on Zero-shot ORE tasks and 15.24% ARI on Unsupervised ORE tasks.

11:00-12:30 (East Foyer / Virtual)
**#159 Language Varieties of Italy: Technology Challenges and Opportunities**
*Alan Ramponi*
Italy is characterized by a one-of-a-kind linguistic diversity landscape in Europe, which implicitly encodes local knowledge, cultural traditions, artistic expressions and history of its speakers. However, most local languages and dialects in Italy are at risk of disappearing within few generations. The NLP community has recently begun to engage with endangered languages, including those of Italy. Yet, most efforts assume that these varieties are under-resourced language monoliths with an established written form and homogeneous functions and needs, and thus highly interchangeable with each other and with high-resource, standardized languages. In this paper, we introduce the linguistic context of Italy and challenge the default machine-centric assumptions of NLP for Italy's language varieties. We advocate for a shift in the paradigm from machine-centric to speaker-centric NLP, and provide recommendations and opportunities for work that prioritizes languages and their speakers over technological advances. To facilitate the process, we finally propose building a local community towards responsible, participatory efforts aimed at supporting vitality of languages and dialects of Italy.

11:00-12:30 (East Foyer / Virtual)
**#160 mGPT: Few-Shot Learners Go Multilingual**
*Tatiana Shavrina, Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov and Tatiana Shavrina*
Pretrained autoregressive language models (LMs) can successfully solve many NLP tasks via zero-shot and few-shot learning. This paper introduces mGPT, a multilingual variant of GPT-3, pretrained on 61 languages from linguistically diverse 25 language families using Wikipedia and C4 Corpus. We detail the design and pretraining procedure. The models undergo an intrinsic and extrinsic evaluation: language modeling in all languages, downstream evaluation on cross-lingual NLU datasets and benchmarks in 33 languages, and world knowledge probing in 23 languages. The in-context learning abilities are on par with the contemporaneous LMs while covering a larger amount of languages, including underrepresented and low-resource languages of the Commonwealth of Independent States and the small peoples in Russia. The source code, pretraining data, and models are either publicly available or will be released upon acceptance.

11:00-12:30 (East Foyer / Virtual)
**#161 Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation**
*Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Martins, Amanda Bertsch, José Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig and André Martins*
Natural language generation has witnessed significant advancements due to the training of large language models on vast internet-scale datasets. Despite these advancements, there exists a critical challenge: these models can inadvertently generate content that is toxic, inaccurate, and unhelpful, and existing automatic evaluation metrics often fall short of identifying these shortcomings.As models become more capable, human feedback is an invaluable signal for evaluating and improving models. This survey aims to provide an overview of recent research that has leveraged human feedback to improve natural language generation. First, we introduce a taxonomy distilled from existing research to categorize and organize the varied forms of feedback. Next, we discuss how feedback can be described by its format and objective, and cover the two approaches proposed to use feedback (either for training or decoding): directly using feedback or training feedback models. We also discuss existing datasets for human-feedback data collection, and concerns surrounding feedback collection. Finally, we provide an overview of the nascent field of AI feedback, which uses large language models to make judgments based on a set of principles and minimize the need for human intervention. We also release a website of this survey at feedback-gap-survey.info

## Findings 1

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**Zero-Shot-BERT-Adapters: a Zero-Shot Pipeline for Unknown Intent Detection**
*Daniele Comi, Dimitrios Christofidellis, Pier Francesco Piazza and Matteo Manica*
Intent discovery is a crucial task in natural language processing, and it is increasingly relevant for various of industrial applications. Identifying novel, unseen intents from user inputs remains one of the biggest challenges in this field. Herein, we propose Zero-Shot-BERT-Adapters, a two-stage method for multilingual intent discovery relying on a Transformer architecture, fine-tuned with Adapters. We train the model for Natural Language Inference (NLI) and later perform unknown intent classification in a zero-shot setting for multiple languages. In our evaluation, we first analyze the quality of the model after adaptive fine-tuning on known classes. Secondly, we evaluate its performance in casting intent classification as an NLI task. Lastly, we test the zero-shot performance of the model on unseen classes, showing how Zero-Shot-BERT-Adapters can effectively perform intent discovery by generating semantically similar intents, if not equal, to the ground-truth ones. Our experiments show how Zero-Shot-BERT-Adapters outperforms various baselines in two zero-shot settings: known intent classification and unseen intent discovery. The proposed pipeline holds the potential for broad application in customer care. It enables automated dynamic triage using a lightweight model that can be easily deployed and scaled in various business scenarios, unlike large language models. Zero-Shot-BERT-Adapters represents an innovative multi-language approach for intent discovery, enabling the online generation of novel intents. A Python package implementing the pipeline and the new datasets we compiled are available at the following link: https://github.com/GT4SD/zero-shot-bert-adapters.

11:00-12:30 (East Foyer / Virtual)
**LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain**
*Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer and Ilias Chalkidis*
Lately, propelled by phenomenal advances around the transformer architecture, the legal NLP field has enjoyed spectacular growth. To measure progress, well-curated and challenging benchmarks are crucial. Previous efforts have produced numerous benchmarks for general NLP models, typically based on news or Wikipedia. However, these may not fit specific domains such as law, with its unique lexicons and intricate

sentence structures. Even though there is a rising need to build NLP systems for languages other than English, many benchmarks are available only in English and no multilingual benchmark exists in the legal NLP field. We survey the legal NLP literature and select 11 datasets covering 24 languages, creating LEXTREME. To fairly compare models, we propose two aggregate scores, i.e., dataset aggregate score and language aggregate score. Our results show that even the best baseline only achieves modest results, and also ChatGPT struggles with many tasks. This indicates that LEXTREME remains a challenging task with ample room for improvement. To facilitate easy use for researchers and practitioners, we release LEXTREME on huggingface along with a public leaderboard and the necessary code to evaluate models. We also provide a public Weights and Biases project containing all runs for transparency.

11:00-12:30 (East Foyer / Virtual)
**FactSpotter: Evaluating the Factual Faithfulness of Graph-to-Text Generation**
*Kun Zhang, Oana Balalau and Ioana Manolescu*
Graph-to-text (G2T) generation takes a graph as input and aims to generate a fluent and faith- ful textual representation of the information in the graph. The task has many applications, such as dialogue generation and question an- swering. In this work, we investigate to what extent the G2T generation problem is solved for previously studied datasets, and how pro- posed metrics perform when comparing generated texts. To help address their limitations, we propose a new metric that correctly identifies factual faithfulness, i.e., given a triple (subject, predicate, object), it decides if the triple is present in a generated text. We show that our metric FactSpotter achieves the highest correlation with human annotations on data correct- ness, data coverage, and relevance. In addition, FactSpotter can be used as a plug-in feature to improve the factual faithfulness of existing models. Finally, we investigate if existing G2T datasets are still challenging for state-of-the-art models. Our code is available online: https://github.com/guihuzhang/FactSpotter.

11:00-12:30 (East Foyer / Virtual)
**CREATOR: Tool Creation for Disentangling Abstract and Concrete Reasoning of Large Language Models**
*Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu and Heng Ji*
Large Language Models (LLMs) have made significant progress in utilizing tools, but their ability is limited by API availability and the instability of implicit reasoning, particularly when both planning and execution are involved. To overcome these limitations, we propose CREATOR, a novel framework that enables LLMs to create their own tools using documentation and code realization. CREATOR disentangles abstract tool creation and concrete decision execution, resulting in improved performance. We evaluate CREATOR on MATH and TabMWP benchmarks, respectively consisting of challenging math competition problems and diverse tabular contents. Remarkably, CREATOR outperforms existing chain-of-thought, program-of-thought, and tool-using baselines. Additionally, we introduce the Creation Challenge dataset, featuring 2K diverse questions, to emphasize the necessity and benefits of LLMs' tool creation ability. Further research demonstrates that leveraging LLMs as tool creators facilitates knowledge transfer, and LLMs exhibit varying levels of tool creation abilities, enabling them to adapt to diverse situations. The tool creation ability revolutionizes the LLM's problem-solving paradigm, driving us closer to the next frontier of artificial intelligence.

11:00-12:30 (East Foyer / Virtual)
**Contrastive Pre-training for Personalized Expert Finding**
*Qiyao Peng, Hongtao Liu, Zhepeng Lv, Qing Yang and Wenjun Wang*
Expert finding could help route questions to potential suitable users to answer in Community Question Answering (CQA) platforms. Hence it is essential to learn accurate representations of experts and questions according to the question text articles. Recently the pre-training and fine-tuning paradigms are powerful for natural language understanding, which has the potential for better question modeling and expert finding. Inspired by this, we propose a CQA-domain Contrastive Pre-training framework for Expert Finding, named CPEF, which could learn more comprehensive question representations. Specifically, considering that there is semantic complementation between question titles and bodies, during the domain pre-training phase, we propose a title-body contrastive learning task to enhance question representations, which directly treats the question title and the corresponding body as positive samples of each other, instead of designing extra data-augmentation strategies. Furthermore, a personalized tuning network is proposed to inject the personalized preferences of different experts during the fine-tuning phase. Extensive experimental results on six real-world datasets demonstrate that our method could achieve superior performance for expert finding.

11:00-12:30 (East Foyer / Virtual)
**Culturally Aware Natural Language Inference**
*Jing Huang and Diyi Yang*
Humans produce and consume language in a particular cultural context, which includes knowledge about specific norms and practices. A listener's awareness of the cultural context is critical for interpreting the speaker's meaning. A simple expression like *I didn't leave a tip* implies a strong sense of dissatisfaction when tipping is assumed to be the norm. As NLP systems reach users from different cultures, achieving culturally aware language understanding becomes increasingly important. However, current research has focused on building cultural knowledge bases without studying how such knowledge leads to contextualized interpretations of texts. In this work, we operationalize cultural variations in language understanding through a natural language inference (NLI) task that surfaces cultural variations as label disagreement between annotators from different cultural groups. We introduce the first Culturally Aware Natural Language Inference (CALI) dataset with 2.7K premise-hypothesis pairs annotated by two cultural groups located in the U.S. and India. With CALI, we categorize how cultural norms affect language understanding and present an evaluation framework to assess at which levels large language models are culturally aware. Our dataset is available at https://github.com/SALT-NLP/CulturallyAwareNLI.

11:00-12:30 (East Foyer / Virtual)
**Mixture-of-Linguistic-Experts Adapters for Improving and Interpreting Pre-trained Language Models**
*Raymond Li, Gabriel Murray and Giuseppe Carenini*
In this work, we propose a method that combines two popular research areas by injecting linguistic structures into pre-trained language models in the parameter-efficient fine-tuning (PEFT) setting. In our approach, parallel adapter modules encoding different linguistic structures are combined using a novel Mixture-of-Linguistic-Experts architecture, where Gumbel-Softmax gates are used to determine the importance of these modules at each layer of the model. To reduce the number of parameters, we first train the model for a fixed small number of steps before pruning the experts based on their important scores. Our experiment results with three different pre-trained models show that our approach can outperform state-of-the-art PEFT methods with a comparable number of parameters. In addition, we provide additional analysis to examine the experts selected by each model at each layer to provide insights for future studies.

11:00-12:30 (East Foyer / Virtual)
**Toward Joint Language Modeling for Speech Units and Text**
*Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski and Michael Auli*
Speech and text are two major forms of human language. The research community has been focusing on mapping speech to text or vice versa for many years. However, in the field of language modeling, very little effort has been made to model them jointly. In light of this, we explore joint language modeling for speech units and text. Specifically, we compare different speech tokenizers to transform continuous

speech signals into discrete units and use different methods to construct mixed speech-text data. We introduce automatic metrics to evaluate how well the joint LM mixes speech and text. We also fine-tune the LM on downstream spoken language understanding (SLU) tasks with different modalities (speech or text) and test its performance to assess the model's learning of shared representations. Our results show that by mixing speech units and text with our proposed mixing techniques, the joint LM improves over a speech-only baseline on SLU tasks and shows zero-shot cross-modal transferability.

11:00-12:30 (East Foyer / Virtual)
**Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate**
*Boshi Wang, Xiang Yue and Huan Sun*
Large language models (LLMs) such as ChatGPT and GPT-4 have shown impressive performance in complex reasoning tasks. However, it is difficult to know whether the models are reasoning based on deep understandings of truth and logic, or leveraging their memorized patterns in a relatively superficial way. In this work, we explore testing LLMs' reasoning by engaging with them in a debate-like conversation, where given a question, the LLM and the user need to discuss to make the correct decision starting from opposing arguments. Upon mitigating the Clever Hans effect, our task requires the LLM to not only achieve the correct answer on its own, but also be able to hold and defend its belief instead of blindly believing or getting misled by the user's (invalid) arguments and critiques, thus testing in greater depth whether the LLM grasps the essence of the reasoning required to solve the problem. Across a range of complex reasoning benchmarks spanning math, commonsense, logic and BIG-Bench tasks, we find that despite their impressive performance as reported in existing work on generating correct step-by-step solutions in the beginning, LLMs like ChatGPT cannot maintain their beliefs in truth for a significant portion of examples when challenged by oftentimes absurdly invalid arguments. Our work points to danger zones of model alignment, and also suggests more careful treatments and interpretations of the recent findings that LLMs can improve their responses based on feedback.

11:00-12:30 (East Foyer / Virtual)
**Three Questions Concerning the Use of Large Language Models to Facilitate Mathematics Learning**
*An-Zi Yen and Wei-Ling Hsu*
Due to the remarkable language understanding and generation abilities of large language models (LLMs), their use in educational applications has been explored. However, little work has been done on investigating the pedagogical ability of LLMs in helping students to learn mathematics. In this position paper, we discuss the challenges associated with employing LLMs to enhance students' mathematical problem-solving skills by providing adaptive feedback. Apart from generating the wrong reasoning processes, LLMs can misinterpret the meaning of the question, and also exhibit difficulty in understanding the given questions' rationales when attempting to correct students' answers. Three research questions are formulated.

11:00-12:30 (East Foyer / Virtual)
**Swap and Predict – Predicting the Semantic Changes in Words across Corpora by Context Swapping**
*Taichi Aida and Danushka Bollegala*
Meanings of words change over time and across domains. Detecting the semantic changes of words is an important task for various NLP applications that must make time-sensitive predictions. We consider the problem of predicting whether a given target word, $w$, changes its meaning between two different text corpora, $\mathcal{C}_1$ and $\mathcal{C}_2$. For this purpose, we propose *Swapping-based Semantic Change Detection* (SSCD), an unsupervised method that randomly swaps contexts between $\mathcal{C}_1$ and $\mathcal{C}_2$ where $w$ occurs. We then look at the distribution of contextualised word embeddings of $w$, obtained from a pretrained masked language model (MLM), representing the meaning of $w$ in its occurrence contexts in $\mathcal{C}_1$ and $\mathcal{C}_2$. Intuitively, if the meaning of $w$ does not change between $\mathcal{C}_1$ and $\mathcal{C}_2$, we would expect the distributions of contextualised word embeddings of $w$ to remain the same before and after this random swapping process. Despite its simplicity, we demonstrate that even by using pretrained MLMs without any fine-tuning, our proposed context swapping method accurately predicts the semantic changes of words in four languages (English, German, Swedish, and Latin) and across different time spans (over 50 years and about five years). Moreover, our method achieves significant performance improvements compared to strong baselines for the English semantic change prediction task. Source code is available at https://github.com/a1da4/svp-swap .

11:00-12:30 (East Foyer / Virtual)
**Improving word mover's distance by leveraging self-attention matrix**
*Hiroaki Yamagiwa, Sho Yokoi and Hidetoshi Shimodaira*
Measuring the semantic similarity between two sentences is still an important task. The word mover's distance (WMD) computes the similarity via the optimal alignment between the sets of word embeddings. However, WMD does not utilize word order, making it challenging to distinguish sentences with significant overlaps of similar words, even if they are semantically very different. Here, we attempt to improve WMD by incorporating the sentence structure represented by BERT's self-attention matrix (SAM). The proposed method is based on the Fused Gromov-Wasserstein distance, which simultaneously considers the similarity of the word embedding and the SAM for calculating the optimal transport between two sentences. Experiments demonstrate the proposed method enhances WMD and its variants in paraphrase identification with near-equivalent performance in semantic textual similarity.

11:00-12:30 (East Foyer / Virtual)
**Global Structure Knowledge-Guided Relation Extraction Method for Visually-Rich Document**
*Xiangnan Chen, Qian Xiao, Juncheng Li, Duo Dong, Jun Lin, Xiaozhong Liu and Siliang Tang*
Visual Relation Extraction (VRE) is a powerful means of discovering relationships between entities within visually-rich documents. Existing methods often focus on manipulating entity features to find pairwise relations, yet neglect the more fundamental structural information that links disparate entity pairs together. The absence of global structure information may make the model struggle to learn long-range relations and easily predict conflicted results. To alleviate such limitations, we propose a GlObal Structure knowledge-guided relation Extraction (GOSE) framework. GOSE initiates by generating preliminary relation predictions on entity pairs extracted from a scanned image of the document. Subsequently, global structural knowledge is captured from the preceding iterative predictions, which are then incorporated into the representations of the entities. This "generate-capture-incorporate" cycle is repeated multiple times, allowing entity representations and global structure knowledge to be mutually reinforced. Extensive experiments validate that GOSE not only outperforms existing methods in the standard fine-tuning setting but also reveals superior cross-lingual learning capabilities; indeed, even yields stronger data-efficient performance in the low-resource setting.

11:00-12:30 (East Foyer / Virtual)
**Ask Language Model to Clean Your Noisy Translation Data**
*Quinten Bolding, Baohao Liao, Brandon James Denis, Jun Luo and Christof Monz*
TTransformer models have demonstrated remarkable performance in neural machine translation (NMT). However, their vulnerability to noisy input poses a significant challenge in practical implementation, where generating clean output from noisy input is crucial. The MTNT dataset is widely used as a benchmark for evaluating the robustness of NMT models against noisy input. Nevertheless, its utility is limited due to the presence of noise in both the source and target sentences. To address this limitation, we focus on cleaning the noise from the target sentences in MTNT, making it more suitable as a benchmark for noise evaluation. Leveraging the capabilities of large language models (LLMs), we

observe their impressive abilities in noise removal. For example, they can remove emojis while considering their semantic meaning. Additionally, we show that LLM can effectively rephrase slang, jargon, and profanities. The resulting datasets, called C-MTNT, exhibit significantly less noise in the target sentences while preserving the semantic integrity of the original sentences. Our human and GPT-4 evaluations also lead to a consistent conclusion that LLM performs well on this task. Lastly, experiments on C-MTNT showcased its effectiveness in evaluating the robustness of NMT models, highlighting the potential of advanced language models for data cleaning and emphasizing C-MTNT as a valuable resource.

11:00-12:30 (East Foyer / Virtual)
**CCIM: Cross-modal Cross-lingual Interactive Image Translation**
*Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou and Chengqing Zong*
Text image machine translation (TIMT) which translates source language text images into target language texts has attracted intensive attention in recent years. Although the end-to-end TIMT model directly generates target translation from encoded text image features with an efficient architecture, it lacks the recognized source language information resulting in a decrease in translation performance. In this paper, we propose a novel Cross-modal Cross-lingual Interactive Model (CCIM) to incorporate source language information by synchronously generating source language and target language results through an interactive attention mechanism between two language decoders. Extensive experimental results have shown the interactive decoder significantly outperforms end-to-end TIMT models and has faster decoding speed with smaller model size than cascade models.

11:00-12:30 (East Foyer / Virtual)
**SimCKP: Simple Contrastive Learning of Keyphrase Representations**
*Minseok Choi, Chaeheon Gwak, Seho Kim, Si Hyeong Kim and Jaegul Choo*
Keyphrase generation (KG) aims to generate a set of summarizing words or phrases given a source document, while keyphrase extraction (KE) aims to identify them from the text. Because the search space is much smaller in KE, it is often combined with KG to predict keyphrases that may or may not exist in the corresponding document. However, current unified approaches adopt sequence labeling and maximization-based generation that primarily operate at a token level, falling short in observing and scoring keyphrases as a whole. In this work, we propose SimCKP, a simple contrastive learning framework that consists of two stages: 1) An extractor-generator that extracts keyphrases by learning context-aware phrase-level representations in a contrastive manner while also generating keyphrases that do not appear in the document; 2) A reranker that adapts scores for each generated phrase by likewise aligning their representations with the corresponding document. Experimental results on multiple benchmark datasets demonstrate the effectiveness of our proposed approach, which outperforms the state-of-the-art models by a significant margin.

11:00-12:30 (East Foyer / Virtual)
**mReFinED: An Efficient End-to-End Multilingual Entity Linking System**
*Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari and Jens Lehmann*
End-to-end multilingual entity linking (MEL) is concerned with identifying multilingual entity mentions and their corresponding entity IDs in a knowledge base. Existing works assumed that entity mentions were given and skipped the entity mention detection step due to a lack of high-quality multilingual training corpora. To overcome this limitation, we propose mReFinED, the first end-to-end multilingual entity linking. Additionally, we propose a bootstrapping mention detection framework that enhances the quality of training corpora. Our experimental results demonstrated that mReFinED outperformed the best existing work in the end-to-end MEL task while being 44 times faster.

11:00-12:30 (East Foyer / Virtual)
**Distilling ChatGPT for Explainable Automated Student Answer Assessment**
*Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi and Yulan He*
Providing explainable and faithful feedback is crucial for automated student answer assessment. In this paper, we introduce a novel framework that explores using ChatGPT, a cutting-edge large language model, for the concurrent tasks of student answer scoring and rationale generation. We identify the appropriate instructions by prompting ChatGPT with different templates to collect the rationales, where inconsistent rationales are refined to align with marking standards. The refined ChatGPT outputs enable us to fine-tune a smaller language model that simultaneously assesses student answers and provides rationales. Extensive experiments on the benchmark dataset show that the proposed method improves the overall QWK score by 11% compared to ChatGPT. Furthermore, our thorough analysis and human evaluation demonstrate that the rationales generated by our proposed method are comparable to those of ChatGPT. Our approach provides a viable solution to achieve explainable automated assessment in education

11:00-12:30 (East Foyer / Virtual)
**A Lightweight Method to Generate Unanswerable Questions in English**
*Vagrant Gautam, Miaoran Zhang and Dietrich Klakow*
If a question cannot be answered with the available information, robust systems for question answering (QA) should know *not* to answer. One way to build QA models that do this is with additional training data comprised of unanswerable questions, created either by employing annotators or through automated methods for unanswerable question generation. To show that the model complexity of existing automated approaches is not justified, we examine a simpler data augmentation method for unanswerable question generation in English: performing antonym and entity swaps on answerable questions. Compared to the prior state-of-the-art, data generated with our training-free and lightweight strategy results in better models (+1.6 F1 points on SQuAD 2.0 data with BERT-large), and has higher human-judged relatedness and readability. We quantify the raw benefits of our approach compared to no augmentation across multiple encoder models, using different amounts of generated data, and also on TydiQA-MinSpan data (+9.3 F1 points with BERT-large). Our results establish swaps as a simple but strong baseline for future work.

11:00-12:30 (East Foyer / Virtual)
**Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation**
*Xi Wang, Hossein A. Rahmani, Jiqun Liu and Emine Yilmaz*
Conversational Recommendation System (CRS) is a rapidly growing research area that has gained significant attention alongside advancements in language modelling techniques. However, the current state of conversational recommendation faces numerous challenges due to its relative novelty and limited existing contributions. In this study, we delve into benchmark datasets for developing CRS models and address potential biases arising from the feedback loop inherent in multi-turn interactions, including selection bias and multiple popularity bias variants. Drawing inspiration from the success of generative data via using language models and data augmentation techniques, we present two novel strategies, 'Once-Aug' and 'PopNudge', to enhance model performance while mitigating biases. Through extensive experiments on ReDial and TG-ReDial benchmark datasets, we show a consistent improvement of CRS techniques with our data augmentation approaches and offer additional insights on addressing multiple newly formulated biases.

11:00-12:30 (East Foyer / Virtual)
**Automatic Evaluation of Attribution by Large Language Models**

*Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su and Huan Sun*
A recent focus of large language model (LLM) development, as exemplified by generative search engines, is to incorporate external references to generate and support its claims. However, evaluating the attribution, i.e., verifying whether the generated statement is fully supported by the cited reference, remains an open problem. Although human evaluation is common practice, it is costly and time-consuming. In this paper, we investigate automatic evaluation of attribution given by LLMs. We begin by defining different types of attribution errors, and then explore two approaches for automatic evaluation: prompting LLMs and fine-tuning smaller LMs. The fine-tuning data is repurposed from related tasks such as question answering, fact-checking, natural language inference, and summarization. We manually curate a set of test examples covering 12 domains from a generative search engine, New Bing. Our results on this curated test set and simulated examples from existing benchmarks highlight both promising signals and challenges. We hope our problem formulation, testbeds, and findings will help lay the foundation for future studies on this important problem.

11:00-12:30 (East Foyer / Virtual)
**TK-KNN: A Balanced Distance-Based Pseudo Labeling Approach for Semi-Supervised Intent Classification**
*Nicholas Botzer, David Vazquez, Tim Weninger and Issam H. Laradji*
The ability to detect intent in dialogue systems has become increasingly important in modern technology. These systems often generate a large amount of unlabeled data, and manually labeling this data requires substantial human effort. Semi-supervised methods attempt to remedy this cost by using a model trained on a few labeled examples and then by assigning pseudo-labels to further a subset of unlabeled examples that has a model prediction confidence higher than a certain threshold. However, one particularly perilous consequence of these methods is the risk of picking an imbalanced set of examples across classes, which could lead to poor labels. In the present work, we describe Top-K K-Nearest Neighbor (TK-KNN), which uses a more robust pseudo-labeling approach based on distance in the embedding space while maintaining a balanced set of pseudo-labeled examples across classes through a ranking-based approach. Experiments on several datasets show that TK-KNN outperforms existing models, particularly when labeled data is scarce on popular datasets such as CLINC150 and Banking77.

11:00-12:30 (East Foyer / Virtual)
**LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis**
*Shih-Chieh Dai, Aiping Xiong and Lun-Wei Ku*
Thematic analysis (TA) has been widely used for analyzing qualitative data in many disciplines and fields. To ensure reliable analysis, the same piece of data is typically assigned to at least two human coders. Moreover, to produce meaningful and useful analysis, human coders develop and deepen their data interpretation and coding over multiple iterations, making TA labor-intensive and time-consuming. Recently the emerging field of large language models (LLMs) research has shown that LLMs have the potential replicate human-like behavior in various tasks: in particular, LLMs outperform crowd workers on text-annotation tasks, suggesting an opportunity to leverage LLMs on TA. We propose a human–LLM collaboration framework (i.e., LLM-in-the-loop) to conduct TA with in-context learning (ICL). This framework provides the prompt to frame discussions with a LLM (e.g., GPT-3.5) to generate the final codebook for TA. We demonstrate the utility of this framework using survey datasets on the aspects of the music listening experience and the usage of a password manager. Results of the two case studies show that the proposed framework yields similar coding quality to that of human coders but reduces TA's labor and time demands.

11:00-12:30 (East Foyer / Virtual)
**PR-MCS: Perturbation Robust Metric for MultiLingual Image Captioning**
*Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, Trung Bui and Kyomin Jung*
Vulnerability to lexical perturbation is a critical weakness of automatic evaluation metrics for image captioning. This paper proposes Perturbation Robust Multi-Lingual CLIPScore(PR-MCS), which exhibits robustness to such perturbations, as a novel reference-free image captioning metric applicable to multiple languages. To achieve perturbation robustness, we fine-tune the text encoder of CLIP with our language-agnostic method to distinguish the perturbed text from the original text. To verify the robustness of PR-MCS, we introduce a new fine-grained evaluation dataset consisting of detailed captions, critical objects, and the relationships between the objects for 3,000 images in five languages. In our experiments, PR-MCS significantly outperforms baseline metrics in capturing lexical noise of all various perturbation types in all five languages, while maintaining a strong correlation with human judgments.

11:00-12:30 (East Foyer / Virtual)
**Do Stochastic Parrots have Feelings Too? Improving Neural Detection of Synthetic Text via Emotion Recognition**
*Alan Cowap, Yvette Graham and Jennifer Foster*
Recent developments in generative AI have shone a spotlight on high-performance synthetic text generation technologies. The now wide availability and ease of use of such models highlights the urgent need to provide equally powerful technologies capable of identifying synthetic text. With this in mind, we draw inspiration from psychological studies which suggest that people can be driven by emotion and encode emotion in the text they compose. We hypothesize that pretrained language models (PLMs) have an affective deficit because they lack such an emotional driver when generating text and consequently may generate synthetic text which has affective incoherence i.e. lacking the kind of emotional coherence present in human-authored text. We subsequently develop an emotionally aware detector by fine-tuning a PLM on emotion. Experiment results indicate that our emotionally-aware detector achieves improvements across a range of synthetic text generators, various sized models, datasets, and domains. Finally, we compare our emotionally-aware synthetic text detector to ChatGPT in the task of identification of its own output and show substantial gains, reinforcing the potential of emotion as a signal to identify synthetic text. Code, models, and datasets are available at https: //github.com/alanagiasi/emoPLMsynth

11:00-12:30 (East Foyer / Virtual)
**Can Large Language Models Fix Data Annotation Errors? An Empirical Study Using Debatepedia for Query-Focused Text Summarization**
*Md Tahmid Rahman Laskar, Mizanur Rahman, Israt Jahan, Enamul Hoque and Jimmy Huang*
Debatepedia is a publicly available dataset consisting of arguments and counter-arguments on controversial topics that has been widely used for the single-document query-focused abstractive summarization task in recent years. However, it has been recently found that this dataset is limited by noise and even most queries in this dataset do not have any relevance to the respective document. In this paper, we study whether large language models (LLMs) can be utilized to clean the Debatepedia dataset to make it suitable for query-focused abstractive summarization. More specifically, we harness the language generation capabilities of two LLMs, namely, ChatGPT and PaLM to regenerate its queries. Based on our experiments, we find that solely depending on large language models for query correction may not be very useful for data cleaning. However, we observe that leveraging a rule-based approach for data sampling followed by query regeneration using LLMs (especially ChatGPT) for the sampled instances may ensure a higher quality version of this dataset suitable for the development of more generalized query-focused text summarization models.

11:00-12:30 (East Foyer / Virtual)
**A Joint Matrix Factorization Analysis of Multilingual Representations**
*Zheng Zhao, Yftah Ziser, Bonnie L. Webber and Shay B Cohen*
We present an analysis tool based on joint matrix factorization for comparing latent representations of multilingual and monolingual mod-

els. An alternative to probing, this tool allows us to analyze multiple sets of representations in a joint manner. Using this tool, we study to what extent and how morphosyntactic features are reflected in the representations learned by multilingual pre-trained models. We conduct a large-scale empirical study of over 33 languages and 17 morphosyntactic categories. Our findings demonstrate variations in the encoding of morphosyntactic information across upper and lower layers, with category-specific differences influenced by language properties. Hierarchical clustering of the factorization outputs yields a tree structure that is related to phylogenetic trees manually crafted by linguists. Moreover, we find the factorization outputs exhibit strong associations with performance observed across different cross-lingual tasks. We release our code to facilitate future research.

11:00-12:30 (East Foyer / Virtual)
**Less than One-shot: Named Entity Recognition via Extremely Weak Supervision**
*Letian Peng, Zihan Wang and Jingbo Shang*
We study the named entity recognition (NER) problem under the extremely weak supervision (XWS) setting, where only one example entity per type is given in a context-free way. While one can see that XWS is *lighter than one-shot* in terms of the amount of supervision, we propose a novel method X-NER that can outperform the state-of-the-art one-shot NER methods. We first mine entity spans that are similar to the example entities from an unlabelled training corpus. Instead of utilizing entity span representations from language models, we find it more effective to compare the context distributions before and after the span is replaced by the entity example. We then leverage the top-ranked spans as pseudo-labels to train an NER tagger. Extensive experiments and analyses on 4 NER datasets show the superior end-to-end NER performance of X-NER, outperforming the state-of-the-art few-shot methods with 1-shot supervision and ChatGPT annotations significantly. Finally, our X-NER possesses several notable properties, such as inheriting the cross-lingual abilities of the underlying language models.

11:00-12:30 (East Foyer / Virtual)
**SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF**
*Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu and Oleksii Kuchaiev*
Model alignment with human preferences is an essential step in making Large Language Models (LLMs) helpful and consistent with human values. It typically consists of supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) stages. However, RLHF faces inherent limitations stemming from a complex training setup and its tendency to align the model with implicit values that end users cannot control at run-time. Moreover, reward models in RLHF stage commonly rely on single-dimensional feedback as opposed to explicit, multifaceted signals that indicate attributes such as helpfulness, humor, and toxicity. To address these limitations, we propose SteerLM, a supervised fine-tuning method that empowers end-users to control responses during inference. SteerLM conditions responses to conform to an explicitly defined multi-dimensional set of attributes, thereby empowering a steerable AI capable of generating helpful and high-quality responses while maintaining customizability. Experiments show that SteerLM trained on open source datasets generates responses that are preferred by human and automatic evaluators to many state-of-the-art baselines trained with RLHF while being much easier to train. Try SteerLM at https://huggingface.co/nvidia/SteerLM-llama2-13B

11:00-12:30 (East Foyer / Virtual)
**A Framework for Bidirectional Decoding: Case Study in Morphological Inflection**
*Marc Canby and Julia Hockenmaier*
Transformer-based encoder-decoder models that generate outputs in a left-to-right fashion have become standard for sequence-to-sequence tasks. In this paper, we propose a framework for decoding that produces sequences from the "outside-in": at each step, the model chooses to generate a token on the left, on the right, or join the left and right sequences. We argue that this is more principled than prior bidirectional decoders. Our proposal supports a variety of model architectures and includes several training methods, such as a dynamic programming algorithm that marginalizes out the latent ordering variable. Our model sets state-of-the-art (SOTA) on the 2022 and 2023 shared tasks, beating the next best systems by over 4.7 and 2.7 points in average accuracy respectively. The model performs particularly well on long sequences, can implicitly learn the split point of words composed of stem and affix, and performs better relative to the baseline on datasets that have fewer unique lemmas.

11:00-12:30 (East Foyer / Virtual)
**Legally Enforceable Hate Speech Detection for Public Forums**
*Chu Fei Luo, Rohan V Bhambhoria, Samuel Dahan and Xiaodan Zhu*
Hate speech causes widespread and deep-seated societal issues. Proper enforcement of hate speech laws is key for protecting groups of people against harmful and discriminatory language. However, determining what constitutes hate speech is a complex task that is highly open to subjective interpretations. Existing works do not align their systems with enforceable definitions of hate speech, which can make their outputs inconsistent with the goals of regulators. This research introduces a new perspective and task for enforceable hate speech detection centred around legal definitions, and a dataset annotated on violations of eleven possible definitions by legal experts. Given the challenge of identifying clear, legally enforceable instances of hate speech, we augment the dataset with expert-generated samples and an automatically mined challenge set. We experiment with grounding the model decision in these definitions using zero-shot and few-shot prompting. We then report results on several large language models (LLMs). With this task definition, automatic hate speech detection can be more closely aligned to enforceable laws, and hence assist in more rigorous enforcement of legal protections against harmful speech in public forums.

11:00-12:30 (East Foyer / Virtual)
**Evaluating Verifiability in Generative Search Engines**
*Nelson F. Liu, Tianyi Zhang and Percy Liang*
Generative search engines directly generate responses to user queries, along with in-line citations. A prerequisite trait of a trustworthy generative search engine is verifiability, i.e., systems should cite comprehensively (high citation recall; all statements are fully supported by citations) and accurately (high citation precision; every cite supports its associated statement). We conduct human evaluation to audit four popular generative search engines—Bing Chat, NeevaAI, perplexity.ai, and YouChat—across a diverse set of queries from a variety of sources (e.g., historical Google user queries, dynamically-collected open-ended questions on Reddit, etc.). We find that responses from existing generative search engines are fluent and appear informative, but frequently contain unsupported statements and inaccurate citations: on average, a mere 51.5% of generated sentences are fully supported by citations and only 74.5% of citations support their associated sentence. We believe that these results are concerningly low for systems that may serve as a primary tool for information-seeking users, especially given their facade of trustworthiness. We hope that our results further motivate the development of trustworthy generative search engines and help researchers and users better understand the shortcomings of existing commercial systems.

11:00-12:30 (East Foyer / Virtual)
**Multi-User MultiWOZ: Task-Oriented Dialogues among Multiple Users**
*Yohan Jo, Xinyan Zhao, Arijit Biswas, Nikoletta Basiou, Vincent Auvray, Nikolaos Malandrakis, Angeliki Metallinou and Alexandros Potamianos*
While most task-oriented dialogues assume conversations between the agent and one user at a time, dialogue systems are increasingly expected to communicate with multiple users simultaneously who make decisions collaboratively. To facilitate development of such systems,

we release the Multi-User MultiWOZ dataset: task-oriented dialogues among two users and one agent. To collect this dataset, each user utterance from MultiWOZ 2.2 was replaced with a small chat between two users that is semantically and pragmatically consistent with the original user utterance, thus resulting in the same dialogue state and system response. These dialogues reflect interesting dynamics of collaborative decision-making in task-oriented scenarios, e.g., social chatter and deliberation. Supported by this data, we propose the novel task of multi-user contextual query rewriting: to rewrite a task-oriented chat between two users as a concise task-oriented query that retains only task-relevant information and that is directly consumable by the dialogue system. We demonstrate that in multi-user dialogues, using predicted rewrites substantially improves dialogue state tracking without modifying existing dialogue systems that are trained for single-user dialogues. Further, this method surpasses training a medium-sized model directly on multi-user dialogues and generalizes to unseen domains.

11:00-12:30 (East Foyer / Virtual)
**An Adaptive Prompt Generation Framework for Task-oriented Dialogue System**
*Jun Gao, Liuyu Xiang, Huijia Wu, Han Zhao, Yiqi Tong and Zhaofeng He*
The de facto way of utilizing black-box large language models (LLMs) to perform various downstream tasks is prompting. However, obtaining suitable prompts for specific tasks is still a challenging problem. While existing LLM-based methods demonstrate promising performance in task-oriented dialogue (TOD) task, they often require manual adjustment in prompt selection, or focus solely on dialogue understanding or generation. To address these issues, we propose an adaptive prompt generation framework to fully unleash the potential of LLMs for the comprehensive TOD system. Firstly, we design a trainable slot generator (TSG) that can generate domain and slot information in the belief state, which serves as prior knowledge for subsequent prompt generation. Next, we propose an adaptive prompt generator (APG) that utilizes the prior knowledge to generate prompts for the LLM, deriving the belief state and system response of the dialogue for evaluation. Finally, we evaluate our framework on the MultiWOZ 2.0 dataset. Extensive experiments demonstrate that our method outperforms existing methods. Our code and data will be released.

11:00-12:30 (East Foyer / Virtual)
**DeTiME: Diffusion-Enhanced Topic Modeling using Encoder-decoder based LLM**
*Weijie Xu, Wenxiang Hu, Fanyou Wu and Srinivasan H. Sengamedu*
In the burgeoning field of natural language processing, Neural Topic Models (NTMs) and Large Language Models (LLMs) have emerged as areas of significant research interest. Despite this, NTMs primarily utilize contextual embeddings from LLMs, which are not optimal for clustering or capable for topic generation. Our study addresses this gap by introducing a novel framework named Diffusion-Enhanced Topic Modeling using Encoder-Decoder-based LLMs (DeTiME). DeTiME leverages Encoder-Decoder-based LLMs to produce highly clusterable embeddings that could generate topics that exhibit both superior clusterability and enhanced semantic coherence compared to existing methods. Additionally, by exploiting the power of diffusion, our framework also provides the capability to generate content relevant to the identified topics. This dual functionality allows users to efficiently produce highly clustered topics and related content simultaneously. DeTiME's potential extends to generating clustered embeddings as well. Notably, our proposed framework proves to be efficient to train and exhibits high adaptability, demonstrating its potential for a wide array of applications.

11:00-12:30 (East Foyer / Virtual)
**Task-Attentive Transformer Architecture for Continual Learning of Vision-and-Language Tasks Using Knowledge Distillation**
*Yuliang Cai, Jesse Thomason and Mohammad Rostami*
The size and the computational load of fine-tuning large-scale pre-trained neural network are becoming two major obstacles in adopting machine learning in many applications. Continual learning (CL) can serve as a remedy through enabling knowledge-transfer across sequentially arriving tasks which relaxes the need to fine-tune all network weights from scratch. However, existing CL algorithms primarily consider learning unimodal vision-only or language-only tasks. We develop a transformer-based CL architecture for learning bimodal vision-and-language tasks based on increasing the number of the learnable parameters dynamically and using knowledge distillation. The new additional parameters are used to specialize the network for each task. Our approach enables sharing information between the tasks while addressing the challenge of catastrophic forgetting. Our approach is scalable learning to a large number of tasks because it requires little memory and time overhead. Our model reaches state-of-the-art performance on challenging vision-and-language tasks.

11:00-12:30 (East Foyer / Virtual)
**Boot and Switch: Alternating Distillation for Zero-Shot Dense Retrieval**
*Fan Jiang, Qiongkai Xu, Tom Drummond and Trevor Cohn*
Neural 'dense' retrieval models are state of the art for many datasets, however these models often exhibit limited domain transfer ability. Existing approaches to adaptation are unwieldy, such as requiring explicit supervision, complex model architectures, or massive external models. We present `ABEL`, a simple but effective unsupervised method to enhance passage retrieval in zero-shot settings. Our technique follows a straightforward loop: a dense retriever learns from supervision signals provided by a reranker, and subsequently, the reranker is updated based on feedback from the improved retriever. By iterating this loop, the two components mutually enhance one another's performance. Experimental results demonstrate that our unsupervised `ABEL` model outperforms both leading supervised and unsupervised retrievers on the BEIR benchmark. Meanwhile, it exhibits strong adaptation abilities to tasks and domains that were unseen during training. By either fine-tuning `ABEL` on labelled data or integrating it with existing supervised dense retrievers, we achieve state-of-the-art results.

11:00-12:30 (East Foyer / Virtual)
**Text Classification via Large Language Models**
*Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang and Guoyin Wang*
Despite the remarkable success of large-scale Language Models (LLMs) such as GPT-3, their performances still significantly underperform fine-tuned models in the task of text classification.This is due to (1) the lack of reasoning ability in addressing complex linguistic phenomena (e.g., intensification, contrast, irony etc); (2) limited number of tokens allowed in in-context learning. In this paper, we introduce **C**lue **A**nd **R**easoning **P**rompting (CARP). CARP adopts a progressive reasoning strategy tailored to addressing the complex linguistic phenomena involved in text classification: CARP first prompts LLMs to find superficial clues (e.g., keywords, tones, semantic relations, references, etc), based on which a diagnostic reasoning process is induced for final decisions. To further address the limited-token issue, CARP uses a fine-tuned model on the supervised dataset for *k*NN demonstration search in the in-context learning, allowing the model to take the advantage of both LLM's generalization ability and the task-specific evidence provided by the full labeled dataset. Remarkably, CARP yields new SOTA performances on 4 out of 5 widely-used text-classification benchmarks, 97.39 (+1.24) on SST-2, 96.40 (+0.72) on AGNews, 98.78 (+0.25) on R8 and 96.95 (+0.6) on R52, and a performance comparable to SOTA on MR (92.39 v.s. 93.3). More importantly, we find that CARP delivers impressive abilities on low-resource and domain-adaptation setups. Specifically, using 16 examples per class, CARP achieves comparable performances to supervised models with 1,024 examples per class.

11:00-12:30 (East Foyer / Virtual)
**Uncovering Limitations in Text-to-Image Generation: A Contrastive Approach with Structured Semantic Alignment**
*Qianyu Feng, Yulei Sui and Hongyu Zhang*
Despite significant advancements in text-to-image generation models, they still face challenges when it comes to producing highly detailed

or complex images based on textual descriptions. In order to explore these limitations, we propose a Structured Semantic Alignment (SSA) method for evaluating text-to-image generation models. SSA focuses on learning structured semantic embeddings across different modalities and aligning them in a joint space. The method employs the following steps to achieve its objective: (i) Generating mutated prompts by substituting words with semantically equivalent or nonequivalent alternatives while preserving the original syntax; (ii) Representing the sentence structure through parsing trees obtained via syntax parsing; (iii) Learning fine-grained structured embeddings that project semantic features from different modalities into a shared embedding space; (iv) Evaluating the semantic consistency between the structured text embeddings and the corresponding visual embeddings. Through experiments conducted on various benchmarks, we have demonstrated that SSA offers improved measurement of semantic consistency of text-to-image generation models. Additionally, it unveils a wide range of generation errors including under-generation, incorrect constituency, incorrect dependency, and semantic confusion. By uncovering these biases and limitations embedded within the models, our proposed method provides valuable insights into their shortcomings when applied to real-world scenarios.

11:00-12:30 (East Foyer / Virtual)
**Proto-lm: A Prototypical Network-Based Framework for Built-in Interpretability in Large Language Models**
*Sean Xie, Soroush Vosoughi and Saeed Hassanpour*
Large Language Models (LLMs) have significantly advanced the field of Natural Language Processing (NLP), but their lack of interpretability has been a major concern. Current methods for interpreting LLMs are post hoc, applied after inference time, and have limitations such as their focus on low-level features and lack of explainability at higher-level text units. In this work, we introduce proto-lm, a prototypical network-based white-box framework that allows LLMs to learn immediately interpretable embeddings during the fine-tuning stage while maintaining competitive performance. Our method's applicability and interpretability are demonstrated through experiments on a wide range of NLP tasks, and our results indicate a new possibility of creating interpretable models without sacrificing performance. This novel approach to interpretability in LLMs can pave the way for more interpretable models without the need to sacrifice performance. We release our code at https://github.com/yx131/proto-lm.

11:00-12:30 (East Foyer / Virtual)
**Improving End-to-End Speech Processing by Efficient Text Data Utilization with Latent Synthesis**
*Jianqiao Lu, Wenyong Huang, Nianzu Zheng, Xingshan Zeng, Yu Ting Yeung and Xiao Chen*
Training a high performance end-to-end speech (E2E) processing model requires an enormous amount of labeled speech data, especially in the era of data-centric artificial intelligence. However, labeled speech data are usually scarcer and more expensive for collection, compared to textual data. We propose Latent Synthesis (LaSyn), an efficient textual data utilization framework for E2E speech processing models. We train a latent synthesizer to convert textual data into an intermediate latent representation of a pre-trained speech model. These pseudo acoustic representations of textual data augment acoustic data for model training. We evaluate LaSyn on low-resource automatic speech recognition (ASR) and spoken language understanding (SLU) tasks. For ASR, LaSyn improves an E2E baseline trained on LibriSpeech train-clean-100, with relative word error rate reductions over 22.3% on different test sets. For SLU, LaSyn improves our E2E baseline by absolute 4.1% for intent classification accuracy and 3.8% for slot filling SLU-F1 on SLURP, and absolute 4.49% and 2.25% for exact match (EM) and EM-Tree accuracies on STOP respectively. With fewer parameters, the results of LaSyn are competitive to published state-of-the-art works. The results demonstrate the quality of the augmented training data.

11:00-12:30 (East Foyer / Virtual)
**You Are What You Annotate: Towards Better Models through Annotator Representations**
*Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang and Rada Mihalcea*
Annotator disagreement is ubiquitous in natural language processing (NLP) tasks. There are multiple reasons for such disagreements, including the subjectivity of the task, difficult cases, unclear guidelines, and so on. Rather than simply aggregating labels to obtain data annotations, we instead try to directly model the diverse perspectives of the annotators, and explicitly account for annotators' idiosyncrasies in the modeling process by creating representations for each annotator (*annotator embeddings*) and also their annotations (*annotation embeddings*). In addition, we propose **TID-8**, **T**he **I**nherent **D**isagreement - **8** dataset, a benchmark that consists of eight existing language understanding datasets that have inherent annotator disagreement. We test our approach on TID-8 and show that our approach helps models learn significantly better from disagreements on six different datasets in TID-8 while increasing model size by fewer than 1% parameters. By capturing the unique tendencies and subjectivity of individual annotators through embeddings, our representations prime AI models to be inclusive of diverse viewpoints.

11:00-12:30 (East Foyer / Virtual)
**Disentangling Extraction and Reasoning in Multi-hop Spatial Reasoning**
*Roshanak Mirzaee and Parisa Kordjamshidi*
Spatial reasoning over text is challenging as the models not only need to extract the direct spatial information from the text but also reason over those and infer implicit spatial relations. Recent studies highlight the struggles even large language models encounter when it comes to performing spatial reasoning over text. In this paper, we explore the potential benefits of disentangling the processes of information extraction and reasoning in models to address this challenge. To explore this, we design various models that disentangle extraction and reasoning(either symbolic or neural) and compare them with state-of-the-art(SOTA) baselines with no explicit design for these parts. Our experimental results consistently demonstrate the efficacy of disentangling, showcasing its ability to enhance models' generalizability within realistic data domains.

11:00-12:30 (East Foyer / Virtual)
**Large-Scale and Multi-Perspective Opinion Summarization with Diverse Review Subsets**
*Han Jiang, Rui Wang, Zhihua Wei, Yu Li and Xinpeng Wang*
Opinion summarization is expected to digest larger review sets and provide summaries from different perspectives. However, most existing solutions are deficient in epitomizing extensive reviews and offering opinion summaries from various angles due to the lack of designs for information selection. To this end, we propose SubSumm, a supervised summarization framework for large-scale multi-perspective opinion summarization. SubSumm consists of a review sampling strategy set and a two-stage training scheme. The sampling strategies take sentiment orientation and contrastive information value into consideration, with which the review subsets from different perspectives and quality levels can be selected. Subsequently, the summarizer is encouraged to learn from the sub-optimal and optimal subsets successively in order to capitalize on the massive input. Experimental results on AmaSum and Rotten Tomatoes datasets demonstrate that SubSumm is adept at generating pros, cons, and verdict summaries from hundreds of input reviews. Furthermore, our in-depth analysis verifies that the advanced selection of review subsets and the two-stage training scheme are vital to boosting the summarization performance.

11:00-12:30 (East Foyer / Virtual)
**Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs**
*Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister and Somesh Jha*
Large language models (LLMs) have recently shown great advances in a variety of tasks, including natural language understanding and generation. However, their use in high-stakes decision-making scenarios is still limited due to the potential for errors. *Selective prediction* is a technique that can be used to improve the reliability of the LLMs by allowing them to abstain from making predictions when they are

unsure of the answer. In this work, we propose a novel framework for adaptation with self-evaluation to improve the selective prediction performance of LLMs. Our framework is based on the idea of using parameter-efficient tuning to adapt the LLM to the specific task at hand while improving its ability to perform self-evaluation. We evaluate our method on a variety of question-answering (QA) datasets and show that it outperforms state-of-the-art selective prediction methods. For example, on the CoQA benchmark, our method improves the AUACC from 91.23% to 92.63% and improves the AUROC from 74.61% to 80.25%.

11:00-12:30 (East Foyer / Virtual)
**MAPO: Boosting Large Language Model Performance with Model-Adaptive Prompt Optimization**
*Yuyan Chen, Zhihao Wen, Ge Fan, Zhengyu Chen, Wei Wu, Dayiheng Liu, Zhixu Li, Bang Liu and Yanghua Xiao*
Prompt engineering, as an efficient and effective way to leverage Large Language Models (LLM), has drawn a lot of attention from the research community. The existing research primarily emphasizes the importance of adapting prompts to specific tasks, rather than specific LLMs. However, a good prompt is not solely defined by its wording, but also binds to the nature of the LLM in question. In this work, we first quantitatively demonstrate that different prompts should be adapted to different LLMs to enhance their capabilities across various downstream tasks in NLP. Then we novelly propose a model-adaptive prompt optimizer (MAPO) method that optimizes the original prompts for each specific LLM in downstream tasks. Extensive experiments indicate that the proposed method can effectively refine prompts for an LLM, leading to significant improvements over various downstream tasks.

11:00-12:30 (East Foyer / Virtual)
**Probing the "Creativity" of Large Language Models: Can models produce divergent semantic association?**
*Honghua Chen and Nai Ding*
Large language models possess remarkable capacity for processing language, but it remains unclear whether these models can further generate creative content. The present study aims to investigate the creative thinking of large language models through a cognitive perspective. We utilize the divergent association task (DAT), an objective measurement of creativity that asks models to generate unrelated words and calculates the semantic distance between them. We compare the results across different models and decoding strategies. Our findings indicate that: (1) When using the greedy search strategy, GPT-4 outperforms 96% of humans, while GPT-3.5-turbo exceeds the average human level. (2) Stochastic sampling and temperature scaling are effective to obtain higher DAT scores for models except GPT-4, but face a trade-off between creativity and stability. These results imply that advanced large language models have divergent semantic associations, which is a fundamental process underlying creativity.

11:00-12:30 (East Foyer / Virtual)
**LEGO: A Multi-agent Collaborative Framework with Role-playing and Iterative Feedback for Causality Explanation Generation**
*Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng li, Mengshu Sun and Jun Zhao*
Causality Explanation Generation refers to generate an explanation in natural language given an initial cause-effect pair. It demands rigorous explicit rationales to demonstrate the acquisition of implicit commonsense knowledge, which is unlikely to be easily memorized, making it challenging for large language models since they are often suffering from spurious causal associations when they encounter the content that does not exist in their memory. In this work, we introduce LEGO, a Multi-agent Collaborative Framework with Role-playing and Iterative Feedback for causality explanation generation. Specifically, we treat LLM as character malleable LEGO block and utilize role-playing to assign specific roles to five LLMs. We firstly devise a Fine-grained World Knowledge Integration Module to augment information about tasks for alleviating the phenomenon of spurious causal associations. Then, we leverage an Iterative Feedback and Refinement Module to improve the generated explanation by multi-aspect feedback. Extensive experiments on widely used WIKIWHY and e-CARE datasets show the superiority of our multi-agent framework in terms of reasoning about the causality among cause and effect.

11:00-12:30 (East Foyer / Virtual)
**MultiCMET: A Novel Chinese Benchmark for Understanding Multimodal Metaphor**
*Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang and Hongfei Lin*
Metaphor is a pervasive aspect of human communication, and its presence in multimodal forms has become more prominent with the progress of mass media. However, there is limited research on multimodal metaphor resources beyond the English language. Furthermore, the existing work in natural language processing does not address the exploration of categorizing the source and target domains in metaphors. This omission is significant considering the extensive research conducted in the fields of cognitive linguistics, which emphasizes that a profound understanding of metaphor relies on recognizing the differences and similarities between domain categories. We, therefore, introduce MultiCMET, a multimodal Chinese metaphor dataset, consisting of 13,820 text-image pairs of advertisements with manual annotations of the occurrence of metaphors, domain categories, and sentiments metaphors convey. We also constructed a domain lexicon that encompasses categorizations of metaphorical source domains and target domains and propose a Cascading Domain Knowledge Integration (CDKI) benchmark to detect metaphors by introducing domain-specific lexical features. Experimental results demonstrate the effectiveness of CDKI. The dataset and code are publicly available.

11:00-12:30 (East Foyer / Virtual)
**Context-faithful Prompting for Large Language Models**
*Wenxuan Zhou, Sheng Zhang, Hoifung Poon and Muhao Chen*
Large language models (LLMs) encode parametric knowledge about world facts and have shown remarkable performance in knowledge-driven NLP tasks. However, their reliance on parametric knowledge may cause them to overlook contextual cues, leading to incorrect predictions in context-sensitive NLP tasks (e.g., knowledge acquisition tasks). In this paper, we seek to assess and enhance LLMs' contextual faithfulness in two aspects: knowledge conflict and prediction with abstention. We demonstrate that LLMs' faithfulness can be significantly improved using carefully designed prompting strategies. In particular, we identify opinion-based prompts and counterfactual demonstrations as the most effective methods. Opinion-based prompts reframe the context as a narrator's statement and inquire about the narrator's opinions, while counterfactual demonstrations use instances containing false facts to improve faithfulness in knowledge conflict situations. Neither technique requires additional training. We conduct experiments on three datasets of two standard NLP tasks, machine reading comprehension and relation extraction, and the results demonstrate significant improvement in faithfulness to contexts. Code and data are released at https://github.com/wzhouad/context-faithful-llm.

11:00-12:30 (East Foyer / Virtual)
**Low-Resource Comparative Opinion Quintuple Extraction by Data Augmentation with Prompting**
*Qingting Xu, Yu Hong, Fubang Zhao, Kaisong Song, Yangyang Kang, Jiaxiang Chen and Guodong Zhou*
Comparative Opinion Quintuple Extraction (COQE) aims to predict comparative opinion quintuples from comparative sentences. These quintuples include subject, object, shareable aspect, comparative opinion, and preference. The existing pipeline-based COQE method fails in error propagation. In addition, the complexity and insufficient amounts of annotated data hinder the performance of COQE models. In this paper, we introduce a novel approach called low-resource comparative opinion quintuple extraction by Data Augmentation with Prompting (DAP). Firstly, we present an end-to-end model architecture better suited to the data augmentation method from triplets to quintuples and can effectively avoid error propagation. Additionally, we introduce a data-centric augmentation approach that leverages the robust generative

abilities of ChatGPT and integrates transfer learning techniques. Experimental results over three datasets (Camera, Car, Ele) demonstrate that our approach yields substantial improvements and achieves state-of-the-art results. The source code and data are publicly released at: https://github.com/qtxu-nlp/COQE-DAP.

11:00-12:30 (East Foyer / Virtual)
**GBT: Generative Boosting Training Approach for Paraphrase Identification**
*Rui Peng, Zhiling Jin and Yu Hong*
Paraphrase Identification (PI), a task of determining whether a pair of sentences express the same meaning, is widely applied in Information Retrieval and Question Answering. Data Augmentation (DA) is proven effective in tackling the PI task. However, the majority of DA methods still suffer from two limitations: inefficiency and poor quality. In this study, we propose the Generative Boosting Training (GBT) approach for PI. GBT designs a boosting learning method for a single model based on the human learning process, utilizing seq2seq model to perform DA on misclassified instances periodically. We conduct experiments on the benchmark corpora QQP and LCQMC, towards both English and Chinese PI tasks. Experimental results show that our method yields significant improvements on a variety of Pre-trained Language Model (PLM) based baselines with good efficiency and effectiveness. It is noteworthy that a single BERT model (with a linear classifier) can outperform the state-of-the-art PI models with the boosting of GBT.

11:00-12:30 (East Foyer / Virtual)
**MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic**
*Damien Sileo and Antoine Lernould*
Theory of Mind (ToM) is a critical component of intelligence but its assessment remains the subject of heated debates. Prior research applied human ToM assessments to natural language processing models using either human-created standardized tests or rule-based templates. However, these methods primarily focus on simplistic reasoning and require further validation. Here, we leverage dynamic epistemic logic to isolate a particular component of ToM and to generate controlled problems. We also introduce new verbalization techniques to express these problems in English natural language. Our findings indicate that some language model scaling (from 70M to 6B and 350M to 174B) does not consistently yield results better than random chance. While GPT-4 demonstrates superior epistemic reasoning capabilities, there is still room for improvement. Our code and datasets are publicly available.

11:00-12:30 (East Foyer / Virtual)
**Selective Demonstrations for Cross-domain Text-to-SQL**
*Shuaichen Chang and Eric Fosler-Lussier*
Large language models (LLMs) with in-context learning have demonstrated impressive generalization capabilities in the cross-domain text-to-SQL task, without the use of in-domain annotations. However, incorporating in-domain demonstration examples has been found to greatly enhance LLMs' performance. In this paper, we delve into the key factors within in-domain examples that contribute to the improvement and explore whether we can harness these benefits without relying on in-domain annotations. Based on our findings, we propose a demonstration selection framework, ODIS, which utilizes both out-of-domain examples and synthetically generated in-domain examples to construct demonstrations. By retrieving demonstrations from hybrid sources, ODIS leverages the advantages of both, showcasing its effectiveness compared to baseline methods that rely on a single data source. Furthermore, ODIS outperforms state-of-the-art approaches on two cross-domain text-to-SQL datasets, with improvements of 1.1 and 11.8 points in execution accuracy, respectively.

11:00-12:30 (East Foyer / Virtual)
**Teacher Perception of Automatically Extracted Grammar Concepts for L2 Language Learning**
*Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos and Graham Neubig*
One of the challenges in language teaching is how best to organize rules regarding syntax, semantics, or phonology in a meaningful manner. This not only requires content creators to have pedagogical skills, but also have that language's deep understanding. While comprehensive materials to develop such curricula are available in English and some broadly spoken languages, for many other languages, teachers need to manually create them in response to their students' needs. This is challenging because i) it requires that such experts be accessible and have the necessary resources, and ii) describing all the intricacies of a language is time-consuming and prone to omission. In this work, we aim to facilitate this process by automatically discovering and visualizing grammar descriptions. We extract descriptions from a natural text corpus that answer questions about morphosyntax (learning of word order, agreement, case marking, or word formation) and semantics (learning of vocabulary). We apply this method for teaching two Indian languages, Kannada and Marathi, which, unlike English, do not have well-developed resources for second language learning. To assess the perceived utility of the extracted material, we enlist the help of language educators from schools in North America to perform a manual evaluation, who find the materials have potential to be used for their lesson preparation and learner evaluation.

### Industry 1

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**An Integrated Search System for Korea Weather Data**
*Jinkyung Jo, Dayeon Ki, Soyoung Yoon and Minjoon Seo*
We introduce WeatherSearch, an integrated search system deployed at the Korea Meteorological Administration (KMA). WeatherSearch enables users to retrieve all the relevant data for weather forecasting from a massive weather database with simple natural language queries. We carefully design and conduct multiple expert surveys and interviews for template creation and apply data augmentation techniques including template filling to collect 4 million data points with minimal human labors. We then finetune mT5 on the collected dataset and achieve an average MRR of 0.66 and an average Recall of 0.82. We also discuss weather-data-specific characteristics that should be taken into account for creating such a system. We hope our paper serves as a simple and effective guideline for those designing similar systems in other regions of the world.

## Lunch
12:30-14:00 - Location: Resort World Sentosa

## Session 3: Oral & Poster - 14:00-15:30

## Discourse and Pragmatics

14:00-15:30 (East Ballroom)

14:00-14:15 (East Ballroom)
**COHESENTIA: A Novel Benchmark of Incremental versus Holistic Assessment of Coherence in Generated Texts**
*Aviya Maimon and Reut Tsarfaty*
Coherence is a linguistic term that refers to the relations between small textual units (sentences, propositions), which make the text logically consistent and meaningful to the reader. With the advances of generative foundational models in NLP, there is a pressing need to automatically assess the human-perceived coherence of automatically generated texts. Up until now, little work has been done on explicitly assessing the coherence of generated texts and analyzing the factors contributing to (in)coherence. Previous work on the topic used other tasks, e.g., sentence reordering, as proxies of coherence, rather than approaching coherence detection heads on. In this paper, we introduce COHESENTIA, a novel benchmark of human-perceived coherence of automatically generated texts. Our annotation protocol reflects two perspectives; one is global, assigning a single coherence score, and the other is incremental, scoring sentence by sentence. The incremental method produces an (in)coherence score for each text fragment and also pinpoints reasons for incoherence at that point. Our benchmark contains 500 automatically-generated and human-annotated paragraphs, each annotated in both methods, by multiple raters. Our analysis shows that the inter-annotator agreement in the incremental mode is higher than in the holistic alternative, and our experiments show that standard LMs fine-tuned for coherence detection show varied performance on the different factors contributing to (in)coherence. All in all, these models yield unsatisfactory performance, emphasizing the need for developing more reliable methods for coherence assessment.

14:15-14:30 (East Ballroom)
**Improving Long Document Topic Segmentation Models With Enhanced Coherence Modeling**
*Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen and Wen Wang*
Topic segmentation is critical for obtaining structured documents and improving down- stream tasks such as information retrieval. Due to its ability of automatically exploring clues of topic shift from abundant labeled data, recent supervised neural models have greatly promoted the development of long document topic segmentation, but leaving the deeper relationship between coherence and topic segmentation underexplored. Therefore, this paper enhances the ability of supervised models to capture coherence from both logical structure and semantic similarity perspectives to further improve the topic segmentation performance, proposing Topic-aware Sentence Structure Prediction (TSSP) and Contrastive Semantic Similarity Learning (CSSL). Specifically, the TSSP task is proposed to force the model to comprehend structural information by learning the original relations between adjacent sentences in a disarrayed document, which is constructed by jointly disrupting the original document at topic and sentence levels. Moreover, we utilize inter- and intra-topic information to construct contrastive samples and design the CSSL objective to ensure that the sentences representations in the same topic have higher similarity, while those in different topics are less similar. Extensive experiments show that the Longformer with our approach significantly outperforms old state-of-the-art (SOTA) methods. Our approach improve $F_1$ of old SOTA by 3.42 (73.74 $\rightarrow$ 77.16) and reduces $P_k$ by 1.11 points (15.0 $\rightarrow$ 13.89) on WIKI-727K and achieves an average relative reduction of 4.3% on $P_k$ on WikiSection. The average relative $P_k$ drop of 8.38% on two out-of-domain datasets also demonstrates the robustness of our approach.

14:30-14:45 (East Ballroom)
**Improving Dialogue Discourse Parsing via Reply-to Structures of Addressee Recognition**
*Yaxin Fan, Feng Jiang, Peifeng Li, Fang Kong and Qiaoming Zhu*
Dialogue discourse parsing aims to reflect the relation-based structure of dialogue by establishing discourse links according to discourse relations. To alleviate data sparsity, previous studies have adopted multitasking approaches to jointly learn dialogue discourse parsing with related tasks (e.g., reading comprehension) that require additional human annotation, thus limiting their generality. In this paper, we propose a multitasking framework that integrates dialogue discourse parsing with its neighboring task addressee recognition. Addressee recognition reveals the reply-to structure that partially overlaps with the relation-based structure, which can be exploited to facilitate relation-based structure learning. To this end, we first proposed a reinforcement learning agent to identify training examples from addressee recognition that are most helpful for dialog discourse parsing. Then, a task-aware structure transformer is designed to capture the shared and private dialogue structure of different tasks, thereby further promoting dialogue discourse parsing. Experimental results on both the Molweni and STAC datasets show that our proposed method can outperform the SOTA baselines. The code will be available at https://github.com/yxfanSuda/RLTST.

14:45-15:00 (East Ballroom)
**QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing**
*Yating Wu, Ritika Rajesh Mangla, Greg Durrett and Junyi Jessy Li*
Questions Under Discussion (QUD) is a versatile linguistic framework in which discourse progresses as continuously asking questions and answering them. Automatic parsing of a discourse to produce a QUD structure thus entails a complex question generation task: given a document and an answer sentence, generate a question that satisfies linguistic constraints of QUD and can be grounded in an anchor sentence in prior context. These questions are known to be curiosity-driven and open-ended. This work introduces the first framework for the automatic evaluation of QUD parsing, instantiating the theoretical constraints of QUD in a concrete protocol. We present QUDeval, a dataset of fine-grained evaluation of 2,190 QUD questions generated from both fine-tuned systems and LLMs. Using QUDeval, we show that satisfying all constraints of QUD is still challenging for modern LLMs, and that existing evaluation metrics poorly approximate parser quality. Encouragingly, human-authored QUDs are scored highly by our human evaluators, suggesting that there is headroom for further progress on language modeling to improve both QUD parsing and QUD evaluation.

15:00-15:15 (East Ballroom)
**Seq2seq is All You Need for Coreference Resolution**
*Wenzheng Zhang, Sam Wiseman and Karl Stratos*
Existing works on coreference resolution suggest that task-specific models are necessary to achieve state-of-the-art performance. In this work, we present compelling evidence that such models are not necessary. We finetune a pretrained seq2seq transformer to map an input document to a tagged sequence encoding the coreference annotation. Despite the extreme simplicity, our model outperforms or closely matches the best coreference systems in the literature on an array of datasets. We consider an even simpler version of seq2seq that generates only the tagged spans and find it highly performant. Our analysis shows that the model size, the amount of supervision, and the choice of sequence representations are key factors in performance.

15:15-15:30 (East Ballroom)
**Prompt-based Logical Semantics Enhancement for Implicit Discourse Relation Recognition**
*Chenxu Wang, Ping Jian and Mu Huang*
Implicit Discourse Relation Recognition (IDRR), which infers discourse relations without the help of explicit connectives, is still a crucial and challenging task for discourse parsing. Recent works tend to exploit the hierarchical structure information from the annotated senses, which demonstrate enhanced discourse relation representations can be obtained by integrating sense hierarchy. Nevertheless, the performance

and robustness for IDRR are significantly constrained by the availability of annotated data. Fortunately, there is a wealth of unannotated utterances with explicit connectives, that can be utilized to acquire enriched discourse relation features. In light of such motivation, we propose a **P**rompt-based **L**ogical **S**emantics **E**nhancement (PLSE) method for IDRR. Essentially, our method seamlessly injects knowledge relevant to discourse relation into pre-trained language models through prompt-based connective prediction. Furthermore, considering the prompt-based connective prediction exhibits local dependencies due to the deficiency of masked language model (MLM) in capturing global semantics, we design a novel self-supervised learning objective based on mutual information maximization to derive enhanced representations of logical semantics for IDRR. Experimental results on PDTB 2.0 and CoNLL16 datasets demonstrate that our method achieves outstanding and consistent performance against the current state-of-the-art models.

## Commonsense Reasoning

14:00-15:30 (Central 1 Ballroom)

---

14:00-14:15 (Central 1 Ballroom)
**Vera: A General-Purpose Plausibility Estimation Model for Commonsense Statements**
*Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi and Hannaneh Hajishirzi*
Today's language models can be remarkably intelligent yet still produce text that contains trivial commonsense errors. Therefore, we seek a retrospective verification approach that can reflect on the commonsense plausibility of the machine text, and introduce Vera, a general-purpose model that learns to estimate the commonsense plausibility of declarative statements. To support diverse commonsense domains, Vera is trained on ∼7M commonsense statements that are automatically converted from 19 QA datasets and two commonsense knowledge bases, and using a combination of three training objectives. When applied to solving commonsense problems in the verification format, Vera substantially outperforms existing models that can be repurposed for commonsense verification, even including GPT-3.5/ChatGPT/GPT-4, and it further exhibits generalization capabilities to unseen tasks and provides well-calibrated outputs. We find that Vera excels at filtering machine-generated commonsense knowledge and is useful in detecting erroneous commonsense statements generated by models like ChatGPT in real-world settings.

14:15-14:30 (Central 1 Ballroom)
**Crystal: Introspective Reasoners Reinforced with Self-Feedback**
*Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi and Asli Celikyilmaz*
Extensive work has shown that the performance and interpretability of commonsense reasoning can be improved via knowledge-augmented reasoning methods, where the knowledge that underpins the reasoning process is explicitly verbalized and utilized. However, existing implementations, including "chain-of-thought" and its variants, fall short in capturing the *introspective* nature of knowledge required in commonsense reasoning, and in accounting for the mutual adaptation between the generation and utilization of knowledge. We propose a novel method to develop an introspective commonsense reasoner, **Crystal**. To tackle commonsense problems, it first introspects for knowledge statements related to the given question, and subsequently makes an informed prediction that is grounded in the previously introspected knowledge. The knowledge introspection and knowledge-grounded reasoning modes of the model are tuned via reinforcement learning to mutually adapt, where the reward derives from the feedback given by the model itself. Experiments show that Crystal significantly outperforms both the standard supervised finetuning and chain-of-thought distilled methods, and enhances the transparency of the commonsense reasoning process. Our work ultimately validates the feasibility and potential of reinforcing a neural model with self-feedback.

14:30-14:45 (Central 1 Ballroom)
**CRoW: Benchmarking Commonsense Reasoning in Real-World Tasks**
*Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva and Antoine Bosselut*
Recent efforts in natural language processing (NLP) commonsense reasoning research have yielded a considerable number of new datasets and benchmarks. However, most of these datasets formulate commonsense reasoning challenges in artificial scenarios that are not reflective of the tasks which real-world NLP systems are designed to solve. In this work, we present CRoW, a manually-curated, multi-task benchmark that evaluates the ability of models to apply commonsense reasoning in the context of six real-world NLP tasks. CRoW is constructed using a multi-stage data collection pipeline that rewrites examples from existing datasets using commonsense-violating perturbations. We use CRoW to study how NLP systems perform across different dimensions of commonsense knowledge, such as physical, temporal, and social reasoning. We find a significant performance gap when NLP systems are evaluated on CRoW compared to humans, showcasing that commonsense reasoning is far from being solved in real-world task settings. We make our dataset and leaderboard available to the research community.

14:45-15:00 (Central 1 Ballroom)
**DialCoT Meets PPO: Decomposing and Exploring Reasoning Paths in Smaller Language Models**
*Chengcheng Han, Xiaowei Du, Che Zhang, Yixin Lian, Xiang Li, Ming Gao and Baoyuan Wang*
Chain-of-Thought (CoT) prompting has successfully enhanced the reasoning capabilities of Large Language Models (LLMs) with at least 100 billion parameters. However, it is ineffective, or even detrimental, to the performance on reasoning tasks in Smaller Language Models (SLMs) with less than 10 billion parameters. In this paper, we propose Dialogue-guided Chain-of-Thought (DialCoT) to improve the reasoning capabilities of SLMs, with the aim of generating intermediate reasoning steps in a dialogue format to guide the model to the final answer. Furthermore, we optimize the model to choose the optimal reasoning path through the Proximal Policy Optimization (PPO) algorithm, further enhancing its reasoning capabilities. Compared to previous methods, our advantages lie in: 1) We transform the process of solving complex reasoning problems into decomposing problems and solving a series of simpler sub-questions, significantly reducing task difficulty and making it more suitable for SLMs. 2) We optimize the model to choose the optimal reasoning path through the PPO algorithm. Comprehensive experiments on four arithmetic reasoning datasets show that our method can achieve significant performance gains over state-of-the-art competitors.

15:00-15:15 (Central 1 Ballroom)
**GD-COMET: A Geo-Diverse Commonsense Inference Model**
*Mehar Bhatia and Vered Shwartz*
With the increasing integration of AI into everyday life, it's becoming crucial to design AI systems to serve users from diverse backgrounds by making them culturally aware. In this paper, we present GD-COMET, a geo-diverse version of the COMET commonsense inference model. GD-COMET goes beyond Western commonsense knowledge and is capable of generating inferences pertaining to a broad range of cultures. We demonstrate the effectiveness of GD-COMET through a comprehensive human evaluation across 5 diverse cultures, as well as extrinsic evaluation on a geo-diverse task. The evaluation shows that GD-COMET captures and generates culturally nuanced commonsense knowledge, demonstrating its potential to benefit NLP applications across the board and contribute to making NLP more inclusive.

15:15-15:30 (Central 1 Ballroom)
**Large Language Models Can Self-Improve**

*Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu and Jiawei Han*
Large Language Models (LLMs) have achieved excellent performances in various tasks. However, fine-tuning an LLM requires extensive supervision. Human, on the other hand, may improve their reasoning abilities by self-thinking without external inputs. In this work, we demonstrate that an LLM is also capable of self-improving with only unlabeled datasets. We use a pre-trained LLM to generate "high-confidence" rationale-augmented answers for unlabeled questions using Chain-of-Though (CoT) prompting and self-consistency, and fine-tune the LLM using those self-generated solutions as target outputs. We show that without any ground truth label, our approach improves the general reasoning ability of a 540B-parameter LLM (74.4%→82.1% on GSM8K, 90.0%→94.4% on OpenBookQA, and 63.4%→67.9% on ANLI-A3) and can also be adapted to extreme low-resource cases where even training questions and CoT prompts are limited. We conduct ablation studies and show that fine-tuning on diverse reasoning paths is critical for self-improvement.

## Efficient Methods for NLP 1

14:00-15:30 (Central 3 Ballroom)

---

14:00-14:15 (Central 3 Ballroom)
**Learning to Predict Task Transferability via Soft Prompt**
*Lingyun Feng*
Fine-tuning pretrained language models on helpful intermediate tasks often greatly improves the performance of target tasks. However, how to efficiently find the source tasks that can successfully transfer still remains under-explored. In this work, we propose to learn an affinity scoring function to predict transferability between tasks. Specifically, we conduct prompt tuning and regard soft prompts as task embeddings that summarize task-specific information. Then we randomly sample task pairs to train an affinity scoring function. The goal is to predict the transfer gain (i.e., affinity) between a task pair, by conditioning on their task embeddings. Once the scoring function is trained, given a novel target task, we use it to predict the most transferable source tasks, without a brute-force search for all possible source-target pairs. Experimental results across 50 tasks show that our method efficiently identifies beneficial tasks for transfer learning.

14:15-14:30 (Central 3 Ballroom)
**Byte Pair Encoding for Symbolic Music**
*Nathan Fradet, Nicolas Gutowski, Fabien Chhel and Jean-Pierre Briot*
When used with deep learning, the symbolic music modality is often coupled with language model architectures. To do so, the music needs to be tokenized, i.e. converted into a sequence of discrete tokens. This can be achieved by different approaches, as music can be composed of simultaneous tracks, of simultaneous notes with several attributes. Until now, the proposed tokenizations rely on small vocabularies of tokens describing the note attributes and time events, resulting in fairly long token sequences, and a sub-optimal use of the embedding space of language models. Recent research has put efforts on reducing the overall sequence length by merging embeddings or combining tokens. In this paper, we show that Byte Pair Encoding, a compression technique widely used for natural language, significantly decreases the sequence length while increasing the vocabulary size. By doing so, we leverage the embedding capabilities of such models with more expressive tokens, resulting in both better results and faster inference in generation and classification tasks. The [source code is shared on Github](https://github.com/Natooz/bpe-symbolic-music), along with a [companion website](https://Natooz.github.io/BPE-Symbolic-Music). Finally, BPE is directly implemented in [MidiTok](https://github.com/Natooz/MidiTok), allowing the reader to easily benefit from this method.

14:30-14:45 (Central 3 Ballroom)
**Understanding the Effect of Model Compression on Social Bias in Large Language Models**
*Gustavo Gonçalves and Emma Strubell*
Large Language Models (LLMs) trained with self-supervision on vast corpora of web text fit to the social biases of that text. Without intervention, these social biases persist in the model's predictions in downstream tasks, leading to representational harm. Many strategies have been proposed to mitigate the effects of inappropriate social biases learned during pretraining. Simultaneously, methods for model compression have become increasingly popular to reduce the computational burden of LLMs. Despite the popularity and need for both approaches, little work has been done to explore the interplay between these two. We perform a carefully controlled study of the impact of model compression via quantization and knowledge distillation on measures of social bias in LLMs. Longer pretraining and larger models led to higher social bias, and quantization showed a regularizer effect with its best trade-off around 20% of the original pretraining time.

14:45-15:00 (Central 3 Ballroom)
**Knowledge Distillation ≈ Label Smoothing: Fact or Fallacy?**
*Md Arafat Sultan*
Originally proposed as a method for knowledge transfer from one model to another, some recent studies have suggested that knowledge distillation (KD) is in fact a form of regularization. Perhaps the strongest argument of all for this new perspective comes from its apparent similarities with label smoothing (LS). Here we re-examine this stated equivalence between the two methods by comparing the predictive confidences of the models they train. Experiments on four text classification tasks involving models of different sizes show that: (a) In most settings, KD and LS drive model confidence in completely opposite directions, and (b) In KD, the student inherits not only its knowledge but also its confidence from the teacher, reinforcing the classical knowledge transfer view.

15:00-15:15 (Central 3 Ballroom)
**The Framework Tax: Disparities Between Inference Efficiency in NLP Research and Deployment**
*Jared Fernandez, Jacob Kahn, Clara Na, Yonatan Bisk and Emma Strubell*
Increased focus on the computational efficiency of systems in natural language processing has motivated the design of efficient model architectures and improvements to underlying hardware accelerators. However, the resulting increases in computational throughput and reductions in floating point operations have not directly translated to improvements in wall-clock inference latency. We demonstrate that these discrepancies can be largely attributed to bottlenecks introduced by deep learning frameworks. We denote this phenomena as the framework tax, and observe that the disparity is growing as hardware speed increases over time. In this work, we examine this phenomena through a series of case studies analyzing the effects of model design decisions, framework paradigms, and hardware platforms on total model latency. Based on our findings, we provide actionable recommendations to researchers and practitioners aimed at narrowing the gap between efficient NLP model research and practice.

15:15-15:30 (Central 3 Ballroom)
**Making Large Language Models Better Data Creators**
*Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W White and Sujay Kumar Jauhar*
Although large language models (LLMs) have advanced the state-of-the-art in NLP significantly, deploying them for downstream applica-

tions is still challenging due to cost, responsiveness, control, or concerns around privacy and security. As such, trainable models are still the preferred option in some cases. However, these models still require human-labeled data for optimal performance, which is expensive and time-consuming to obtain. In order to address this issue, several techniques to reduce human effort involve labeling or generating data using LLMs. Although these methods are effective for certain applications, in practice they encounter difficulties in real-world scenarios. Labeling data requires careful data selection, while generating data necessitates task-specific prompt engineering. In this paper, we propose a unified data creation pipeline that requires only a single formatting example, and which is applicable to a broad range of tasks, including traditionally problematic ones with semantically devoid label spaces. In our experiments we demonstrate that instruction-following LLMs are highly cost-effective data creators, and that models trained with these data exhibit performance better than those trained with human-labeled data (by up to 17.5%) on out-of-distribution evaluation, while maintaining comparable performance on in-distribution tasks. These results have important implications for the robustness of NLP systems deployed in the real-world.

## Ethics in NLP

14:00-15:30 (West 1 Ballroom)

---

14:00-14:15 (West 1 Ballroom)
**TrojanSQL: SQL Injection against Natural Language Interface to Database**
*Jinchuan Zhang, Yan Zhou, Binyuan Hui, Yaxin Liu, Ziming Li and Songlin Hu*
The technology of text-to-SQL has significantly enhanced the efficiency of accessing and manipulating databases. However, limited research has been conducted to study its vulnerabilities emerging from malicious user interaction. By proposing TrojanSQL, a backdoor-based SQL injection framework for text-to-SQL systems, we show how state-of-the-art text-to-SQL parsers can be easily misled to produce harmful SQL statements that can invalidate user queries or compromise sensitive information about the database. The study explores two specific injection attacks, namely *boolean-based injection* and *union-based injection*, which use different types of triggers to achieve distinct goals in compromising the parser. Experimental results demonstrate that both medium-sized models based on fine-tuning and LLM-based parsers using prompting techniques are vulnerable to this type of attack, with attack success rates as high as 99% and 89%, respectively. We hope that this study will raise more concerns about the potential security risks of building natural language interfaces to databases.

14:15-14:30 (West 1 Ballroom)
**ToViLaG: Your Visual-Language Generative Model is Also An Evildoer**
*Xinpeng Wang, Xiaoyuan Yi, Han Jiang, Shanlin Zhou, Zhihua Wei and Xing Xie*
Recent large-scale Visual-Language Generative Models (VLGMs) have achieved unprecedented improvement in multimodal image/text generation. However, these models might also generate toxic content, e.g., offensive text and pornography images, raising significant ethical risks. Despite exhaustive studies on toxic degeneration of language models, this problem remains largely unexplored within the context of visual-language generation. This work delves into the propensity for toxicity generation and susceptibility to toxic data across various VLGMs. For this purpose, we built ToViLaG, a dataset comprising 32K co-toxic/mono-toxic text-image pairs and 1K innocuous but evocative text that tends to stimulate toxicity. Furthermore, we propose WInToRe, a novel toxicity metric tailored to visual-language generation, which theoretically reflects different aspects of toxicity considering both input and output. On such a basis, we benchmarked the toxicity of a diverse spectrum of VLGMs and discovered that some models do more evil than expected while some are more vulnerable to infection, underscoring the necessity of VLGMs detoxification. Therefore, we develop an innovative bottleneck-based detoxification method. Our method could reduce toxicity while maintaining comparable generation quality, providing a promising initial solution to this line of research.

14:30-14:45 (West 1 Ballroom)
**ROBBIE: Robust Bias Evaluation of Large Generative Language Models**
*David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams and Eric Michael Smith*
As generative large language models (LLMs) grow more performant and prevalent, we must develop comprehensive enough tools to measure and improve their fairness. Different prompt-based datasets can be used to measure social bias across multiple text domains and demographic axes, meaning that testing LLMs on more datasets can potentially help us characterize their biases more fully, and better ensure equal and equitable treatment of marginalized demographic groups. In this work, our focus is two-fold: (1) Benchmarking: a comparison of 6 different prompt-based bias and toxicity metrics across 12 demographic axes and 5 families of generative LLMs. Out of those 6 metrics, AdvPromptSet and HolisticBiasR are novel datasets proposed in the paper. The comparison of those benchmarks gives us insights about the bias and toxicity of the compared models. Therefore, we explore the frequency of demographic terms in common LLM pre-training corpora and how this may relate to model biases. (2) Mitigation: we conduct a comprehensive study of how well 3 bias/toxicity mitigation techniques perform across our suite of measurements. ROBBIE aims to provide insights for practitioners while deploying a model, emphasizing the need to not only measure potential harms, but also understand how they arise by characterizing the data, mitigate harms once found, and balance any trade-offs. We open-source our analysis code in hopes of encouraging broader measurements of bias in future LLMs.

14:45-15:00 (West 1 Ballroom)
**We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields**
*Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp and Saif M. Mohammad*
Natural Language Processing (NLP) is poised to substantially influence the world. However, significant progress comes hand-in-hand with substantial risks. Addressing them requires broad engagement with various fields of study. Yet, little empirical work examines the state of such engagement (past or current). In this paper, we quantify the degree of influence between 23 fields of study and NLP (on each other). We analyzed ∼77k NLP papers, ∼3.1m citations from NLP papers to other papers, and ∼1.8m citations from other papers to NLP papers. We show that, unlike most fields, the cross-field engagement of NLP, measured by our proposed Citation Field Diversity Index (CFDI), has declined from 0.58 in 1980 to 0.31 in 2022 (an all-time low). In addition, we find that NLP has grown more insular—citing increasingly more NLP papers and having fewer papers that act as bridges between fields. NLP citations are dominated by computer science; Less than 8% of NLP citations are to linguistics, and less than 3% are to math and psychology. These findings underscore NLP's urgent need to reflect on its engagement with various fields.

15:00-15:15 (West 1 Ballroom)
**Deciphering Stereotypes in Pre-Trained Language Models**
*Weicheng Ma, Henry Scheible, Brian C Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang and Soroush Vosoughi*
Warning: This paper contains content that is stereotypical and may be upsetting. This paper addresses the issue of demographic stereotypes present in Transformer-based pre-trained language models (PLMs) and aims to deepen our understanding of how these biases are encoded in these models. To accomplish this, we introduce an easy-to-use framework for examining the stereotype-encoding behavior of PLMs through

a combination of model probing and textual analyses. Our findings reveal that a small subset of attention heads within PLMs are primarily responsible for encoding stereotypes and that stereotypes toward specific minority groups can be identified using attention maps on these attention heads. Leveraging these insights, we propose an attention-head pruning method as a viable approach for debiasing PLMs, without compromising their language modeling capabilities or adversely affecting their performance on downstream tasks.

15:15-15:30 (West 1 Ballroom)
**Copyright Violations and Large Language Models**
*Antonia Karamolegkou, Jiaang Li, Li Zhou and Anders Søgaard*
Language models may memorize more than just facts, including entire chunks of texts seen during training. Fair use exemptions to copyright laws typically allow for limited use of copyrighted material without permission from the copyright holder, but typically for extraction of information from copyrighted materials, rather than *verbatim* reproduction. This work explores the issue of copyright violations and large language models through the lens of verbatim memorization, focusing on possible redistribution of copyrighted text. We present experiments with a range of language models over a collection of popular books and coding problems, providing a conservative characterization of the extent to which language models can redistribute these materials. Overall, this research highlights the need for further examination and the potential impact on future developments in natural language processing to ensure adherence to copyright regulations. Code is at https://github.com/coastalcph/CopyrightLLMs.

## Phonology, Morphology, and Word Segmentation

14:00-15:30 (West 2 Ballroom)

---

14:00-14:15 (West 2 Ballroom)
**Cognate Transformer for Automated Phonological Reconstruction and Cognate Reflex Prediction**
*V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya*
Phonological reconstruction is one of the central problems in historical linguistics where a proto-word of an ancestral language is determined from the observed cognate words of daughter languages. Computational approaches to historical linguistics attempt to automate the task by learning models on available linguistic data. Several ideas and techniques drawn from computational biology have been successfully applied in this area of computational historical linguistics. Following these lines, we adapt MSA Transformer, a protein language model, to the problem of automated phonological reconstruction. MSA Transformer trains on multiple sequence alignments as input and is, thus, apt for application on aligned cognate words. We, hence, name our model as Cognate Transformer. We also apply the model on another associated task, namely, cognate reflex prediction where a reflex word in a daughter language is predicted based on cognate words from other daughter languages. We show that our model outperforms the existing models on both the tasks, especially when it is pre-trained on masked word prediction task.

14:15-14:30 (West 2 Ballroom)
**Understanding Compositional Data Augmentation in Typologically Diverse Morphological Inflection**
*Farhan Samir and Miikka Silfverberg*
Data augmentation techniques are widely used in low-resource automatic morphological inflection to address the issue of data sparsity. However, the full implications of these techniques remain poorly understood. In this study, we aim to shed light on the theoretical aspects of the data augmentation strategy StemCorrupt, a method that generates synthetic examples by randomly substituting stem characters in existing gold standard training examples. Our analysis uncovers that StemCorrupt brings about fundamental changes in the underlying data distribution, revealing inherent compositional concatenative structure. To complement our theoretical analysis, we investigate the data-efficiency of StemCorrupt. Through evaluation across a diverse set of seven typologically distinct languages, we demonstrate that selecting a subset of datapoints with both high diversity *and* high predictive uncertainty significantly enhances the data-efficiency of compared to competitive baselines. Furthermore, we explore the impact of typological features on the choice of augmentation strategy and find that languages incorporating non-concatenativity, such as morphonological alternations, derive less benefit from synthetic examples with high predictive uncertainty. We attribute this effect to phonotactic violations induced by StemCorrupt, emphasizing the need for further research to ensure optimal performance across the entire spectrum of natural language morphology.

14:30-14:45 (West 2 Ballroom)
**TopWORDS-Poetry: Simultaneous Text Segmentation and Word Discovery for Classical Chinese Poetry via Bayesian Inference**
*Changzai Pan, Feiyue Li and Ke Deng*
As a precious cultural heritage of human beings, classical Chinese poetry has a very unique writing style and often contains special words that rarely appear in general Chinese texts, posting critical challenges for natural language processing. Little effort has been made in the literature for processing texts from classical Chinese poetry. This study fills in this gap with TopWORDS-Poetry, an unsupervised method that can achieve reliable text segmentation and word discovery for classical Chinese poetry simultaneously without pre-given vocabulary or training corpus. Experimental studies confirm that TopWORDS-Poetry can successfully recognize unique poetry words, such as named entities and literary allusions, from metrical poems of *Complete Tang Poetry* and segment these poetry lines into sequences of meaningful words with high quality.

14:45-15:00 (West 2 Ballroom)
**Improved Unsupervised Chinese Word Segmentation Using Pre-trained Knowledge and Pseudo-labeling Transfer**
*Hsiu-Wen Li, Ying-Jia Lin, Yi-Ting Li, Chun Yi Lin and Hung-Yu Kao*
Unsupervised Chinese word segmentation (UCWS) has made progress by incorporating linguistic knowledge from pre-trained language models using parameter-free probing techniques. However, such approaches suffer from increased training time due to the need for multiple inferences using a pre-trained language model to perform word segmentation. This work introduces a novel way to enhance UCWS performance while maintaining training efficiency. Our proposed method integrates the segmentation signal from the unsupervised segmental language model to the pre-trained BERT classifier under a pseudo-labeling framework. Experimental results demonstrate that our approach achieves state-of-the-art performance on the eight UCWS tasks while considerably reducing the training time compared to previous approaches.

15:00-15:15 (West 2 Ballroom)
**Exploring Linguistic Probes for Morphological Inflection**
*Jordan Kodner, Salam Khalifa and Sarah Ruth Brogden Payne*
Modern work on the cross-linguistic computational modeling of morphological inflection has typically employed language-independent data splitting algorithms. In this paper, we supplement that approach with language-specific probes designed to test aspects of morphological generalization. Testing these probes on three morphologically distinct languages, English, Spanish, and Swahili, we find evidence that three leading morphological inflection systems employ distinct generalization strategies over conjugational classes and feature sets on both ortho-

graphic and phonologically transcribed inputs.

15:15-15:30 (West 2 Ballroom)
**On the Role of Morphological Information for Contextual Lemmatization**
*Olia Toporkov and Rodrigo Agerri*
Lemmatization is a natural language processing (NLP) task which consists of producing, from a given inflected word, its canonical form or lemma. Lemmatization is one of the basic tasks that facilitate downstream NLP applications, and is of particular importance for high-inflected languages. Given that the process to obtain a lemma from an inflected word can be explained by looking at its morphosyntactic category, including fine-grained morphosyntactic information to train contextual lemmatizers has become common practice, without considering whether that is the optimum in terms of downstream performance. In order to address this issue, in this paper we empirically investigate the role of morphological information to develop contextual lemmatizers in six languages within a varied spectrum of morphological complexity: Basque, Turkish, Russian, Czech, Spanish and English. Furthermore, and unlike the vast majority of previous work, we also evaluate lemmatizers in out-of-domain settings, which constitutes, after all, their most common application use. The results of our study are rather surprising. It turns out that providing lemmatizers with fine-grained morphological features during training is not that beneficial, not even for agglutinative languages. In fact, modern contextual word representations seem to implicitly encode enough morphological information to obtain competitive contextual lemmatizers without seeing any explicit morphological signal. Moreover, our experiments suggest that the best lemmatizers out-of-domain are those using simple UPOS tags or those trained without morphology and, finally, that current evaluation practices for lemmatization are not adequate to clearly discriminate between models.

## Information Extraction 2

14:00-15:30 (West 3 Ballroom)

---

14:00-14:15 (West 3 Ballroom)
**ViStruct: Visual Structural Knowledge Extraction via Curriculum Guided Code-Vision Representation**
*Yangyi Chen, Xingyao Wang, Manling Li, Derek Hoiem and Heng Ji*
State-of-the-art vision-language models (VLMs) still have limited performance in structural knowledge extraction, such as relations between objects. In this work, we present ViStruct, a training framework to learn VLMs for effective visual structural knowledge extraction. Two novel designs are incorporated. First, we propose to leverage the inherent structure of programming language to depict visual structural information. This approach enables explicit and consistent representation of visual structural information of multiple granularities, such as concepts, relations, and events, in a well-organized structured format. Second, we introduce curriculum-based learning for VLMs to progressively comprehend visual structures, from fundamental visual concepts to intricate event structures. Our intuition is that lower-level knowledge may contribute to complex visual structure understanding. Furthermore, we compile and release a collection of datasets tailored for visual structural knowledge extraction. We adopt a weakly-supervised approach to directly generate visual event structures from captions for ViStruct training, capitalizing on abundant image-caption pairs from the web. In experiments, we evaluate ViStruct on visual structure prediction tasks, demonstrating its effectiveness in improving the understanding of visual structures. The code will be made public to facilitate future research.

14:15-14:30 (West 3 Ballroom)
**CorefPrompt: Prompt-based Event Coreference Resolution by Measuring Event Type and Argument Compatibilities**
*Sheng Xu, Peifeng Li and Qiaoming Zhu*
Event coreference resolution (ECR) aims to group event mentions referring to the same real-world event into clusters. Most previous studies adopt the "encoding first, then scoring" framework, making the coreference judgment rely on event encoding. Furthermore, current methods struggle to leverage human-summarized ECR rules, e.g., coreferential events should have the same event type, to guide the model. To address these two issues, we propose a prompt-based approach, CorefPrompt, to transform ECR into a cloze-style MLM (masked language model) task. This allows for simultaneous event modeling and coreference discrimination within a single template, with a fully shared context. In addition, we introduce two auxiliary prompt tasks, event-type compatibility and argument compatibility, to explicitly demonstrate the reasoning process of ECR, which helps the model make final predictions. Experimental results show that our method CorefPrompt performs well in a state-of-the-art (SOTA) benchmark.

14:30-14:45 (West 3 Ballroom)
**Continual Event Extraction with Semantic Confusion Rectification**
*Zitao Wang, Xinyi Wang and Wei Hu*
We study continual event extraction, which aims to extract incessantly emerging event information while avoiding forgetting. We observe that the semantic confusion on event types stems from the annotations of the same text being updated over time. The imbalance between event types even aggravates this issue. This paper proposes a novel continual event extraction model with semantic confusion rectification. We mark pseudo labels for each sentence to alleviate semantic confusion. We transfer pivotal knowledge between current and previous models to enhance the understanding of event types. Moreover, we encourage the model to focus on the semantics of long-tailed event types by leveraging other associated types. Experimental results show that our model outperforms state-of-the-art baselines and is proficient in imbalanced datasets.

14:45-15:00 (West 3 Ballroom)
**CORE: A Few-Shot Company Relation Classification Dataset for Robust Domain Adaptation.**
*Philipp Borchert, Jochen De Weerdt, Kristof Coussement, Arno De Caigny and Marie-Francine Moens*
We introduce CORE, a dataset for few-shot relation classification (RC) focused on company relations and business entities. CORE includes 4,708 instances of 12 relation types with corresponding textual evidence extracted from company Wikipedia pages. Company names and business entities pose a challenge for few-shot RC models due to the rich and diverse information associated with them. For example, a company name may represent the legal entity, products, people, or business divisions depending on the context. Therefore, deriving the relation type between entities is highly dependent on textual context. To evaluate the performance of state-of-the-art RC models on the CORE dataset, we conduct experiments in the few-shot domain adaptation setting. Our results reveal substantial performance gaps, confirming that models trained on different domains struggle to adapt to CORE. Interestingly, we find that models trained on CORE showcase improved out-of-domain performance, which highlights the importance of high-quality data for robust domain generalization. Specifically, the information richness embedded in business entities allows models to focus on contextual nuances, reducing their reliance on superficial clues such as relation-specific verbs. In addition to the dataset, we provide relevant code snippets to facilitate reproducibility and encourage further research in the field. The CORE dataset and code are publicly available at `https://anonymous.4open.science/r/CORE-D377`.

15:00-15:15 (West 3 Ballroom)
**SpEL: Structured Prediction for Entity Linking**

*Hassan Shavarani and Anoop Sarkar*
Entity linking is a prominent thread of research focused on structured data creation by linking spans of text to an ontology or knowledge source. We revisit the use of structured prediction for entity linking which classifies each individual input token as an entity, and aggregates the token predictions. Our system, called SpEL (Structured prediction for Entity Linking) is a state-of-the-art entity linking system that uses some new ideas to apply structured prediction to the task of entity linking including: two refined fine-tuning steps; a context sensitive prediction aggregation strategy; reduction of the size of the model's output vocabulary, and; we address a common problem in entity-linking systems where there is a training vs. inference tokenization mismatch. Our experiments show that we can outperform the state-of-the-art on the commonly used AIDA benchmark dataset for entity linking to Wikipedia. Our method is also very compute efficient in terms of number of parameters and speed of inference.

15:15-15:30 (West 3 Ballroom)
**TIMELINE: Exhaustive Annotation of Temporal Relations Supporting the Automatic Ordering of Events in News Articles**
*Sarah Alsayyahi and Riza Batista-Navarro*
Temporal relation extraction models have thus far been hindered by a number of issues in existing temporal relation-annotated news datasets, including: (1) low inter-annotator agreement due to the lack of specificity of their annotation guidelines in terms of what counts as a temporal relation; (2) the exclusion of long-distance relations within a given document (those spanning across different paragraphs); and (3) the exclusion of events that are not centred on verbs. This paper aims to alleviate these issues by presenting a new annotation scheme that clearly defines the criteria based on which temporal relations should be annotated. Additionally, the scheme includes events even if they are not expressed as verbs (e.g., nominalised events). Furthermore, we propose a method for annotating all temporal relations—including long-distance ones—which automates the process, hence reducing time and manual effort on the part of annotators. The result is a new dataset, the TIMELINE corpus, in which improved inter-annotator agreement was obtained, in comparison with previously reported temporal relation datasets. We report the results of training and evaluating two baseline temporal relation extraction models on the new corpus, and compare them with results obtained on the widely used MATRES corpus.

## Demo session 2

14:00-15:30 (East Foyer / Virtual)

---

14:00-15:30 (East Foyer / Virtual)
**Spacerini: Plug-and-play Search Engines with Pyserini and Hugging Face**
*Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, Xinyu Zhang, Akintunde Oladipo, Jimmy Lin and Martin Potthast*
We present Spacerini, a tool that integrates the Pyserini toolkit for reproducible information retrieval research with Hugging Face to enable the seamless construction and deployment of interactive search engines. Spacerini makes state-of-the-art sparse and dense retrieval models more accessible to non-IR practitioners while minimizing deployment effort. This is useful for NLP researchers who want to better understand and validate their research by performing qualitative analyses of training corpora, for IR researchers who want to demonstrate new retrieval models integrated into the growing Pyserini ecosystem, and for third parties reproducing the work of other researchers. Spacerini is open source and includes utilities for loading, preprocessing, indexing, and deploying search engines locally and remotely. We demonstrate a portfolio of 13 search engines created with Spacerini for different use cases.

14:00-15:30 (East Foyer / Virtual)
**Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning**
*Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vuli, Sebastian Ruder, Iryna Gurevych and Jonas Pfeiffer*
We introduce Adapters, an open-source library that unifies parameter-efficient and modular transfer learning in large language models. By integrating 10 diverse adapter methods into a unified interface, Adapters offers ease of use and flexible configuration. Our library allows researchers and practitioners to leverage adapter modularity through composition blocks, enabling the design of complex adapter setups. We demonstrate the library's efficacy by evaluating its performance against full fine-tuning on various NLP tasks. Adapters provides a powerful tool for addressing the challenges of conventional fine-tuning paradigms and promoting more efficient and modular transfer learning. The library is available via https://adapterhub.ml/adapters.

14:00-15:30 (East Foyer / Virtual)
**Humanoid Agents: Platform for Simulating Human-like Generative Agents**
*Zhilin Wang, Yu Ying Chiu and Yu Cheung Chiu*
Just as computational simulations of atoms, molecules and cells have shaped the way we study the sciences, true-to-life simulations of human-like agents can be valuable tools for studying human behavior. We propose Humanoid Agents, a system that guides Generative Agents to behave more like humans by introducing three elements of System 1 processing: Basic needs (e.g. hunger, health and energy), Emotion and Closeness in Relationships. Humanoid Agents are able to use these dynamic elements to adapt their daily activities and conversations with other agents, as supported with empirical experiments. Our system is designed to be extensible to various settings, three of which we demonstrate, as well as to other elements influencing human behavior (e.g. empathy, moral values and cultural background). Our platform also includes a Unity WebGL game interface for visualization and an interactive analytics dashboard to show agent statuses over time. Our platform is available on https://www.humanoidagents.com/ and code is on https://github.com/HumanoidAgents/HumanoidAgents

14:00-15:30 (East Foyer / Virtual)
**CLEVA: Chinese Language Models EVAluation Platform**
*Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael Lyu and Liwei Wang*
With the continuous emergence of Chinese Large Language Models (LLMs), how to evaluate a model's capabilities has become an increasingly significant issue. The absence of a comprehensive Chinese benchmark that thoroughly assesses a model's performance, the unstandardized and incomparable prompting procedure, and the prevalent risk of contamination pose major challenges in the current evaluation of Chinese LLMs. We present CLEVA, a user-friendly platform crafted to holistically evaluate Chinese LLMs. Our platform employs a standardized workflow to assess LLMs' performance across various dimensions, regularly updating a competitive leaderboard. To alleviate contamination, CLEVA curates a significant proportion of new data and develops a sampling strategy that guarantees a unique subset for each leaderboard round. Empowered by an easy-to-use interface that requires just a few mouse clicks and a model API, users can conduct a thorough evaluation with minimal coding. Large-scale experiments featuring 23 Chinese LLMs have validated CLEVA's efficacy.

14:00-15:30 (East Foyer / Virtual)
**DOPA METER – A Tool Suite for Metrical Document Profiling and Aggregation**

*Christina Lohr and Udo Hahn*
We present DOPA METER, a tool suite for the metrical investigation of written language, that provides diagnostic means for its division into discourse categories, such as registers, genres, and style. The quantitative basis of our system are 120 metrics covering a wide range of lexical, syntactic, and semantic features relevant for language profiling. The scores can be summarized, compared, and aggregated using visualization tools that can be tailored according to the users' needs. We also showcase an application scenario for DOPA METER.

14:00-15:30 (East Foyer / Virtual)
**Muted: Multilingual Targeted Offensive Speech Identification and Visualization**
*Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse, Rong Zhang, Avirup Sil and Bishwaranjan Bhattacharjee*
Offensive language such as hate, abuse, and profanity (HAP) occurs in various content on the web. While previous work has mostly dealt with sentence level annotations, there have been a few recent attempts to identify offensive spans as well. We build upon this work and introduce MUTED, a system to identify multilingual HAP content by displaying offensive arguments and their targets using heat maps to indicate their intensity. MUTED can leverage any transformer-based HAP-classification model and its attention mechanism out-of-the-box to identify toxic spans, without further fine-tuning. In addition, we use the spaCy library to identify the specific targets and arguments for the words predicted by the attention heatmaps. We present the model's performance on identifying offensive spans and their targets in existing datasets and present new annotations on German text. Finally, we demonstrate our proposed visualization tool on multilingual inputs.

## Poster session 2

14:00-15:30 (East Foyer / Virtual)

---

14:00-15:30 (East Foyer / Virtual)
**#1 VivesDebate-Speech: A Corpus of Spoken Argumentation to Leverage Audio Features for Argument Mining**
*Ramon Ruiz-Dolz and Javier Iranzo Sanchez*
In this paper, we describe VivesDebate-Speech, a corpus of spoken argumentation created to leverage audio features for argument mining tasks. The creation of this corpus represents an important contribution to the intersection of speech processing and argument mining communities, and one of the most complete publicly available resources in this topic. Moreover, we have performed a set of first-of-their-kind experiments which show an improvement when integrating audio features into the argument mining pipeline. The provided results can be used as a baseline for future research.

14:00-15:30 (East Foyer / Virtual)
**#2 CQE: A Comprehensive Quantity Extractor**
*Satya Almasian, Vivian Kazakova, Philipp Göldner and Michael Gertz*
Quantities are essential in documents to describe factual information. They are ubiquitous in application domains such as finance, business, medicine, and science in general. Compared to other information extraction approaches, interestingly only a few works exist that describe methods for a proper extraction and representation of quantities in text. In this paper, we present such a comprehensive quantity extraction framework from text data. It efficiently detects combinations of values and units, the behavior of a quantity (e.g., rising or falling), and the concept a quantity is associated with. Our framework makes use of dependency parsing and a dictionary of units, and it provides for a proper normalization and standardization of detected quantities. Using a novel dataset for evaluation, we show that our open source framework outperforms other systems and – to the best of our knowledge – is the first to detect concepts associated with identified quantities. The code and data underlying our framework are available at https://github.com/vivkaz/CQE.

14:00-15:30 (East Foyer / Virtual)
**#3 Automatic Debate Evaluation with Argumentation Semantics and Natural Language Argument Graph Networks**
*Ramon Ruiz-Dolz, Stella Heras and Ana Garcia*
The lack of annotated data on professional argumentation and complete argumentative debates has led to the oversimplification and the inability of approaching more complex natural language processing tasks. Such is the case of the automatic evaluation of complete professional argumentative debates. In this paper, we propose an original hybrid method to automatically predict the winning stance in this kind of debates. For that purpose, we combine concepts from argumentation theory such as argumentation frameworks and semantics, with Transformer-based architectures and neural graph networks. Furthermore, we obtain promising results that lay the basis on an unexplored new instance of the automatic analysis of natural language arguments.

14:00-15:30 (East Foyer / Virtual)
**#4 Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis**
*Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu and Tianshu Yu*
Though Multimodal Sentiment Analysis (MSA) proves effective by utilizing rich information from multiple sources (*e.g.,* language, video, and audio), the potential sentiment-irrelevant and conflicting information across modalities may hinder the performance from being further improved. To alleviate this, we present Adaptive Language-guided Multimodal Transformer (ALMT), which incorporates an Adaptive Hyper-modality Learning (AHL) module to learn an irrelevance/conflict-suppressing representation from visual and audio features under the guidance of language features at different scales. With the obtained hyper-modality representation, the model can obtain a complementary and joint representation through multimodal fusion for effective MSA. In practice, ALMT achieves state-of-the-art performance on several popular datasets (*e.g.,* MOSI, MOSEI and CH-SIMS) and an abundance of ablation demonstrates the validity and necessity of our irrelevance/conflict suppression mechanism.

14:00-15:30 (East Foyer / Virtual)
**#5 How to Enhance Causal Discrimination of Utterances: A Case on Affective Reasoning**
*Hang Chen, Xinyu Yang, Jing Luo and Wenjing Zhu*
Our investigation into the Affective Reasoning in Conversation (ARC) task highlights the challenge of causal discrimination. Almost all existing models, including large language models (LLMs), excel at capturing semantic correlations within utterance embeddings but fall short in determining the specific causal relationships. To overcome this limitation, we propose the incorporation of *i.i.d.* noise terms into the conversation process, thereby constructing a structural causal model (SCM). It explores how distinct causal relationships of fitted embeddings can be discerned through independent conditions. To facilitate the implementation of deep learning, we introduce the cogn frameworks to handle unstructured conversation data, and employ an autoencoder architecture to regard the unobservable noise as learnable "implicit causes." Moreover, we curate a synthetic dataset that includes i.i.d. noise. Through comprehensive experiments, we validate the effectiveness and interpretability of our approach. Our code is available in https://github.com/Zodiark-ch/mater-of-our-EMNLP2023-paper.

14:00-15:30 (East Foyer / Virtual)
**#6 Symbol tuning improves in-context learning in language models**
*Jerry Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma and Quoc V Le*
We present symbol tuning - finetuning language models on in-context input-label pairs where natural language labels (e.g., "positive/negative sentiment") are replaced with arbitrary symbols (e.g., "foo/bar"). Symbol tuning leverages the intuition that when a model cannot use instructions or natural language labels to figure out a task, it must instead do so by learning the input-label mappings. We experiment with symbol tuning across PaLM models up to 540B parameters and observe benefits across various settings. First, symbol tuning boosts performance on unseen in-context learning tasks and is much more robust to underspecified prompts, such as those without instructions or without natural language labels. Second, symbol-tuned models are much stronger at algorithmic reasoning tasks, with up to 18.2% better performance on the List Functions benchmark and up to 15.3% better performance on the Simple Turing Concepts benchmark. Finally, symbol-tuned models show large improvements in following flipped-labels presented in-context, meaning that they are more capable of using in-context information to override prior knowledge.

14:00-15:30 (East Foyer / Virtual)
**#7 Collaborative Generative AI: Integrating GPT-k for Efficient Editing in Text-to-Image Generation**
*Wanrong Zhu, Xinyi Wang, Yujie Lu, Tsu-Jui Fu, Xin Eric Wang, Miguel Eckstein and William Yang Wang*
The field of text-to-image (T2I) generation has garnered significant attention both within the research community and among everyday users. Despite the advancements of T2I models, a common issue encountered by users is the need for repetitive editing of input prompts in order to receive a satisfactory image, which is time-consuming and labor-intensive. Given the demonstrated text generation power of large-scale language models, such as GPT-k, we investigate the potential of utilizing such models to improve the prompt editing process for T2I generation. We conduct a series of experiments to compare the common edits made by humans and GPT-k, evaluate the performance of GPT-k in prompting T2I, and examine factors that may influence this process. We found that GPT-k models focus more on inserting modifiers while humans tend to replace words and phrases, which includes changes to the subject matter. Experimental results show that GPT-k are more effective in adjusting modifiers rather than predicting spontaneous changes in the primary subject matters. Adopting the edit suggested by GPT-k models may reduce the percentage of remaining edits by 20-30%.

14:00-15:30 (East Foyer / Virtual)
**#8 Automatic Prompt Optimization with "Gradient Descent" and Beam Search**
*Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu and Michael Zeng*
Large Language Models (LLMs) have shown impressive performance as general purpose agents, but their abilities remain highly dependent on prompts which are hand written with onerous trial-and-error effort. We propose a simple and nonparametric solution to this problem, Prompt Optimization with Textual Gradients (ProTeGi), which is inspired by numerical gradient descent to automatically improve prompts, assuming access to training data and an LLM API. The algorithm uses minibatches of data to form natural language "gradients" that criticize the current prompt, much like how numerical gradients point in the direction of error ascent. The natural language gradients are then "propagated" into the prompt by editing the prompt in the opposite semantic direction of the gradient. These gradient descent steps are guided by a beam search and bandit selection procedure which significantly improves algorithmic efficiency. Preliminary results across three benchmark NLP tasks and the novel problem of LLM jailbreak detection suggest that Automatic Prompt Optimization can outperform prior prompt editing techniques and improve an initial prompt's performance by up to 31%, by using data to rewrite vague task descriptions into more precise annotation instructions.

14:00-15:30 (East Foyer / Virtual)
**#9 CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations**
*Myra Cheng, Tiziano Piccardi and Diyi Yang*
Recent work has aimed to capture nuances of human behavior by using LLMs to simulate responses from particular demographics in settings like social science experiments and public opinion surveys. However, there are currently no established ways to discuss or evaluate the quality of such LLM simulations. Moreover, there is growing concern that these LLM simulations are flattened caricatures of the personas that they aim to simulate, failing to capture the multidimensionality of people and perpetuating stereotypes. To bridge these gaps, we present CoMPosT, a framework to characterize LLM simulations using four dimensions: Context, Model, Persona, and Topic. We use this framework to measure open-ended LLM simulations' susceptibility to caricature, defined via two criteria: individuation and exaggeration. We evaluate the level of caricature in scenarios from existing work on LLM simulations. We find that for GPT-4, simulations of certain demographics (political and marginalized groups) and topics (general, uncontroversial) are highly susceptible to caricature.

14:00-15:30 (East Foyer / Virtual)
**#10 Self-Ensemble of $N$-best Generation Hypotheses by Lexically Constrained Decoding**
*Ryota Miyano, Tomoyuki Kajiwara and Yuki Arase*
We propose a method that ensembles $N$-best hypotheses to improve natural language generation. Previous studies have achieved notable improvements in generation quality by explicitly reranking $N$-best candidates. These studies assume that there exists a hypothesis of higher quality. We expand the assumption to be more practical as there exist *partly* higher quality hypotheses in the $N$-best yet they may be imperfect as the entire sentences. By merging these high-quality fragments, we can obtain a higher-quality output than the single-best sentence. Specifically, we first obtain $N$-best hypotheses and conduct token-level quality estimation. We then apply tokens that should or should not be present in the final output as lexical constraints in decoding. Empirical experiments on paraphrase generation, summarisation, and constrained text generation confirm that our method outperforms the strong $N$-best reranking methods.

14:00-15:30 (East Foyer / Virtual)
**#11 The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis**
*Pranav Narayanan Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J. Passonneau and Shomir Wilson*
We conduct an inquiry into the sociotechnical aspects of sentiment analysis (SA) by critically examining 189 peer-reviewed papers on their applications, models, and datasets. Our investigation stems from the recognition that SA has become an integral component of diverse sociotechnical systems, exerting influence on both social and technical users. By delving into sociological and technological literature on sentiment, we unveil distinct conceptualizations of this term in domains such as finance, government, and medicine. Our study exposes a lack of explicit definitions and frameworks for characterizing sentiment, resulting in potential challenges and biases. To tackle this issue, we propose an ethics sheet encompassing critical inquiries to guide practitioners in ensuring equitable utilization of SA. Our findings underscore the significance of adopting an interdisciplinary approach to defining sentiment in SA and offer a pragmatic solution for its implementation.

14:00-15:30 (East Foyer / Virtual)
**#12 Characterizing and Verifying Scientific Claims: Qualitative Causal Structure is All You Need**
*Jinxuan Wu, Wenhan Chao, Xian Zhou and Zhunchen Luo*
A scientific claim typically begins with the formulation of a research question or hypothesis, which is a tentative statement or proposition

about a phenomenon or relationship between variables. Within the realm of scientific claim verification, considerable research efforts have been dedicated to attention architectures and leveraging the text comprehension capabilities of Pre-trained Language Models (PLMs), yielding promising performances. However, these models overlook the causal structure information inherent in scientific claims, thereby failing to establish a comprehensive chain of causal inference. This paper delves into the exploration to highlight the crucial role of qualitative causal structure in characterizing and verifying scientific claims based on evidence. We organize the qualitative causal structure into a heterogeneous graph and propose a novel attention-based graph neural network model to facilitate causal reasoning across relevant causally-potent factors. Our experiments demonstrate that by solely utilizing the qualitative causal structure, the proposed model achieves comparable performance to PLM-based models. Furthermore, by incorporating semantic features, our model outperforms state-of-the-art approaches comprehensively.

14:00-15:30 (East Foyer / Virtual)
**#13 Rationale-Enhanced Language Models are Better Continual Relation Learners**
*Weimin Xiong, Yifan Song, Peiyi Wang and Sujian Li*
Continual relation extraction (CRE) aims to solve the problem of catastrophic forgetting when learning a sequence of newly emerging relations. Recent CRE studies have found that catastrophic forgetting arises from the model's lack of robustness against future analogous relations. To address the issue, we introduce rationale, i.e., the explanations of relation classification results generated by Large Language Models (LLM), into CRE task. Specifically, we design the multi-task rationale tuning strategy to help the model learn current relations robustly. We also conduct contrastive rationale replay to further distinguish analogous relations. Experimental results on two standard benchmarks demonstrate that our method outperforms the state-of-the-art CRE models.

14:00-15:30 (East Foyer / Virtual)
**#14 Towards A Unified View of Sparse Feed-Forward Network in Pretraining Large Language Model**
*Zeyu Liu, Tim Dettmers, Xi Victoria Lin, Veselin Stoyanov and Xian Li*
Large and sparse feed-forward layers (S-FFN) such as Mixture-of-Experts (MoE) have proven effective in scaling up Transformers model size for pretraining large language models. By only activating part of the FFN parameters conditioning on input, S-FFN improves generalization performance while keeping training and inference costs (in FLOPs) fixed. In this work, we analyzed two major design choices of S-FFN: the memory block (a.k.a. expert) size and the memory block selection method under a general conceptual framework of sparse neural memory. Using this unified framework, we compare several S-FFN architectures for language modeling and provide insights into their relative efficacy and efficiency. We found a simpler selection method — Avg-K that selects blocks through their mean aggregated hidden states, achieving lower perplexity in language model pretraining compared to existing MoE architectures including Switch Transformer (Fedus et al., 2021) and HashLayer (Roller et al., 2021).

14:00-15:30 (East Foyer / Virtual)
**#15 Event Causality Extraction via Implicit Cause-Effect Interactions**
*Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin and Xiaoyu Li*
Event Causality Extraction (ECE) aims to extract the cause-effect event pairs from the given text, which requires the model to possess a strong reasoning ability to capture event causalities. However, existing works have not adequately exploited the interactions between the cause and effect event that could provide crucial clues for causality reasoning. To this end, we propose an Implicit Cause-Effect interaction (ICE) framework, which formulates ECE as a template-based conditional generation problem. The proposed method captures the implicit intra- and inter-event interactions by incorporating the privileged information (ground truth event types and arguments) for reasoning, and a knowledge distillation mechanism is introduced to alleviate the unavailability of privileged information in the test stage. Furthermore, to facilitate knowledge transfer from teacher to student, we design an event-level alignment strategy named Cause-Effect Optimal Transport (CEOT) to strengthen the semantic interactions of cause-effect event types and arguments. Experimental results indicate that ICE achieves state-of-the-art performance on the ECE-CCKS dataset.

14:00-15:30 (East Foyer / Virtual)
**#16 Self-Improvement of Non-autoregressive Model via Sequence-Level Distillation**
*Yusheng Liao, Shuyang Jiang, Yiqi Li, Yu Wang and Yanfeng Wang*
Although Non-autoregressive Transformer (NAT) models have achieved great success in terms of fast inference speed, this speedup comes with a performance drop due to the inherent *multi-modality* problem of the NAT model. Previous works commonly alleviate this problem by replacing the target side of the raw data with distilled data generated by Autoregressive Transformer (AT) models. However, the multi-modality problem in the distilled data is still significant and thus limits further improvement of the NAT models. In this paper, we propose a method called Sequence-Level Self-Distillation (SLSD), which aims to generate distilled data by the NAT model itself, eliminating the need for additional teacher networks. Furthermore, SLSD can adapt to different NAT models without precise adjustments since the self-distilled data is generated from the same types of NAT models. We conduct extensive experiments on WMT14 EN↔DE and WMT16 EN↔RO and choose four classic NAT models as the backbones to validate the generality and effectiveness of SLSD. The results show that our approach can consistently improve all models on both raw data and distilled data without sacrificing the inference speed.

14:00-15:30 (East Foyer / Virtual)
**#17 Primacy Effect of ChatGPT**
*Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang and Bryan Hooi*
Instruction-tuned large language models (LLMs), such as ChatGPT, have led to promising zero-shot performance in discriminative natural language understanding (NLU) tasks. This involves querying the LLM using a prompt containing the question, and the candidate labels to choose from. The question-answering capabilities of ChatGPT arise from its pre-training on large amounts of human-written text, as well as its subsequent fine-tuning on human preferences, which motivates us to ask: Does ChatGPT also inherit humans' cognitive biases? In this paper, we study the primacy effect of ChatGPT: the tendency of selecting the labels at earlier positions as the answer. We have two main findings: i) ChatGPT's decision is sensitive to the order of labels in the prompt; ii) ChatGPT has a clearly higher chance to select the labels at earlier positions as the answer. We hope that our experiments and analyses provide additional insights into building more reliable ChatGPT-based solutions. We release the source code at https://github.com/wangywUST/PrimacyEffectGPT.

14:00-15:30 (East Foyer / Virtual)
**#18 Biomedical Named Entity Recognition via Dictionary-based Synonym Generalization**
*Zihao Fu, Yixuan Su, Zaiqiao Meng and Nigel Collier*
Biomedical named entity recognition is one of the core tasks in biomedical natural language processing (BioNLP). To tackle this task, numerous supervised/distantly supervised approaches have been proposed. Despite their remarkable success, these approaches inescapably demand laborious human effort. To alleviate the need of human effort, dictionary-based approaches have been proposed to extract named entities simply based on a given dictionary. However, one downside of existing dictionary-based approaches is that they are challenged to identify concept synonyms that are not listed in the given dictionary, which we refer as the synonym generalization problem. In this study, we propose a novel Synonym Generalization (SynGen) framework that recognizes the biomedical concepts contained in the input text using span-based predictions. In particular, SynGen introduces two regularization terms, namely, (1) a synonym distance regularizer; and (2) a noise pertur-

bation regularizer, to minimize the synonym generalization error. To demonstrate the effectiveness of our approach, we provide a theoretical analysis of the bound of synonym generalization error. We extensively evaluate our approach on a wide range of benchmarks and the results verify that SynGen outperforms previous dictionary-based models by notable margins. Lastly, we provide a detailed analysis to further reveal the merits and inner-workings of our approach.

14:00-15:30 (East Foyer / Virtual)
**#19 Target-to-Source Augmentation for Aspect Sentiment Triplet Extraction**
*Yice Zhang, Yifan Yang, Meng Li, Bin Liang, Shiwei Chen and Ruifeng Xu*
Aspect Sentiment Triplet Extraction (ASTE) is an important task in sentiment analysis, aiming to extract aspect-level opinions and sentiments from user-generated reviews. The fine-grained nature of ASTE incurs a high annotation cost, while the scarcity of annotated data limits the performance of existing methods. This paper exploits data augmentation to address this issue. Traditional augmentation methods typically modify the input sentences of existing samples via heuristic rules or language models, which have shown success in text classification tasks. However, applying these methods to fine-grained tasks like ASTE poses challenges in generating diverse augmented samples while maintaining alignment between modified sentences and origin labels. Therefore, this paper proposes a target-to-source augmentation approach for ASTE. Our approach focuses on learning a generator that can directly generate new sentences based on labels and syntactic templates. With this generator, we can generate a substantial number of diverse augmented samples by mixing labels and syntactic templates from different samples. Besides, to ensure the quality of the generated sentence, we introduce fluency and alignment discriminators to provide feedback on the generated sentence and then use this feedback to optimize the generator via a reinforcement learning framework. Experiments demonstrate that our approach significantly enhances the performance of existing ASTE models.

14:00-15:30 (East Foyer / Virtual)
**#20 Self-Detoxifying Language Models via Toxification Reversal**
*Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang and Wenjie Li*
Language model detoxification aims to minimize the risk of generating offensive or harmful content in pretrained language models (PLMs) for safer deployment. Existing methods can be roughly categorized as finetuning-based and decoding-based. However, the former is often resource-intensive, while the latter relies on additional components and potentially compromises the generation fluency. In this paper, we propose a more lightweight approach that enables the PLM itself to achieve "self-detoxification". Our method is built upon the observation that prepending a negative steering prompt can effectively induce PLMs to generate toxic content. At the same time, we are inspired by the recent research in the interpretability field, which formulates the evolving contextualized representations within the PLM as an information stream facilitated by the attention layers. Drawing on this idea, we devise a method to identify the toxification direction from the normal generation process to the one prompted with the negative prefix, and then steer the generation to the reversed direction by manipulating the information movement within the attention layers. Experimental results show that our approach, without any fine-tuning or extra components, can achieve comparable performance with state-of-the-art methods.

14:00-15:30 (East Foyer / Virtual)
**#21 Explanation Selection Using Unlabeled Data for Chain-of-Thought Prompting**
*Xi Ye and Greg Durrett*
Recent work has shown how to prompt large language models with explanations to obtain strong performance on textual reasoning tasks, i.e., the chain-of-thought paradigm. However, subtly different explanations can yield widely varying downstream task accuracy. Explanations that have not been "tuned" for a task, such as off-the-shelf explanations written by non-experts, may lead to mediocre performance. This paper tackles the problem of how to optimize explanation-infused prompts in a blackbox fashion. We first generate sets of candidate explanations for each example in the prompt using a leave-one-out scheme, then find an effective combination of these explanations with a two-stage framework. We first evaluate explanations for each in-context example in isolation according to two proxy metrics, log likelihood and accuracy on new examples. Then, we search over combinations of explanations to find one that yields high performance against a silver-labeled development set. Across four textual reasoning tasks spanning question answering, mathematical reasoning, and natural language inference, results show that our proxy metrics correlate with ground truth accuracy and our overall method can effectively improve prompts over crowdworker annotations and naive search strategies

14:00-15:30 (East Foyer / Virtual)
**#22 Open-world Semi-supervised Generalized Relation Discovery Aligned in a Real-world Setting**
*William P Hogan, Jiacheng Li and Jingbo Shang*
Open-world Relation Extraction (OpenRE) has recently garnered significant attention. However, existing approaches tend to oversimplify the problem by assuming that all instances of unlabeled data belong to novel classes, thereby limiting the practicality of these methods. We argue that the OpenRE setting should be more aligned with the characteristics of real-world data. Specifically, we propose two key improvements: (a) unlabeled data should encompass known and novel classes, including negative instances; and (b) the set of novel classes should represent long-tail relation types. Furthermore, we observe that popular relations can often be implicitly inferred through specific patterns, while long-tail relations tend to be explicitly expressed. Motivated by these insights, we present a method called KNoRD (Known and Novel Relation Discovery), which effectively classifies explicitly and implicitly expressed relations from known and novel classes within unlabeled data. Experimental evaluations on several Open-world RE benchmarks demonstrate that KNoRD consistently outperforms other existing methods, achieving significant performance gains.

14:00-15:30 (East Foyer / Virtual)
**#23 SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization**
*Philippe Laban, Wojciech Maciej Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty and Chien-Sheng Wu*
With the recent appearance of LLMs in practical settings, having methods that can effectively detect factual inconsistencies is crucial to reduce the propagation of misinformation and improve trust in model outputs. When testing on existing factual consistency benchmarks, we find that a few large language models (LLMs) perform competitively on classification benchmarks for factual inconsistency detection compared to traditional non-LLM methods. However, a closer analysis reveals issues with existing evaluation benchmarks, affecting evaluation precision. To address this, we propose a new protocol for inconsistency detection benchmark creation and implement it in a 10-domain benchmark called SummEdits. This new benchmark is 20 times more cost-effective per sample than previous benchmarks and highly reproducible, as we estimate inter-annotator agreement at about 0.9. Most LLMs struggle on SummEdits, with performance close to random chance. The best-performing model, GPT-4, is still 8% below estimated human performance, highlighting the gaps in LLMs' ability to reason about facts and detect inconsistencies when they occur.

14:00-15:30 (East Foyer / Virtual)
**#24 Video-Helpful Multimodal Machine Translation**
*Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi and Wei Li*
Existing multimodal machine translation (MMT) datasets consist of images and video captions or instructional video subtitles, which rarely contain linguistic ambiguity, making visual information ineffective in generating appropriate translations. Recent work has constructed an

ambiguous subtitles dataset to alleviate this problem but is still limited to the problem that videos do not necessarily contribute to disambiguation. We introduce EVA (Extensive training set and Video-helpful evaluation set for Ambiguous subtitles translation), an MMT dataset containing 852k Japanese-English parallel subtitle pairs, 520k Chinese-English parallel subtitle pairs, and corresponding video clips collected from movies and TV episodes. In addition to the extensive training set, EVA contains a video-helpful evaluation set in which subtitles are ambiguous, and videos are guaranteed helpful for disambiguation. Furthermore, we propose SAFA, an MMT model based on the Selective Attention model with two novel methods: Frame attention loss and Ambiguity augmentation, aiming to use videos in EVA for disambiguation fully. Experiments on EVA show that visual information and the proposed methods can boost translation performance, and our model performs significantly better than existing MMT models.

14:00-15:30 (East Foyer / Virtual)
**#25 An Investigation of LLMs' Inefficacy in Understanding Converse Relations**
*Chengwen Qi, Bowen Li, Binyuan Hui, Bailin Wang, Jinyang Li, Jinwang Wu and Yuanjun Laili*
Large Language Models (LLMs) have achieved remarkable success in many formal language oriented tasks, such as structural data-to-text and semantic parsing. However current benchmarks mostly follow the data distribution of the pre-training data of LLMs. Therefore, a natural question rises that do LLMs really understand the structured semantics of formal languages. In this paper, we investigate this problem on a special case, converse binary relation. We introduce a new benchmark ConvRe focusing on converse relations, which contains 17 relations and 1240 triples extracted from popular knowledge graph completion datasets. Our ConvRE features two tasks, Re2Text and Text2Re, which are formulated as multi-choice question answering to evaluate LLMs' ability to determine the matching between relations and associated text. For the evaluation protocol, apart from different prompting methods, we further introduce variants to the test text and few-shot example text. We conduct experiments on three popular LLM families and have observed various scaling trends. The results suggest that LLMs often resort to shortcut learning and still face challenges on our proposed benchmark.

14:00-15:30 (East Foyer / Virtual)
**#26 Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT**
*Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun and Ming Gu*
Large language models (LLMs), e.g., ChatGPT, have revolutionized the domain of natural language processing because of their excellent performance on various tasks. Despite their great potential, LLMs also incur serious concerns as they are likely to be misused. There are already reported cases of academic cheating by using LLMs. Thus, it is a pressing problem to identify LLM-generated texts. In this work, we design a zero-shot black-box method for detecting LLM-generated texts. The key idea is to revise the text to be detected using the ChatGPT model. Our method is based on the intuition that the ChatGPT model will make fewer revisions to LLM-generated texts than it does to human-written texts, because the texts generated by LLMs are more in accord with the generation logic and statistical patterns learned by LLMs like ChatGPT. Thus, if the text to be detected and its ChatGPT-revised version have a higher degree of similarity, the text is more likely to be LLM-generated. Extensive experiments on various datasets and tasks show that our method can effectively detect LLM-generated texts. Moreover, compared with other detection methods, our method has better generalization ability and is more stable across various datasets. The codes are publicly available at https://github.com/thunlp/LLM-generated-text-detection.

14:00-15:30 (East Foyer / Virtual)
**#27 Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation**
*Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian and Yunshi Lan*
The task of Question Generation over Knowledge Bases (KBQG) aims to convert a logical form into a natural language question. For the sake of expensive cost of large-scale question annotation, the methods of KBQG under low-resource scenarios urgently need to be developed. However, current methods heavily rely on annotated data for fine-tuning, which is not well-suited for few-shot question generation. The emergence of Large Language Models (LLMs) has shown their impressive generalization ability in few-shot tasks. Inspired by Chain-of-Thought (CoT) prompting, which is an in-context learning strategy for reasoning, we formulate KBQG task as a reasoning problem, where the generation of a complete question is splitted into a series of sub-question generation. Our proposed prompting method KQG-CoT first retrieves supportive logical forms from the unlabeled data pool taking account of the characteristics of the logical form. Then, we write a prompt to explicit the reasoning chain of generating complicated questions based on the selected demonstrations. To further ensure prompt quality, we extend KQG-CoT into KQG-CoT+ via sorting the logical forms by their complexity. We conduct extensive experiments over three public KBQG datasets. The results demonstrate that our prompting method consistently outperforms other prompting baselines on the evaluated datasets. Remarkably, our KQG-CoT+ method could surpass existing few-shot SoTA results of the PathQuestions dataset by 18.25, 10.72, and 10.18 absolute points on BLEU-4, METEOR, and ROUGE-L, respectively.

14:00-15:30 (East Foyer / Virtual)
**#28 DEPN: Detecting and Editing Privacy Neurons in Pretrained Language Models**
*Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian and Deyi Xiong*
Pretrained language models have learned a vast amount of human knowledge from large-scale corpora, but their powerful memorization capability also brings the risk of data leakage. Some risks may only be discovered after the model training is completed, such as the model memorizing a specific phone number and frequently outputting it. In such cases, model developers need to eliminate specific data influences from the model to mitigate legal and ethical penalties. To effectively mitigate these risks, people often have to spend a significant amount of time and computational costs to retrain new models instead of finding ways to cure the 'sick' models. Therefore, we propose a method to locate and erase risky neurons in order to eliminate the impact of privacy data in the model. We use a new method based on integrated gradients to locate neurons associated with privacy texts, and then erase these neurons by setting their activation values to zero.Furthermore, we propose a risky neuron aggregation method to eliminate the influence of privacy data in the model in batches. Experimental results show that our method can effectively and quickly eliminate the impact of privacy data without affecting the model's performance. Additionally, we demonstrate the relationship between model memorization and neurons through experiments, further illustrating the robustness of our method.

14:00-15:30 (East Foyer / Virtual)
**#29 A Unified View of Evaluation Metrics for Structured Prediction**
*Yunmo Chen, William Gantt, Tongfei Chen, Aaron Steven White and Benjamin Van Durme*
We present a conceptual framework that unifies a variety of evaluation metrics for different structured prediction tasks (e.g. event and relation extraction, syntactic and semantic parsing). Our framework requires representing the outputs of these tasks as objects of certain data types, and derives metrics through matching of common substructures, possibly followed by normalization. We demonstrate how commonly used metrics for a number of tasks can be succinctly expressed by this framework, and show that new metrics can be naturally derived in a bottom-up way based on an output structure. We release a library that enables this derivation to create new metrics. Finally, we consider how specific characteristics of tasks motivate metric design decisions, and suggest possible modifications to existing metrics in line with those motivations.

14:00-15:30 (East Foyer / Virtual)
**#30 Air-Decoding: Attribute Distribution Reconstruction for Decoding-Time Controllable Text Generation**
*Tianqi Zhong, Quan Wang, Jingxuan Han, Yongdong Zhang and Zhendong Mao*

Controllable text generation (CTG) aims to generate text with desired attributes, and decoding-time-based methods have shown promising performance on this task. However, in this paper, we identify the phenomenon of Attribute Collapse for the first time. It causes the fluency of generated text to rapidly decrease when the control strength exceeds a critical value, rendering the text completely unusable. This limitation hinders the effectiveness of decoding methods in achieving high levels of controllability. To address this problem, we propose a novel lightweight decoding framework named Air-Decoding. Its main idea is reconstructing the attribute distributions to balance the weights between attribute words and non-attribute words to generate more fluent text. Specifically, we train prefixes by prefix-tuning to obtain attribute distributions. Then we design a novel attribute distribution reconstruction method to balance the obtained distributions and use the reconstructed distributions to guide language models for generation, effectively avoiding the issue of Attribute Collapse. Experiments on multiple CTG tasks prove that our method achieves a new state-of-the-art control performance.

14:00-15:30 (East Foyer / Virtual)
**#31 GreedyCAS: Unsupervised Scientific Abstract Segmentation with Normalized Mutual Information**
*Yingqiang Gao, Jessica Lam, Nianlong Gu and Richard Hahnloser*
The abstracts of scientific papers typically contain both premises (e.g., background and observations) and conclusions. Although conclusion sentences are highlighted in structured abstracts, in non-structured abstracts the concluding information is not explicitly marked, which makes the automatic segmentation of conclusions from scientific abstracts a challenging task. In this work, we explore Normalized Mutual Information (NMI) as a means for abstract segmentation. We consider each abstract as a recurrent cycle of sentences and place two segmentation boundaries by greedily optimizing the NMI score between the two segments, assuming that conclusions are strongly semantically linked with preceding premises. On non-structured abstracts, our proposed unsupervised approach GreedyCAS achieves the best performance across all evaluation metrics; on structured abstracts, GreedyCAS outperforms all baseline methods measured by $P_k$. The strong correlation of NMI to our evaluation metrics reveals the effectiveness of NMI for abstract segmentation.

14:00-15:30 (East Foyer / Virtual)
**#32 E-CORE: Emotion Correlation Enhanced Empathetic Dialogue Generation**
*Fengyi Fu, Lei Zhang, Quan Wang and Zhendong Mao*
Achieving empathy is a crucial step toward humanized dialogue systems. Current approaches for empathetic dialogue generation mainly perceive an emotional label to generate an empathetic response conditioned on it, which simply treat emotions independently, but ignore the intrinsic emotion correlation in dialogues, resulting in inaccurate emotion perception and unsuitable response generation. In this paper, we propose a novel emotion correlation enhanced empathetic dialogue generation framework, which comprehensively realizes emotion correlation learning, utilization, and supervising. Specifically, a multi-resolution emotion graph is devised to capture context-based emotion interactions from different resolutions, further modeling emotion correlation. Then we propose an emotion correlation enhanced decoder, with a novel correlation-aware aggregation and soft/hard strategy, respectively improving the emotion perception and response generation. Experimental results on the benchmark dataset demonstrate the superiority of our model in both empathetic perception and expression.

14:00-15:30 (East Foyer / Virtual)
**#33 Exploring Jiu-Jitsu Argumentation for Writing Peer Review Rebuttals**
*Sukannya Purkayastha, Anne Lauscher and Iryna Gurevych*
In many domains of argumentation, people's arguments are driven by so-called attitude roots, i.e., underlying beliefs and world views, and their corresponding attitude themes. Given the strength of these latent drivers of arguments, recent work in psychology suggests that instead of directly countering surface-level reasoning (e.g., falsifying the premises), one should follow an argumentation style inspired by the Jiu-Jitsu "soft" combat system: first, identify an arguer's attitude roots and themes, and then choose a prototypical rebuttal that is aligned with those drivers instead of trying to invalidate those. In this work, we are the first to explore Jiu-Jitsu argumentation for peer reviews by proposing the novel task of attitude and theme-guided rebuttal generation. To this end, we enrich an existing dataset for discourse structure in peer reviews with attitude roots, attitude themes, and canonical rebuttals. To facilitate this process, we recast established annotation concepts from the domain of peer reviews (e.g., aspects a review sentence is relating to) and train domain-specific models. We then propose strong rebuttal generation strategies, which we benchmark on our novel dataset for the task of end-to-end attitude and theme-guided rebuttal generation and two subtasks.

14:00-15:30 (East Foyer / Virtual)
**#34 Effects of sub-word segmentation on performance of transformer language models**
*Jue Hou, Anisia Katinskaia, Anh-Duc Vu and Roman Yangarber*
Language modeling is a fundamental task in natural language processing, which has been thoroughly explored with various architectures and hyperparameters. However, few studies focus on the effect of sub-word segmentation on the performance of language models (LMs). In this paper, we compare GPT and BERT models trained with the statistical segmentation algorithm BPE vs. two unsupervised algorithms for morphological segmentation — Morfessor and StateMorph. We train the models for several languages — including ones with very rich morphology — and compare their performance with different segmentation algorithms, vocabulary sizes, and model sizes. The results show that training with morphological segmentation allows the LMs to: (1) achieve lower perplexity, (2) converge more efficiently in terms of training time, and (3) achieve equivalent or better evaluation scores on downstream tasks. Lastly, we show that (4) LMs of smaller size using morphological segmentation can perform comparably to models of larger size trained with BPE — both in terms of (1) perplexity and (3) scores on downstream tasks. Points (2) and (4) impact on sustainability, since they reduce the model cost; and while 2 reduces cost only in the training phase, 4 does so also in the inference phase.

14:00-15:30 (East Foyer / Virtual)
**#35 GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation**
*Alexander Jones, Isaac Rayburn Caswell, Orhan Firat and Ishank Saxena*
Modern machine translation models and language models are able to translate without having been trained on parallel data, greatly expanding the set of languages that they can serve. However, these models still struggle in a variety of predictable ways, a problem that cannot be overcome without at least some trusted bilingual data. This work expands on a cheap and abundant resource to combat this problem: bilingual lexica. We test the efficacy of bilingual lexica in a real-world set-up, on 200-language translation models trained on web-crawled text. We present several findings: (1) using lexical data augmentation, we demonstrate sizable performance gains for unsupervised translation; (2) we compare several families of data augmentation, demonstrating that they yield similar improvements, and can be combined for even greater improvements; (3) we demonstrate the importance of carefully curated lexica over larger, noisier ones, especially with larger models; and (4) we compare the efficacy of multilingual lexicon data versus human-translated parallel data. Based on results from (3), we develop and open-source GATITOS, a high-quality, curated dataset in 168 tail languages, one of the first human-translated resources to cover many of these languages.

14:00-15:30 (East Foyer / Virtual)
**#36 Generating Commonsense Counterfactuals for Stable Relation Extraction**
*Xin Miao, Yongqi Li and Tieyun Qian*

Recent studies on counterfactual augmented data have achieved great success in the coarse-grained natural language processing tasks. However, existing methods encounter two major problems when dealing with the fine-grained relation extraction tasks. One is that they struggle to accurately identify causal terms under the invariant entity constraint. The other is that they ignore the commonsense constraint. To solve these problems, we propose a novel framework to generate commonsense counterfactuals for stable relation extraction. Specifically, to identify causal terms accurately, we introduce an intervention-based strategy and leverage a constituency parser for correction. To satisfy the commonsense constraint, we introduce the concept knowledge base WordNet and design a bottom-up relation expansion algorithm on it to uncover commonsense relations between entities. We conduct a series of comprehensive evaluations, including the low-resource, out-of-domain, and adversarial-attack settings. The results demonstrate that our framework significantly enhances the stability of base relation extraction models.

14:00-15:30 (East Foyer / Virtual)
**#37 Improving Biomedical Abstractive Summarisation with Knowledge Aggregation from Citation Papers**
*Chen Tang, Shun Wang, Tomas Goldsack and Chenghua Lin*
Abstracts derived from biomedical literature possess distinct domain-specific characteristics, including specialised writing styles and biomedical terminologies, which necessitate a deep understanding of the related literature. As a result, existing language models struggle to generate technical summaries that are on par with those produced by biomedical experts, given the absence of domain-specific background knowledge. This paper aims to enhance the performance of language models in biomedical abstractive summarisation by aggregating knowledge from external papers cited within the source article. We propose a novel attention-based citation aggregation model that integrates domain-specific knowledge from citation papers, allowing neural networks to generate summaries by leveraging both the paper content and relevant knowledge from citation papers. Furthermore, we construct and release a large-scale biomedical summarisation dataset that serves as a foundation for our research. Extensive experiments demonstrate that our model outperforms state-of-the-art approaches and achieves substantial improvements in abstractive biomedical text summarisation.

14:00-15:30 (East Foyer / Virtual)
**#38 Knowledge Graph Compression Enhances Diverse Commonsense Generation**
*EunJeong Hwang, Veronika Thost, Vered Shwartz and Tengfei Ma*
Generating commonsense explanations requires reasoning about commonsense knowledge beyond what is explicitly mentioned in the context. Existing models use commonsense knowledge graphs such as ConceptNet to extract a subgraph of relevant knowledge pertaining to concepts in the input. However, due to the large coverage and, consequently, vast scale of ConceptNet, the extracted subgraphs may contain loosely related, redundant and irrelevant information, which can introduce noise into the model. We propose to address this by applying a differentiable graph compression algorithm that focuses on the relevant knowledge for the task. The compressed subgraphs yield considerably more diverse outputs when incorporated into models for the tasks of generating commonsense and abductive explanations. Moreover, our model achieves better quality-diversity tradeoff than a large language model with 100 times the number of parameters. Our generic approach can be applied to additional NLP tasks that can benefit from incorporating external knowledge.

14:00-15:30 (East Foyer / Virtual)
**#39 Comparing Styles across Languages**
*Shreya Havaldar, Matthew Pressimone, Eric Wong and Lyle Ungar*
Understanding how styles differ across languages is advantageous for training both humans and computers to generate culturally appropriate text. We introduce an explanation framework to extract stylistic differences from multilingual LMs and compare styles across languages. Our framework (1) generates comprehensive style lexica in any language and (2) consolidates feature importances from LMs into comparable lexical categories. We apply this framework to compare politeness, creating the first holistic multilingual politeness dataset and exploring how politeness varies across four languages. Our approach enables an effective evaluation of how distinct linguistic categories contribute to stylistic variations and provides interpretable insights into how people communicate differently around the world.

14:00-15:30 (East Foyer / Virtual)
**#40 DecoMT: Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models**
*Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw and Nancy F. Chen*
This study investigates machine translation between related languages i.e., languages within the same family that share linguistic characteristics such as word order and lexical similarity. Machine translation through few-shot prompting leverages a small set of translation pair examples to generate translations for test sentences. This procedure requires the model to learn how to generate translations while simultaneously ensuring that token ordering is maintained to produce a fluent and accurate translation. We propose that for related languages, the task of machine translation can be simplified by leveraging the monotonic alignment characteristic of such languages. We introduce DecoMT, a novel approach of few-shot prompting that decomposes the translation process into a sequence of word chunk translations. Through automatic and human evaluation conducted on multiple related language pairs across various language families, we demonstrate that our proposed approach of decomposed prompting surpasses multiple established few-shot baseline approaches. For example, DecoMT outperforms the strong few-shot prompting BLOOM model with an average improvement of 8 chrF++ scores across the examined languages.

14:00-15:30 (East Foyer / Virtual)
**#41 SentiStream: A Co-Training Framework for Adaptive Online Sentiment Analysis in Evolving Data Streams**
*Yuhao Wu, Karthick Sharma, Chun Wei Seah and Shuhao Zhang*
Online sentiment analysis has emerged as a crucial component in numerous data-driven applications, including social media monitoring, customer feedback analysis, and online reputation management. Despite their importance, current methodologies falter in effectively managing the continuously evolving nature of data streams, largely due to their reliance on substantial, pre-existing labelled datasets. This paper presents **sentistream**, a novel co-training framework specifically designed for efficient sentiment analysis within dynamic data streams. Comprising unsupervised, semi-supervised, and stream merge modules, **sentistream** guarantees constant adaptability to evolving data landscapes. This research delves into the continuous adaptation of language models for online sentiment analysis, focusing on real-world applications. Experimental evaluations using data streams derived from three benchmark sentiment analysis datasets confirm that our proposed methodology surpasses existing approaches in terms of both accuracy and computational efficiency.

14:00-15:30 (East Foyer / Virtual)
**#42 Enhancing Low-resource Fine-grained Named Entity Recognition by Leveraging Coarse-grained Datasets**
*Su Ah Lee, Seokjin Oh and Woohwan Jung*
Named Entity Recognition (NER) frequently suffers from the problem of insufficient labeled data, particularly in fine-grained NER scenarios. Although $K$-shot learning techniques can be applied, their performance tends to saturate when the number of annotations exceeds several tens of labels. To overcome this problem, we utilize existing coarse-grained datasets that offer a large number of annotations. A straightforward approach to address this problem is pre-finetuning, which employs coarse-grained data for representation learning. However, it cannot directly utilize the relationships between fine-grained and coarse-grained entities, although a fine-grained entity type is likely to be a subcategory of a coarse-grained entity type. We propose a fine-grained NER model with a Fine-to-Coarse(F2C) mapping matrix to leverage the hierarchical structure explicitly. In addition, we present an inconsistency filtering method to eliminate coarse-grained entities that are inconsistent with

fine-grained entity types to avoid performance degradation. Our experimental results show that our method outperforms both $K$-shot learning and supervised learning methods when dealing with a small number of fine-grained annotations.

14:00-15:30 (East Foyer / Virtual)
**#43 Target-Agnostic Gender-Aware Contrastive Learning for Mitigating Bias in Multilingual Machine Translation**
*Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim and Kyomin Jung*
Gender bias is a significant issue in machine translation, leading to ongoing research efforts in developing bias mitigation techniques. However, most works focus on debiasing bilingual models without much consideration for multilingual systems. In this paper, we specifically target the gender bias issue of multilingual machine translation models for unambiguous cases where there is a single correct translation, and propose a bias mitigation method based on a novel approach. Specifically, we propose Gender-Aware Contrastive Learning, GACL, which encodes contextual gender information into the representations of non-explicit gender words. Our method is target language-agnostic and is applicable to pre-trained multilingual machine translation models via fine-tuning. Through multilingual evaluation, we show that our approach improves gender accuracy by a wide margin without hampering translation performance. We also observe that incorporated gender information transfers and benefits other target languages regarding gender accuracy. Finally, we demonstrate that our method is applicable and beneficial to models of various sizes.

14:00-15:30 (East Foyer / Virtual)
**#44 Bridging the Gap between Synthetic and Authentic Images for Multimodal Machine Translation**
*Wenyu Guo, Qingkai Fang, Dong Yu and Yang Feng*
Multimodal machine translation (MMT) simultaneously takes the source sentence and a relevant image as input for translation. Since there is no paired image available for the input sentence in most cases, recent studies suggest utilizing powerful text-to-image generation models to provide image inputs. Nevertheless, synthetic images generated by these models often follow different distributions compared to authentic images. Consequently, using authentic images for training and synthetic images for inference can introduce a distribution shift, resulting in performance degradation during inference. To tackle this challenge, in this paper, we feed synthetic and authentic images to the MMT model, respectively. Then we minimize the gap between the synthetic and authentic images by drawing close the input image representations of the Transformer Encoder and the output distributions of the Transformer Decoder. Therefore, we mitigate the distribution disparity introduced by the synthetic images during inference, thereby freeing the authentic images from the inference process. Experimental results show that our approach achieves state-of-the-art performance on the Multi30K En-De and En-Fr datasets, while remaining independent of authentic images during inference.

14:00-15:30 (East Foyer / Virtual)
**#45 Exploring All-In-One Knowledge Distillation Framework for Neural Machine Translation**
*Zhongjian Miao, Wen Zhang, Jinsong Su, Xiang Li, Jian Luan, Yidong Chen, Bin Wang and Min Zhang*
Conventional knowledge distillation(KD) approaches are commonly employed to compress neural machine translation(NMT) models. However, they only obtain one lightweight student each time. Consequently, we have to conduct KD multiple times when different students are required at the same time, which could be resource-intensive. Additionally, these students are individually optimized, and thus lack interactions with each other, leading to their potential not being fully exerted. In this work, we propose a novel All-In-One Knowledge Distillation(AIO-KD) framework for NMT, which generates multiple satisfactory students at once. Under AIO-KD, we first randomly extract fewer-layer subnetworks from the teacher as the sample students. Then, we jointly optimize the teacher and these students, where the students simultaneously learn the knowledge from the teacher and interact with other students via mutual learning. When utilized, we re-extract the candidate students, satisfying the specifications of various devices. Particularly, we adopt carefully-designed strategies for AIO-KD: 1) we dynamically detach gradients to prevent poorly-performed students from negatively affecting the teacher during the knowledge transfer, which could subsequently impact other students; 2) we design a two-stage mutual learning strategy, which alleviates the negative impacts of poorly-performed students on the early-stage student interactions. Extensive experiments and in-depth analyses on three benchmarks demonstrate the effectiveness and eco-friendliness of AIO-KD. Our source code is available at https://github.com/DeepLearnXMU/AIO-KD.

14:00-15:30 (East Foyer / Virtual)
**#46 Linking Surface Facts to Large-Scale Knowledge Graphs**
*Gorjan Radevski, Kiril Gashteovski, Chia-Chien Hung, Carolin Lawrence and Goran Glavaš*
Open Information Extraction (OIE) methods extract facts from natural language text in the form of ("subject"; "relation"; "object") triples. These facts are, however, merely surface forms, the ambiguity of which impedes their downstream usage; e.g., the surface phrase "Michael Jordan" may refer to either the former basketball player or the university professor. Knowledge Graphs (KGs), on the other hand, contain facts in a canonical (i.e., unambiguous) form, but their coverage is limited by a static schema (i.e., a fixed set of entities and predicates). To bridge this gap, we need the best of both worlds: (i) high coverage of free-text OIEs, and (ii) semantic precision (i.e., monosemy) of KGs. In order to achieve this goal, we propose a new benchmark with novel evaluation protocols that can, for example, measure fact linking performance on a granular triple slot level, while also measuring if a system has the ability to recognize that a surface form has no match in the existing KG. Our extensive evaluation of several baselines show that detection of out-of-KG entities and predicates is more difficult than accurate linking to existing ones, thus calling for more research efforts on this difficult task. We publicly release all resources (data, benchmark and code) on https://github.com/nec-research/fact-linking.

14:00-15:30 (East Foyer / Virtual)
**#47 ReTAG: Reasoning Aware Table to Analytic Text Generation**
*Deepanway Ghosal, Preksha Nema and Aravindan Raghuveer*
The task of table summarization involves generating text that both succinctly and accurately represents the table or a specific set of highlighted cells within a table. While significant progress has been made in table to text generation techniques, models still mostly generate descriptive summaries, which reiterates the information contained within the table in sentences. Through analysis of popular table to text benchmarks (ToTTo (Parikh et al., 2020 and InfoTabs (Gupta et al., 2020) we observe that in order to generate the ideal summary, multiple types of reasoning is needed coupled with access to knowledge beyond the scope of the table. To address this gap, we propose ReTAG, a table and reasoning aware model that uses vector-quantization to infuse different types of analytical reasoning into the output. ReTAG achieves 2.2%, 2.9% improvement on the PARENT metric in the relevant slice of ToTTo and InfoTabs for the table to text generation task over state of the art baselines. Through human evaluation, we observe that output from ReTAG is upto 12% more faithful and analytical compared to a strong table-aware model. To the best of our knowledge, ReTAG is the first model that can controllably use multiple reasoning methods within a structure-aware sequence to sequence model to surpass state of the art performance in multiple table to text tasks. We extend (and open source 35.6K analytical, 55.9k descriptive instances) the ToTTo, InfoTabs datasets with the reasoning categories used in each reference sentences.

14:00-15:30 (East Foyer / Virtual)
**#48 ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness**
*Jan Cegin, Jakub Simko and Peter Brusilovsky*

The emergence of generative large language models (LLMs) raises the question: what will be its impact on crowdsourcing? Traditionally, crowdsourcing has been used for acquiring solutions to a wide variety of human-intelligence tasks, including ones involving text generation, modification or evaluation. For some of these tasks, models like ChatGPT can potentially substitute human workers. In this study, we investigate whether this is the case for the task of paraphrase generation for intent classification. We apply data collection methodology of an existing crowdsourcing study (similar scale, prompts and seed data) using ChatGPT and Falcon-40B. We show that ChatGPT-created paraphrases are more diverse and lead to at least as robust models.

14:00-15:30 (East Foyer / Virtual)
**#49 Nearest Neighbor Machine Translation is Meta-Optimizer on Output Projection Layer**
*Ruize Gao, Zhirui Zhang, Yichao Du, Lemao Liu and Rui Wang*
Nearest Neighbor Machine Translation ($k$NN-MT) has achieved great success in domain adaptation tasks by integrating pre-trained Neural Machine Translation (NMT) models with domain-specific token-level retrieval. However, the reasons underlying its success have not been thoroughly investigated. In this paper, we comprehensively analyze $k$NN-MT through theoretical and empirical studies. Initially, we provide new insights into the working mechanism of $k$NN-MT as an efficient technique to implicitly execute gradient descent on the output projection layer of NMT, indicating that it is a specific case of model fine-tuning. Subsequently, we conduct multi-domain experiments and word-level analysis to examine the differences in performance between $k$NN-MT and entire-model fine-tuning. Our findings suggest that: ($i$) Incorporating $k$NN-MT with adapters yields comparable translation performance to fine-tuning on in-domain test sets, while achieving better performance on out-of-domain test sets; ($ii$) Fine-tuning significantly outperforms $k$NN-MT on the recall of in-domain low-frequency words, but this gap could be bridged by optimizing the context representations with additional adapter layers.

14:00-15:30 (East Foyer / Virtual)
**#50 Controlling Pre-trained Language Models for Grade-Specific Text Simplification**
*Sweta Agrawal and Marine Carpuat*
Text simplification systems rewrite text to make it more readable while preserving its content. However, what makes a text easy to read depends on the intended readers. Recent work has shown that pre-trained language models can simplify text using a wealth of techniques to control output simplicity, ranging from specifying only the desired reading grade level, to directly specifying low-level edit operations. Yet it remains unclear how to set these control parameters in practice. Existing approaches set them at the corpus level, disregarding the complexity of individual inputs and considering only one level of output complexity. In this work, we conduct an empirical study to understand how different control mechanisms impact the adequacy and simplicity of text simplification systems. Based on these insights, we introduce a simple method that predicts the edit operations required for simplifying a text for a specific grade level on an instance-per-instance basis. This approach improves the quality of the simplified outputs over corpus-level search-based heuristics.

14:00-15:30 (East Foyer / Virtual)
**#51 Multilingual $k$-Nearest-Neighbor Machine Translation**
*David Stap and Christof Monz*
$k$-nearest-neighbor machine translation has demonstrated remarkable improvements in machine translation quality by creating a datastore of cached examples. However, these improvements have been limited to high-resource language pairs, with large datastores, and remain a challenge for low-resource languages. In this paper, we address this issue by combining representations from multiple languages into a single datastore. Our results consistently demonstrate substantial improvements not only in low-resource translation quality (up to $+3.6$ BLEU), but also for high-resource translation quality (up to $+0.5$ BLEU). Our experiments show that it is possible to create multilingual datastores that are a quarter of the size, achieving a 5.3x speed improvement, by using linguistic similarities for datastore creation.

14:00-15:30 (East Foyer / Virtual)
**#52 Mirror: A Universal Framework for Various Information Extraction Tasks**
*Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai and Min Zhang*
Sharing knowledge between information extraction tasks has always been a challenge due to the diverse data formats and task variations. Meanwhile, this divergence leads to information waste and increases difficulties in building complex applications in real scenarios. Recent studies often formulate IE tasks as a triplet extraction problem. However, such a paradigm does not support multi-span and n-ary extraction, leading to weak versatility. To this end, we reorganize IE problems into unified multi-slot tuples and propose a universal framework for various IE tasks, namely Mirror. Specifically, we recast existing IE tasks as a multi-span cyclic graph extraction problem and devise a non-autoregressive graph decoding algorithm to extract all spans in a single step. It is worth noting that this graph structure is incredibly versatile, and it supports not only complex IE tasks, but also machine reading comprehension and classification tasks. We manually construct a corpus containing 57 datasets for model pretraining, and conduct experiments on 30 datasets across 8 downstream tasks. The experimental results demonstrate that our model has decent compatibility and outperforms or reaches competitive performance with SOTA systems under few-shot and zero-shot settings. The code, model weights, and pretraining corpus are available at https://github.com/Spico197/Mirror .

14:00-15:30 (East Foyer / Virtual)
**#53 Lazy-k Decoding: Constrained Decoding for Information Extraction**
*Arthur Hemmer, Mickael Coustaty, Nicola Bartolo, Jerome Brachat and Jean-marc Ogier*
We explore the possibility of improving probabilistic models in structured prediction. Specifically, we combine the models with constrained decoding approaches in the context of token classification for information extraction. The decoding methods search for constraint-satisfying label-assignments while maximizing the total probability. To do this, we evaluate several existing approaches, as well as propose a novel decoding method called Lazy-$k$. Our findings demonstrate that constrained decoding approaches can significantly improve the models' performances, especially when using smaller models. The Lazy-$k$ approach allows for more flexibility between decoding time and accuracy. The code for using Lazy-$k$ decoding can be found at https://github.com/ArthurDevNL/lazyk.

14:00-15:30 (East Foyer / Virtual)
**#54 Improved Pseudo Data for Machine Translation Quality Estimation with Constrained Beam Search**
*Xiang Geng, Yu Zhang, Zhejian Lai, Shuaijie She, Wei Zou, Shimin Tao, Hao Yang, Jiajun Chen and Shujian Huang*
Machine translation (MT) quality estimation (QE) is a crucial task to estimate the quality of MT outputs when reference translations are unavailable. Many studies focus on generating pseudo data using large parallel corpus and achieve remarkable success in the supervised setting. However, pseudo data solutions are less satisfying in unsupervised scenarios because the pseudo labels are inaccurate or the pseudo translations differ from the real ones. To address these problems, we propose to generate pseudo data using the MT model with constrained beam search (CBSQE). CBSQE preserves the reference parts with high MT probabilities as correct translations, while the rest parts as the wrong ones for MT generation. Therefore, CBSQE can reduce the false negative labels caused by synonyms. Overall, beam search will prefer a more real hypothesis with a higher MT generation likelihood. Extensive experiments demonstrate that CBSQE outperforms strong baselines in both supervised and unsupervised settings. Analyses further show the superiority of CBSQE. The code is available at https://github.com/NJUNLP/njuqe.

14:00-15:30 (East Foyer / Virtual)
**#55 Adapting Language Models to Compress Contexts**
*Alexis Chevalier, Alexander Wettig, Anirudh Ajith and Danqi Chen*
Transformer-based language models (LMs) are powerful and widely-applicable tools, but their usefulness is constrained by a finite context window and the expensive computational cost of processing long text documents. We propose to adapt pre-trained LMs into AutoCompressors. These language models are capable of compressing long contexts into summary vectors, which are then accessible to the model as soft prompts. Summary vectors are trained with an unsupervised objective, whereby long documents are processed in segments, and summary vectors from all previous segments are used in language modeling. We fine-tune OPT and Llama-2 models on sequences of up to 30,720 tokens and show that AutoCompressors can utilize long contexts to improve perplexity. We evaluate AutoCompressors on in-context learning by compressing task demonstrations and find that summary vectors are good substitutes for plain-text demonstrations, increasing accuracy while reducing inference costs. Finally, we explore the benefits of pre-computing summary vectors for large corpora by applying summary vectors to retrieval-augmented language modeling and a passage re-ranking task. Overall, AutoCompressors emerge as a simple and inexpensive solution to extend the context window of LMs while speeding up inference over long contexts.

14:00-15:30 (East Foyer / Virtual)
**#56 Task-Adaptive Tokenization: Enhancing Long-Form Text Generation Efficacy in Mental Health and Beyond**
*Siyang Liu, Naihao Deng, Sahand Sabour, Yilin Jia, Minlie Huang and Rada Mihalcea*
We propose task-adaptive tokenization as a way to adapt the generation pipeline to the specifics of a downstream task and enhance long-form generation in mental health. Inspired by insights from cognitive science, our task-adaptive tokenizer samples variable segmentations from multiple outcomes, with sampling probabilities optimized based on task-specific data. We introduce a strategy for building a specialized vocabulary and introduce a vocabulary merging protocol that allows for the integration of task-specific tokens into the pre-trained model's tokenization step. Through extensive experiments on psychological question-answering tasks in both Chinese and English, we find that our task-adaptive tokenization approach brings a significant improvement in generation performance while using up to 60% fewer tokens. Preliminary experiments point to promising results when using our tokenization approach with very large language models.

14:00-15:30 (East Foyer / Virtual)
**#57 Towards a Better Understanding of Variations in Zero-Shot Neural Machine Translation Performance**
*Shaomu Tan and Christof Monz*
Multilingual Neural Machine Translation (MNMT) facilitates knowledge sharing but often suffers from poor zero-shot (ZS) translation qualities. While prior work has explored the causes of overall low zero-shot translation qualities, our work introduces a fresh perspective: the presence of significant variations in zero-shot performance. This suggests that MNMT does not uniformly exhibit poor zero-shot capability; instead, certain translation directions yield reasonable results. Through systematic experimentation, spanning 1,560 language directions across 40 languages, we identify three key factors contributing to high variations in ZS NMT performance: 1) target-side translation quality, 2) vocabulary overlap, and 3) linguistic properties. Our findings highlight that the target side translation quality is the most influential factor, with vocabulary overlap consistently impacting zero-shot capabilities. Additionally, linguistic properties, such as language family and writing system, play a role, particularly with smaller models. Furthermore, we suggest that the off-target issue is a symptom of inadequate performance, emphasizing that zero-shot translation challenges extend beyond addressing the off-target problem. To support future research, we release the data and models as a benchmark for the study of ZS NMT.

14:00-15:30 (East Foyer / Virtual)
**#58 Personalized Distillation: Empowering Open-Sourced LLMs with Adaptive Learning for Code Generation**
*Hailin Chen, Amrita Saha, Steven Hoi and Shafiq Joty*
With the rise of powerful closed-sourced LLMs (ChatGPT, GPT-4), there are increasing interests in distilling the capabilies of close-sourced LLMs to smaller open-sourced LLMs. Previous distillation methods usually prompt ChatGPT to generate a set of instructions and answers, for the student model to learn. However, such standard distillation approach neglects the merits and conditions of the student model. Inspired by modern teaching principles, we design a personalised distillation process, in which the student attempts to solve a task first, then the teacher provides an adaptive refinement for the student to improve. Instead of feeding the student with teacher's prior, personalised distillation enables personalised learning for the student model, as it only learns on examples it makes mistakes upon and learns to improve its own solution. On code generation, personalised distillation consistently outperforms standard distillation with only one third of the data. With only 2.5-3K personalised examples that incur a data-collection cost of 4-6$, we boost CodeGen-mono-16B by 7% to achieve 36.4% pass@1 and StarCoder by 12.2% to achieve 45.8% pass@1 on HumanEval.

14:00-15:30 (East Foyer / Virtual)
**#59 Detecting Propaganda Techniques in Code-Switched Social Media Text**
*Muhammad Umar Salman, Asif Hanif, Shady Shehata and Preslav Nakov*
Propaganda is a form of communication intended to influence the opinions and the mindset of the public to promote a particular agenda. With the rise of social media, propaganda has spread rapidly, leading to the need for automatic propaganda detection systems. Most work on propaganda detection has focused on high-resource languages, such as English, and little effort has been made to detect propaganda for low-resource languages. Yet, it is common to find a mix of multiple languages in social media communication, a phenomenon known as code-switching. Code-switching combines different languages within the same text, which poses a challenge for automatic systems. Considering this premise, we propose a novel task of detecting propaganda techniques in code-switched text. To support this task, we create a corpus of 1,030 texts code-switching between English and Roman Urdu, annotated with 20 propaganda techniques at fragment-level. We perform a number of experiments contrasting different experimental setups, and we find that it is important to model the multilinguality directly rather than using translation as well as to use the right fine-tuning strategy. We plan to publicly release our code and dataset.

14:00-15:30 (East Foyer / Virtual)
**#60 Beyond Shared Vocabulary: Increasing Representational Word Similarities across Languages for Multilingual Machine Translation**
*Di Wu and Christof Monz*
Using a shared vocabulary is common practice in Multilingual Neural Machine Translation (MNMT). In addition to its simple design, shared tokens play an important role in positive knowledge transfer, which manifests naturally when the shared tokens refer to similar meanings across languages. However, when words overlap is small, e.g., using different writing systems, transfer is inhibited. In this paper, we propose a re-parameterized method for building embeddings to alleviate this problem. More specifically, we define word-level information transfer pathways via word equivalence classes and rely on graph networks to fuse word embeddings across languages. Our experiments demonstrate the advantages of our approach: 1) the semantics of embeddings are better aligned across languages, 2) our method achieves evident BLEU improvements on high- and low-resource MNMT, and 3) only less than 1.0% additional trainable parameters are required with a limited increase in computational costs, while the inference time is identical to baselines.

14:00-15:30 (East Foyer / Virtual)

**#61 RainProof: An Umbrella to Shield Text Generator from Out-Of-Distribution Data**
*Maxime Darrin, Pablo Piantanida and Pierre Colombo*
Implementing effective control mechanisms to ensure the proper functioning and security of deployed NLP models, from translation to chatbots, is essential. A key ingredient to ensure safe system behaviour is Out-Of-Distribution (OOD) detection, which aims to detect whether an input sample is statistically far from the training distribution. Although OOD detection is a widely covered topic in classification tasks, most methods rely on hidden features output by the encoder. In this work, we focus on leveraging soft-probabilities in a black-box framework, i.e. we can access the soft-predictions but not the internal states of the model. Our contributions include: (i) RAINPROOF a Relative informAItioN Projection OOD detection framework; and (ii) a more operational evaluation setting for OOD detection. Surprisingly, we find that OOD detection is not necessarily aligned with task-specific measures. The OOD detector may filter out samples well processed by the model and keep samples that are not, leading to weaker performance. Our results show that RAINPROOF provides OOD detection methods more aligned with task-specific performance metrics than traditional OOD detectors.

14:00-15:30 (East Foyer / Virtual)
**#62 Elaborative Simplification as Implicit Questions Under Discussion**
*Yating Wu, William Berkeley Sheffield, Kyle Mahowald and Junyi Jessy Li*
Automated text simplification, a technique useful for making text more accessible to people such as children and emergent bilinguals, is often thought of as a monolingual translation task from complex sentences to simplified sentences using encoder-decoder models. This view fails to account for elaborative simplification, where new information is added into the simplified text. This paper proposes to view elaborative simplification through the lens of the Question Under Discussion (QUD) framework, providing a robust way to investigate what writers elaborate upon, how they elaborate, and how elaborations fit into the discourse context by viewing elaborations as explicit answers to implicit questions. We introduce ELABQUD, consisting of 1.3K elaborations accompanied with implicit QUDs, to study these phenomena. We show that explicitly modeling QUD (via question generation) not only provides essential understanding of elaborative simplification and how the elaborations connect with the rest of the discourse, but also substantially improves the quality of elaboration generation.

14:00-15:30 (East Foyer / Virtual)
**#63 trlX: A Framework for Large Scale Reinforcement Learning from Human Feedback**
*Alexander Havrilla, Maksym Zhuravinskyi, Duy Van Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Gregory Anthony and Louis Castricato*
Reinforcement learning from human feedback (**RLHF**) utilizes human feedback to better align large language models with human preferences via online optimization against a learned reward model. Current RLHF paradigms rely on Proximal Policy Optimization (**PPO**), which quickly becomes a challenge to implement and scale up to large architectures. To address this difficulty we present the **AutoRLHF** library as a feature complete open-source framework for RLHF fine-tuning of models up to and exceeding 70 billion parameters. To do so we implement support for multiple types of distributed training including distributed data parallel, model sharded, as well as tensor, sequential, and pipeline parallelism. Additionally, we implement compute and memory saving features, giving AutoRLHF the flexibility to support users with a wide range of compute resources. This includes offline RL methods like Implicit Language Q Learning (**ILQL**) as a compute efficient alternative to PPO. We find offline fine-tuning offers competitive performance relative to online algorithms while being easier to implement, train, and scale. To evaluate our framework we train RLHF models on two separate well-known tasks using publicly available human preference data. Models trained with AutoRLHF achieve preference win-rates over baselines at rates comparable to the original works.

14:00-15:30 (East Foyer / Virtual)
**#64 We Are What We Repeatedly Do: Inducing and Deploying Habitual Schemas in Persona-Based Responses**
*Benjamin Kane and Lenhart K. Schubert*
Many practical applications of dialogue technology require the generation of responses according to a particular developer-specified persona. While a variety of personas can be elicited from recent large language models, the opaqueness and unpredictability of these models make it desirable to be able to specify personas in an explicit form. In previous work, personas have typically been represented as sets of one-off pieces of self-knowledge that are retrieved by the dialogue system for use in generation. However, in realistic human conversations, personas are often revealed through story-like narratives that involve rich habitual knowledge – knowledge about kinds of events that an agent often participates in (e.g., work activities, hobbies, sporting activities, favorite entertainments, etc.), including typical goals, sub-events, preconditions, and postconditions of those events. We capture such habitual knowledge using an explicit schema representation, and propose an approach to dialogue generation that retrieves relevant schemas to condition a large language model to generate persona-based responses. Furthermore, we demonstrate a method for bootstrapping the creation of such schemas by first generating generic passages from a set of simple facts, and then inducing schemas from the generated passages.

14:00-15:30 (East Foyer / Virtual)
**#65 Don't Take This Out of Context!: On the Need for Contextual Models and Evaluations for Stylistic Rewriting**
*Akhila Yerukola, Xuhui Zhou, Elizabeth Clark and Maarten Sap*
Most existing stylistic text rewriting methods and evaluation metrics operate on a sentence level, but ignoring the broader context of the text can lead to preferring generic, ambiguous, and incoherent rewrites. In this paper, we investigate integrating the preceding textual context into both the *rewriting* and *evaluation* stages of stylistic text rewriting, and introduce a new composite contextual evaluation metric `CtxSimFit` that combines similarity to the original sentence with contextual cohesiveness. We comparatively evaluate non-contextual and contextual rewrites in formality, toxicity, and sentiment transfer tasks. Our experiments show that humans significantly prefer contextual rewrites as more fitting and natural over non-contextual ones, yet existing sentence-level automatic metrics (e.g., ROUGE, SBERT) correlate poorly with human preferences ($\rho$=0–0.3). In contrast, human preferences are much better reflected by both our novel `CtxSimFit` ($\rho$=0.7–0.9) as well as proposed context-infused versions of common metrics ($\rho$=0.4–0.7). Overall, our findings highlight the importance of integrating context into the generation and especially the evaluation stages of stylistic text rewriting.

14:00-15:30 (East Foyer / Virtual)
**#66 Multilingual Simplification of Medical Texts**
*Sebastian Antony Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J Ramanathan, Wei Xu, Byron C Wallace and Junyi Jessy Li*
Automated text simplification aims to produce simple versions of complex texts. This task is especially useful in the medical domain, where the latest medical findings are typically communicated via complex and technical articles. This creates barriers for laypeople seeking access to up-to-date medical findings, consequently impeding progress on health literacy. Most existing work on medical text simplification has focused on monolingual settings, with the result that such evidence would be available only in just one language (most often, English). This work addresses this limitation via multilingual simplification, i.e., directly simplifying complex texts into simplified texts in multiple languages. We introduce MultiCochrane, the first sentence-aligned multilingual text simplification dataset for the medical domain in four languages: English, Spanish, French, and Farsi. We evaluate fine-tuned and zero-shot models across these languages with extensive human assessments and analyses. Although models can generate viable simplified texts, we identify several outstanding challenges that this dataset might be used to address.

14:00-15:30 (East Foyer / Virtual)

**#67 A Comprehensive Evaluation of Biomedical Entity Linking Models**
*David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández and Cassie S. Mitchell*
Biomedical entity linking (BioEL) is the process of connecting entities referenced in documents to entries in biomedical databases such as the Unified Medical Language System (UMLS) or Medical Subject Headings (MeSH). The study objective was to comprehensively evaluate nine recent state-of-the-art biomedical entity linking models under a unified framework. We compare these models along axes of (1) accuracy, (2) speed, (3) ease of use, (4) generalization, and (5) adaptability to new ontologies and datasets. We additionally quantify the impact of various preprocessing choices such as abbreviation detection. Systematic evaluation reveals several notable gaps in current methods. In particular, current methods struggle to correctly link genes and proteins and often have difficulty effectively incorporating context into linking decisions. To expedite future development and baseline testing, we release our unified evaluation framework and all included models on GitHub at https://github.com/davidkartchner/biomedical-entity-linking

14:00-15:30 (East Foyer / Virtual)
**#68 GLEN: General-Purpose Event Detection for Thousands of Types**
*Sha Li, Qiusi Zhan, Kathryn Conger, Martha Palmer, Heng Ji and Jiawei Han*
The progress of event extraction research has been hindered by the absence of wide-coverage, large-scale datasets. To make event extraction systems more accessible, we build a general-purpose event detection dataset GLEN, which covers 205K event mentions with 3,465 different types, making it more than 20x larger in ontology than today's largest event dataset. GLEN is created by utilizing the DWD Overlay, which provides a mapping between Wikidata Qnodes and PropBank rolesets. This enables us to use the abundant existing annotation for PropBank as distant supervision. In addition, we also propose a new multi-stage event detection model specifically designed to handle the large ontology size in GLEN. We show that our model exhibits superior performance compared to a range of baselines including InstructGPT. Finally, we perform error analysis and show that label noise is still the largest challenge for improving performance for this new dataset.

14:00-15:30 (East Foyer / Virtual)
**#69 MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions**
*Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts and Danqi Chen*
The information stored in large language models (LLMs) falls out of date quickly, and retraining from scratch is often not an option. This has recently given rise to a range of techniques for injecting new facts through updating model weights. Current evaluation paradigms are extremely limited, mainly validating the recall of edited facts, but changing one fact should cause rippling changes to the model's related beliefs. If we edit the UK Prime Minister to now be Rishi Sunak, then we should get a different answer to Who is married to the British Prime Minister? In this work, we present a benchmark MQuAKE (Multi-hop Question Answering for Knowledge Editing) comprising multi-hop questions that assess whether edited models correctly answer questions where the answer should change as an entailed consequence of edited facts. While we find that current knowledge-editing approaches can recall edited facts accurately, they fail catastrophically on the constructed multi-hop questions. We thus propose a simple memory-based approach, MeLLo, which stores all edited facts externally while prompting the language model iteratively to generate answers that are consistent with the edited facts. While MQuAKE remains challenging, we show that MeLLo scales well with LLMs (up to 175B) and outperforms previous model editors by a large margin.

14:00-15:30 (East Foyer / Virtual)
**#70 Aligning Large Language Models through Synthetic Feedback**
*Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo and Minjoon Seo*
Aligning large language models (LLMs) to human values has become increasingly important as it enables sophisticated steering of LLMs. However, it requires significant human demonstrations and feedback or distillation from proprietary LLMs such as ChatGPT. In this work, we propose a novel alignment learning framework with synthetic feedback not dependent on extensive human annotations and proprietary LLMs. First, we perform reward modeling (RM) with synthetic feedback by contrasting responses from vanilla LLMs with various sizes and prompts. Then, we use the RM to simulate high-quality demonstrations to train a supervised policy and further optimize the model with reinforcement learning. Our resulting model, Aligned Language Model with Synthetic Training dataset (ALMoST), outperforms recent open-sourced models, which are trained on the outputs of InstructGPT or human-annotated demonstrations, in alignment benchmarks. In human evaluation, our model is preferred to Alpaca and Dolly-v2, 55.0% and 58.5% of the time, respectively. Further analyses demonstrate the efficacy and importance of synthetic feedback in our framework.

14:00-15:30 (East Foyer / Virtual)
**#71 Expository Text Generation: Imitate, Retrieve, Paraphrase**
*Nishant Balepur, Jie Huang and Kevin Chang*
Expository documents are vital resources for conveying complex information to readers. Despite their usefulness, writing expository text by hand is a challenging process that requires careful content planning, obtaining facts from multiple sources, and the ability to clearly synthesize these facts. To ease these burdens, we propose the task of expository text generation, which seeks to automatically generate an accurate and stylistically consistent expository text for a topic by intelligently searching a knowledge source. We solve our task by developing IRP, a framework that overcomes the limitations of retrieval-augmented models and iteratively performs content planning, fact retrieval, and rephrasing. Through experiments on three diverse, newly-collected datasets, we show that IRP produces factual and organized expository texts that accurately inform readers.

14:00-15:30 (East Foyer / Virtual)
**#72 Text Fact Transfer**
*Nishant Balepur, Jie Huang and Kevin Chang*
Text style transfer is a prominent task that aims to control the style of text without inherently changing its factual content. To cover more text modification applications, such as adapting past news for current events and repurposing educational materials, we propose the task of text fact transfer, which seeks to transfer the factual content of a source text between topics without modifying its style. We find that existing language models struggle with text fact transfer, due to their inability to preserve the specificity and phrasing of the source text, and tendency to hallucinate errors. To address these issues, we design ModQGA, a framework that minimally modifies a source text with a novel combination of end-to-end question generation and specificity-aware question answering. Through experiments on four existing datasets adapted for text fact transfer, we show that ModQGA can accurately transfer factual content without sacrificing the style of the source text.

14:00-15:30 (East Foyer / Virtual)
**#73 Multi-level Contrastive Learning for Script-based Character Understanding**
*Dawei Li, Hengyuan Zhang, Yanran Li and Shiping Yang*
In this work, we tackle the scenario of understanding characters in scripts, which aims to learn the characters' personalities and identities from their utterances. We begin by analyzing several challenges in this scenario, and then propose a multi-level contrastive learning framework to capture characters' global information in a fine-grained manner. To validate the proposed framework, we conduct extensive experiments on three character understanding sub-tasks by comparing with strong pre-trained language models, including SpanBERT, Longformer, BigBird and ChatGPT-3.5. Experimental results demonstrate that our method improves the performances by a considerable margin. Through further

in-depth analysis, we show the effectiveness of our method in addressing the challenges and provide more hints on the scenario of character understanding. We will open-source our work in this URL.

14:00-15:30 (East Foyer / Virtual)
**#74 MMNMT: Modularizing Multilingual Neural Machine Translation with Flexibly Assembled MoE and Dense Blocks**
*Shangjie Li, Xiangpeng Wei, Shaolin Zhu, Jun Xie, Baosong Yang and Deyi Xiong*
Mixture-of-Experts (MoE) based sparse architectures can significantly increase model capacity with sublinear computational overhead, which are hence widely used in massively multilingual neural machine translation (MNMT). However, they are prone to overfitting on low-resource language translation. In this paper, we propose a modularized MNMT framework that is able to flexibly assemble dense and MoE-based sparse modules to achieve the best of both worlds. The training strategy of the modularized MNMT framework consists of three stages: (1) Pre-training basic MNMT models with different training objectives or model structures, (2) Initializing modules of the framework with pre-trained couterparts (e.g., encoder, decoder and embedding layers) from the basic models and (3) Fine-tuning the modularized MNMT framework to fit modules from different models together. We pre-train three basic MNMT models from scratch: a dense model, an MoE-based sparse model and a new MoE model, termed as MoE-LGR that explores multiple Language-Group-specifc Routers to incorporate language group knowledge into MNMT. The strengths of these pre-trained models are either on low-resource language translation, high-resource language translation or zero-shot translation. Our modularized MNMT framework attempts to incorporate these advantages into a single model with reasonable initialization and fine-tuning. Experiments on widely-used benchmark datasets demonstrate that the proposed modularized MNMT framwork substantially outperforms both MoE and dense models on high- and low-resource language translation as well as zero-shot translation. Our framework facilitates the combination of different methods with their own strengths and recycling off-the-shelf models for multilingual neural machine translation. Codes are available at https://github.com/lishangjie1/MMNMT.

14:00-15:30 (East Foyer / Virtual)
**#75 Rethinking Word-Level Auto-Completion in Computer-Aided Translation**
*Xingyu Chen, Lemao Liu, Guoping Huang, Zhirui Zhang, Mingming Yang, Shuming Shi and Rui Wang*
Word-level auto-completion (WLAC) plays a crucial role in Computer-Assisted Translation. While previous studies have primarily focused on designing complex model architectures, this paper takes a different perspective by rethinking the fundamental question: what kind of words are good auto-completions? We introduce a measurable criterion to address this question and discover that existing WLAC models often fail to meet this criterion. Building upon this observation, we propose an effective approach to enhance WLAC performance by promoting adherence to the criterion. Notably, the proposed approach is general and can be applied to various encoder-based architectures. Through extensive experiments, we demonstrate that our approach outperforms the top-performing system submitted to the WLAC shared tasks in WMT2022, while utilizing significantly smaller model sizes.

14:00-15:30 (East Foyer / Virtual)
**#76 PROSE: A Pronoun Omission Solution for Chinese-English Spoken Language Translation**
*Ke Wang, Xiutian Zhao, Yanghui Li and Wei Peng*
Neural Machine Translation (NMT) systems encounter a significant challenge when translating a pro-drop ('pronoun-dropping') language (e.g., Chinese) to a non-pro-drop one (e.g., English), since the pro-drop phenomenon demands NMT systems to recover omitted pronouns. This unique and crucial task, however, lacks sufficient datasets for benchmarking. To bridge this gap, we introduce PROSE, a new benchmark featured in diverse pro-drop instances for document-level Chinese-English spoken language translation. Furthermore, we conduct an in-depth investigation of the pro-drop phenomenon in spoken Chinese on this dataset, reconfirming that pro-drop reduces the performance of NMT systems in Chinese-English translation. To alleviate the negative impact introduced by pro-drop, we propose Mention-Aware Semantic Augmentation, a novel approach that leverages the semantic embedding of dropped pronouns to augment training pairs. Results from the experiments on four Chinese-English translation corpora show that our proposed method outperforms existing methods regarding omitted pronoun retrieval and overall translation quality.

14:00-15:30 (East Foyer / Virtual)
**#77 MailEx: Email Event and Argument Extraction**
*Saurabh Srivastava, Gaurav Singh, Shou Matsumoto, Ali K Raz, Paulo Costa, Joshua Campbell Poore and Ziyu Yao*
In this work, we present the first dataset, MailEx, for performing event extraction from conversational email threads. To this end, we first proposed a new taxonomy covering 10 event types and 76 arguments in the email domain. Our final dataset includes 1.5K email threads and ∼4K emails, which are annotated with a total of ∼8K event instances. To understand the task challenges, we conducted a series of experiments comparing three types of approaches, i.e., fine-tuned sequence labeling, fine-tuned generative extraction, and few-shot in-context learning. Our results showed that the task of email event extraction is far from being addressed, due to challenges lying in, e.g., extracting non-continuous, shared trigger spans, extracting non-named entity arguments, and modeling the email conversational history. Our work thus suggests more future investigations in this domain-specific event extraction task.

14:00-15:30 (East Foyer / Virtual)
**#78 Generating Data for Symbolic Language with Large Language Models**
*Jiacheng Ye, Chengzu Li, Lingpeng Kong and Tao Yu*
While large language models (LLMs) bring not only performance but also complexity, recent work has started to turn LLMs into data generators rather than task inferencers, where another affordable task model is trained for efficient deployment and inference. However, such an approach has primarily been applied to natural language tasks, and has not yet been explored for symbolic language tasks with complex structured outputs (e.g., semantic parsing and code generation). In this paper, we propose SymGen which utilizes LLMs for generating various annotation-expensive symbolic language data. SymGen consists of an informative prompt to steer generation and an agreement-based verifier to improve data correctness. We conduct extensive experiments on six symbolic language tasks across various settings. Compared with the LLMs, we demonstrate the 1%-sized task model can achieve comparable or better performance, largely cutting inference and deployment costs. We also show that generated data with only a few human demonstrations can be as effective as over 10 times the amount of human-annotated data when training the task model, saving a considerable amount of annotation effort. SymGen takes a step toward data generation for annotation-expensive complex tasks, and we release the code at URL.

14:00-15:30 (East Foyer / Virtual)
**#79 Empathy Intent Drives Empathy Detection**
*Liting Jiang, Di Wu, Bohui Mao, Yanbing Li and Wushour Slamu*
Empathy plays an important role in the human dialogue. Detecting the empathetic direction expressed by the user is necessary for empathetic dialogue systems because it is highly relevant to understanding the user's needs. Several studies have shown that empathy intent information improves the ability to response capacity of empathetic dialogue. However, the interaction between empathy detection and empathy intent recognition has not been explored. To this end, we invite 3 experts to manually annotate the healthy empathy detection datasets IEMPATHIZE and TwittEmp with 8 empathy intent labels, and perform joint training for the two tasks. Empirical study has shown that the introduction of empathy intent recognition task can improve the accuracy of empathy detection task, and we analyze possible reasons for this improvement.

To make joint training of the two tasks more challenging, we propose a novel framework, Cascaded Label Signal Network, which uses the cascaded interactive attention module and the label signal enhancement module to capture feature exchange information between empathy and empathy intent representations. Experimental results show that our framework outperforms all baselines under both settings on the two datasets.

14:00-15:30 (East Foyer / Virtual)
**#80 Condensing Multilingual Knowledge with Lightweight Language-Specific Modules**
*Haoran Xu, Weiting Tan, Shuyue Stella Li, Yunmo Chen, Benjamin Van Durme, Philipp Koehn and Kenton Murray*
Incorporating language-specific (LS) modules or Mixture-of-Experts (MoE) are proven methods to boost performance in multilingual model performance, but the scalability of these approaches to hundreds of languages or experts tends to be hard to manage. We present Language-specific Matrix Synthesis (LMS), a novel method that addresses the issue. LMS utilizes parameter-efficient and lightweight modules, reducing the number of parameters while outperforming existing methods, e.g., +1.73 BLEU over Switch Transformer on OPUS-100 multilingual translation. Additionally, we introduce Fuse Distillation (FD) to condense multilingual knowledge from multiple LS modules into a single shared module, improving model inference and storage efficiency. Our approach demonstrates superior scalability and performance compared to state-of-the-art methods.

14:00-15:30 (East Foyer / Virtual)
**#81 Fast and Accurate Factual Inconsistency Detection Over Long Documents**
*Barrett Martin Lattimer, Patrick CHen, Xinyuan Zhang and Yi Yang*
Generative AI models exhibit remarkable potential; however, hallucinations across various tasks present a significant challenge, particularly for longer inputs that current approaches struggle to address effectively. We introduce SCALE (Source Chunking Approach for Large-scale inconsistency Evaluation), a task-agnostic model for detecting factual inconsistencies using a novel chunking strategy. Specifically, SCALE is a Natural Language Inference (NLI) based model that uses large text chunks to condition over long texts. This approach achieves state-of-the-art performance in factual inconsistency detection for diverse tasks and long inputs. Additionally, we leverage the chunking mechanism and employ a novel algorithm to explain SCALE's decisions through relevant source sentence retrieval. Our evaluations reveal that SCALE outperforms existing methods on both standard benchmarks and a new long-form dialogue dataset ScreenEval we constructed. Moreover, SCALE surpasses competitive systems in efficiency and model explanation evaluations. We have released our code and data publicly to GitHub.

14:00-15:30 (East Foyer / Virtual)
**#82 Be Selfish, But Wisely: Investigating the Impact of Agent Personality in Mixed-Motive Human-Agent Interactions**
*Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas and Jonathan Gratch*
A natural way to design a negotiation dialogue system is via self-play RL: train an agent that learns to maximize its performance by interacting with a simulated user that has been designed to imitate human-human dialogue data. Although this procedure has been adopted in prior work, we find that it results in a fundamentally flawed system that fails to learn the value of compromise in a negotiation, which can often lead to no agreements (i.e., the partner walking away without a deal), ultimately hurting the model's overall performance. We investigate this observation in the context of DealOrNoDeal task, a multi-issue negotiation over books, hats, and balls. Grounded in negotiation theory from Economics, we modify the training procedure in two novel ways to design agents with diverse personalities and analyze their performance with human partners. We find that although both techniques show promise, a selfish agent, which maximizes its own performance while also avoiding walkaways, performs superior to other variants by implicitly learning to generate value for both itself and the negotiation partner. We discuss the implications of our findings for what it means to be a successful negotiation dialogue system and how these systems should be designed in the future.

14:00-15:30 (East Foyer / Virtual)
**#83 MAF: Multi-Aspect Feedback for Improving Reasoning in Large Language Models**
*Deepak Nathani, David Wang, Liangming Pan and William Yang Wang*
Language Models (LMs) have shown impressive performance in various natural language tasks. However, when it comes to natural language reasoning, LMs still face challenges such as hallucination, generating incorrect intermediate reasoning steps, and making mathematical errors. Recent research has focused on enhancing LMs through *self-improvement* using feedback. Nevertheless, existing approaches relying on a single generic feedback source fail to address the diverse error types found in LM-generated reasoning chains. In this work, we propose **Multi-Aspect Feedback**, an iterative refinement framework that integrates multiple feedback modules, including frozen LMs and external tools, each focusing on a specific error category. Our experimental results demonstrate the efficacy of our approach to addressing several errors in the LM-generated reasoning chain and thus improving the overall performance of an LM in several reasoning tasks. We see an improvement of up to 20% in Mathematical Reasoning and up to 18% in Logical Entailment.

14:00-15:30 (East Foyer / Virtual)
**#84 Does the Correctness of Factual Knowledge Matter for Factual Knowledge-Enhanced Pre-trained Language Models?**
*Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han and Le Sun*
In recent years, the injection of factual knowledge has been observed to have a significant positive correlation to the downstream task performance of pre-trained language models. However, existing work neither demonstrates that pre-trained models successfully learn the injected factual knowledge nor proves that there is a causal relation between injected factual knowledge and downstream performance improvements. In this paper, we introduce a counterfactual-based analysis framework to explore the causal effects of factual knowledge injection on the performance of language models within pretrain-finetune paradigm. Instead of directly probing the language model or exhaustively enumerating potential confounding factors, we analyze this issue by perturbing the factual knowledge sources at different scales and comparing the performance of pre-trained language models before and after the perturbation. Surprisingly, throughout our experiments, we find that although the knowledge seems to be successfully injected, the correctness of injected knowledge only has a very limited effect on the models' downstream performance. This finding strongly challenges previous assumptions that the injected factual knowledge is the key for language models to achieve performance improvements on downstream tasks in pretrain-finetune paradigm.

14:00-15:30 (East Foyer / Virtual)
**#85 Penalty Decoding: Well Suppress the Self-Reinforcement Effect in Open-Ended Text Generation**
*Wenhong Zhu, Hongkun Hao and Rui Wang*
The decoding algorithm is critical for open-ended text generation, transforming latent representations into coherent and meaningful outputs. This paper investigates the self-reinforcement effect in text generation and the effectiveness of a repetition penalty to mitigate it. However, determining the optimal repetition penalty value is challenging. To tackle this, we propose a forgetting mechanism that disregards distant tokens, reducing the burden of penalty selection. In addition, we introduce a length penalty to address overly short sentences caused by excessive penalties. Our penalty decoding approach incorporating three strategies helps resolve issues with sampling methods deviating from factual information. Experimental results demonstrate the efficacy of our approach in generating high-quality sentences resembling human output.

14:00-15:30 (East Foyer / Virtual)
**#86 Fidelity-Enriched Contrastive Search: Reconciling the Faithfulness-Diversity Trade-Off in Text Generation**
*Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen and Chung-Chi Chen*
In this paper, we address the hallucination problem commonly found in natural language generation tasks. Language models often generate fluent and convincing content but can lack consistency with the provided source, resulting in potential inaccuracies. We propose a new decoding method called Fidelity-Enriched Contrastive Search (FECS), which augments the contrastive search framework with context-aware regularization terms. FECS promotes tokens that are semantically similar to the provided source while penalizing repetitiveness in the generated text. We demonstrate its effectiveness across two tasks prone to hallucination: abstractive summarization and dialogue generation. Results show that FECS consistently enhances faithfulness across various language model sizes while maintaining output diversity comparable to well-performing decoding algorithms.

14:00-15:30 (East Foyer / Virtual)
**#87 Specialist or Generalist? Instruction Tuning for Specific NLP Tasks**
*Chufan Shi, Yixuan Su, Cheng Yang, Yujiu Yang and Deng Cai*
The potential of large language models (LLMs) to simultaneously perform a wide range of natural language processing (NLP) tasks has been the subject of extensive research. Although instruction tuning has proven to be a data-efficient method for transforming LLMs into such generalist models, their performance still lags behind specialist models trained exclusively for specific tasks. In this paper, we investigate whether incorporating broadcoverage generalist instruction tuning can contribute to building a specialist model. We hypothesize that its efficacy depends on task specificity and skill requirements. Our experiments assess four target tasks with distinct coverage levels, revealing that integrating generalist instruction tuning consistently enhances model performance when the task coverage is broad. The effect is particularly pronounced when the amount of task-specific training data is limited. Further investigation into three target tasks focusing on different capabilities demonstrates that generalist instruction tuning improves understanding and reasoning abilities. However, for tasks requiring factual knowledge, generalist data containing hallucinatory information may negatively affect the model's performance. Overall, our work provides a systematic guide for developing specialist models with general instruction tuning.

14:00-15:30 (East Foyer / Virtual)
**#88 Challenges in Context-Aware Neural Machine Translation**
*Linghao Jin, Jacqueline He, Jonathan May and Xuezhe Ma*
Context-aware neural machine translation, a paradigm that involves leveraging information beyond sentence-level context to resolve intersentential discourse dependencies and improve document-level translation quality, has given rise to a number of recent techniques. However, despite well-reasoned intuitions, most context-aware translation models show only modest improvements over sentence-level systems. In this work, we investigate and present several core challenges that impede progress within the field, relating to discourse phenomena, context usage, model architectures, and document-level evaluation. To address these problems, we propose a more realistic setting for document-level translation, called paragraph-to-paragraph (PARA2PARA) translation, and collect a new dataset of Chinese-English novels to promote future research.

14:00-15:30 (East Foyer / Virtual)
**#89 SKD-NER: Continual Named Entity Recognition via Span-based Knowledge Distillation with Reinforcement Learning**
*Yi Chen and Liang He*
Continual learning for named entity recognition (CL-NER) aims to enable models to continuously learn new entity types while retaining the ability to recognize previously learned ones. However, the current strategies fall short of effectively addressing the catastrophic forgetting of previously learned entity types. To tackle this issue, we propose the SKD-NER model, an efficient continual learning NER model based on the span-based approach, which innovatively incorporates reinforcement learning strategies to enhance the model's ability against catastrophic forgetting. Specifically, we leverage knowledge distillation (KD) to retain memory and employ reinforcement learning strategies during the KD process to optimize the soft labeling and distillation losses generated by the teacher model to effectively prevent catastrophic forgetting during continual learning. This approach effectively prevents or mitigates catastrophic forgetting during continuous learning, allowing the model to retain previously learned knowledge while acquiring new knowledge. Our experiments on two benchmark datasets demonstrate that our model significantly improves the performance of the CL-NER task, outperforming state-of-the-art methods.

14:00-15:30 (East Foyer / Virtual)
**#90 Lifelong Sequence Generation with Dynamic Module Expansion and Adaptation**
*Chengwei Qin, Chen Chen and Shafiq Joty*
Lifelong sequence generation (LSG), a problem in continual learning, aims to continually train a model on a sequence of generation tasks to learn constantly emerging new generation patterns while avoiding the forgetting of previous knowledge. Existing LSG methods mainly focus on maintaining old knowledge while paying little attention to knowledge transfer across tasks. In contrast, humans can better learn new tasks by leveraging previously acquired knowledge from similar tasks. Inspired by the learning paradigm of humans, we propose Dynamic Module Expansion and Adaptation (DMEA), which enables the model to dynamically determine the architecture for acquiring new knowledge based on task correlation and select the most similar previous tasks to facilitate adaptation to new tasks. In addition, as the learning process can easily be biased towards the current task which might cause more severe forgetting of previously learned knowledge, we propose dynamic gradient scaling to balance the learning of the current task and replayed tasks. With extensive experiments, we demonstrate that DMEA can consistently outperform existing methods in different LSG settings.

14:00-15:30 (East Foyer / Virtual)
**#91 FedTherapist: Mental Health Monitoring with User-Generated Linguistic Expressions on Smartphones via Federated Learning**
*Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho D. Choi and Sung-Ju Lee*
Psychiatrists diagnose mental disorders via the linguistic use of patients. Still, due to data privacy, existing passive mental health monitoring systems use alternative features such as activity, app usage, and location via mobile devices. We propose FedTherapist, a mobile mental health monitoring system that utilizes continuous speech and keyboard input in a privacy-preserving way via federated learning. We explore multiple model designs by comparing their performance and overhead for FedTherapist to overcome the complex nature of on-device language model training on smartphones. We further propose a Context-Aware Language Learning (CALL) methodology to effectively utilize smartphones' large and noisy text for mental health signal sensing. Our IRB-approved evaluation of the prediction of self-reported depression, stress, anxiety, and mood from 46 participants shows higher accuracy of FedTherapist compared with the performance with non-language features, achieving 0.15 AUROC improvement and 8.21% MAE reduction.

14:00-15:30 (East Foyer / Virtual)
**#92 Program Translation via Code Distillation**
*Yufan Huang, Mengnan Qi, Yongqiang Yao, Maoquan Wang, Bin Gu, Colin Clement and Neel Sundaresan*
Software version migration and program translation are an important and costly part of the lifecycle of large codebases. Traditional machine

translation relies on parallel corpora for supervised translation, which is not feasible for program translation due to a dearth of aligned data. Recent unsupervised neural machine translation techniques have overcome data limitations by included techniques such as back translation and low level compiler intermediate representations (IR). These methods face significant challenges due to the noise in code snippet alignment and the diversity of IRs respectively. In this paper we propose a novel model called Code Distillation (CoDist) whereby we capture the semantic and structural equivalence of code in a language agnostic intermediate representation. Distilled code serves as a translation pivot for any programming language, leading by construction to parallel corpora which scale to all available source code by simply applying the distillation compiler. We demonstrate that our approach achieves state-of-the-art performance on CodeXGLUE and TransCoder GeeksForGeeks translation benchmarks, with an average absolute increase of 12.7% on the TransCoder GeeksforGeeks translation benchmark compare to TransCoder-ST.

14:00-15:30 (East Foyer / Virtual)
**#93 Hi-ArG: Exploring the Integration of Hierarchical Argumentation Graphs in Language Pretraining**
*Jingcong Liang, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang and Zhongyu Wei*
The knowledge graph is a structure to store and represent knowledge, and recent studies have discussed its capability to assist language models for various applications. Some variations of knowledge graphs aim to record arguments and their relations for computational argumentation tasks. However, many must simplify semantic types to fit specific schemas, thus losing flexibility and expression ability. In this paper, we propose the **Hi**erarchical **Ar**gumentation **G**raph (Hi-ArG), a new structure to organize arguments. We also introduce two approaches to exploit Hi-ArG, including a text-graph multi-modal model GreaseArG and a new pre-training framework augmented with graph information. Experiments on two argumentation tasks have shown that after further pre-training and fine-tuning, GreaseArG supersedes same-scale language models on these tasks, while incorporating graph information during further pre-training can also improve the performance of vanilla language models. Code for this paper is available at <https://github.com/ljcleo/Hi-ArG>.

14:00-15:30 (East Foyer / Virtual)
**#94 Privacy Implications of Retrieval-Based Language Models**
*Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li and Danqi Chen*
Retrieval-based language models (LMs) have demonstrated improved interpretability, factuality, and adaptability compared to their parametric counterparts by incorporating retrieved text from external datastores. While it is well known that parametric models are prone to leaking private data, it remains unclear how the addition of a retrieval datastore impacts model privacy. In this work, we present the first study of privacy risks in retrieval-based LMs, particularly $k$NN-LMs. Our goal is to explore the optimal design and training procedure in domains where privacy is of concern, aiming to strike a balance between utility and privacy. Crucially, we find that $k$NN-LMs are more susceptible to leaking private information from their private datastore than parametric models. We further explore mitigations of privacy risks: When privacy information is targeted and readily detected in the text, we find that a simple sanitization step would eliminate the risks while decoupling query and key encoders achieves an even better utility-privacy trade-off. Otherwise, we consider strategies of mixing public and private data in both datastore and encoder training. While these methods offer modest improvements, they leave considerable room for future work. Together, our findings provide insights for practitioners to better understand and mitigate privacy risks in retrieval-based LMs.

14:00-15:30 (East Foyer / Virtual)
**#95 Addressing NER Annotation Noises with Uncertainty-Guided Tree-Structured CRFs**
*Jian Liu, Weichang Liu, Yufeng Chen, Jinan Xu and Zhe Zhao*
Real-world named entity recognition (NER) datasets are notorious for their noisy nature, attributed to annotation errors, inconsistencies, and subjective interpretations. Such noises present a substantial challenge for traditional supervised learning methods. In this paper, we present a new and unified approach to tackle annotation noises for NER. Our method considers NER as a constituency tree parsing problem, utilizing a tree-structured Conditional Random Fields (CRFs) with uncertainty evaluation for integration. Through extensive experiments conducted on four real-world datasets, we demonstrate the effectiveness of our model in addressing both partial and incorrect annotation errors. Remarkably, our model exhibits superb performance even in extreme scenarios with 90% annotation noise.

14:00-15:30 (East Foyer / Virtual)
**#96 Learning from Mistakes via Cooperative Study Assistant for Large Language Models**
*Danqing Wang and Lei Li*
Large language models (LLMs) have demonstrated their potential to refine their generation based on their own feedback. However, the feedback from LLM itself is often inaccurate, thereby limiting its benefits. In this paper, we propose Study Assistant for Large LAnguage Model (SALAM), a novel framework with an auxiliary agent to assist the main LLM in learning from mistakes through interactive cooperation. In the gathering phase, the student assistant agent probes the main LLM, analyzes its errors, and collects the interaction in a mistake memory. During the examination phase, the study assistant provides guidelines by retrieving relevant cases to help the main LLM anticipate and avoid similar errors. We first investigate the effectiveness of a general study assistant and then customize it to provide LLM-specific guidance through imitation learning from successful guidance experiences. Our experiments on three LLMs using two challenging frameworks demonstrate that SALAM can significantly boost LLMs by an accuracy margin of up to 6.6 on BBH and 12.6 on BBQ.

14:00-15:30 (East Foyer / Virtual)
**#97 Sentiment Analysis on Streaming User Reviews via Dual-Channel Dynamic Graph Neural Network**
*Xin Zhang, Linhai Zhang and Deyu Zhou*
Sentiment analysis on user reviews has achieved great success thanks to the rapid growth of deep learning techniques. The large number of online streaming reviews also provides the opportunity to model temporal dynamics for users and products on the timeline. However, existing methods model users and products in the real world based on a static assumption and neglect their time-varying characteristics. In this paper, we present DC-DGNN, a dual-channel framework based on a dynamic graph neural network (DGNN) that models temporal user and product dynamics for sentiment analysis. Specifically, a dual-channel text encoder is employed to extract current local and global contexts from review documents for users and products. Moreover, user review streams are integrated into the dynamic graph neural network by treating users and products as nodes and reviews as new edges. Node representations are dynamically updated along with the evolution of the dynamic graph and used for the final score prediction. Experimental results on five real-world datasets demonstrate the superiority of the proposed method.

14:00-15:30 (East Foyer / Virtual)
**#98 Axiomatic Preference Modeling for Longform Question Answering**
*Corby Rosset, Guoqing Zheng, Victor Dibia, Ahmed Hassan Awadallah and Paul N. Bennett*
The remarkable abilities of large language models (LLMs) like ChatGPT and GPT-4 partially stem from the post-training processes involving human preferences encoded within a reward model as part of a Reinforcement Learning from Human Feedback (RLHF) regimen. These reward models (RMs) often lack direct knowledge of why, or under what principles, the preferences annotations were made. In this study, we identify principles that guide RMs to better align with human preferences, and then develop an axiomatic framework to generate a rich variety of preference signals to uphold them. We use these axiomatic signals to train a model for the scoring answers to longform questions. Our approach yields a **Preference Model** with only about 220M parameters that agrees with gold human-annotated preference labels more

often than GPT-4. The contributions of this work include: training a standalone preference model that can score human- and LLM-generated answers on the same scale; developing an axiomatic framework for generating training data pairs tailored to certain principles; and showing that a small amount of axiomatic signals can help small models outperform GPT-4 in preference scoring. We intend to release our axiomatic data and model.

14:00-15:30 (East Foyer / Virtual)
**#99 CLAD-ST: Contrastive Learning with Adversarial Data for Robust Speech Translation**
*Sathish Reddy Indurthi, Shamil Chollampatt, Ravi Agrawal and Marco Turchi*
The cascaded approach continues to be the most popular choice for speech translation (ST). This approach consists of an automatic speech recognition (ASR) model and a machine translation (MT) model that are used in a pipeline to translate speech in one language to text in another language. MT models are often trained on the well-formed text and therefore lack robustness while translating noisy ASR outputs in the cascaded approach, degrading the overall translation quality significantly. We address this robustness problem in downstream MT models by forcing the MT encoder to bring the representations of a noisy input closer to its clean version in the semantic space. This is achieved by introducing a contrastive learning method that leverages adversarial examples in the form of ASR outputs paired with their corresponding human transcripts to optimize the network parameters. In addition, a curriculum learning strategy is then used to stabilize the training by alternating the standard MT log-likelihood loss and the contrastive losses. Our approach achieves significant gains of up to 3 BLEU scores in English-German and English-French speech translation without hurting the translation quality on clean text.

14:00-15:30 (East Foyer / Virtual)
**#100 Towards Example-Based NMT with Multi-Levenshtein Transformers**
*Maxime Bouthors, Josep Crego and François Yvon*
Retrieval-Augmented Machine Translation (RAMT) is attracting growing attention. This is because RAMT not only improves translation metrics, but is also assumed to implement some form of domain adaptation. In this contribution, we study another salient trait of RAMT, its ability to make translation decisions more transparent by allowing users to go back to examples that contributed to these decisions. For this, we propose a novel architecture aiming to increase this transparency. This model adapts a retrieval-augmented version of the Levenshtein Transformer and makes it amenable to simultaneously edit multiple fuzzy matches found in memory. We discuss how to perform training and inference in this model, based on multi-way alignment algorithms and imitation learning. Our experiments show that editing several examples positively impacts translation scores, notably increasing the number of target spans that are copied from existing instances.

14:00-15:30 (East Foyer / Virtual)
**#101 Interventional Rationalization**
*Linan Yue, Qi Liu, Li Wang, Yanqing An, Yichao Du and Zhenya Huang*
Selective rationalizations improve the explainability of neural networks by selecting a subsequence of the input (i.e., rationales) to explain the prediction results. Although existing methods have achieved promising results, they still suffer from adopting the spurious correlations in data (aka., shortcuts) to compose rationales and make predictions. Inspired by the causal theory, in this paper, we develop an interventional rationalization (Inter-RAT) to discover the causal rationales. Specifically, we first analyse the causalities among the input, rationales and results with a structural causal model. Then, we discover spurious correlations between the input and rationales, and between rationales and results, respectively, by identifying the confounder in the causalities. Next, based on the backdoor adjustment, we propose a causal intervention method to remove the spurious correlations between input and rationales. Further, we discuss reasons why spurious correlations between the selected rationales and results exist by analysing the limitations of the sparsity constraint in the rationalization, and employ the causal intervention method to remove these correlations. Extensive experimental results on three real-world datasets clearly validate the effectiveness of our proposed method. The source code of Inter-RAT is available at https://github.com/yuelinan/Codes-of-Inter-RAT.

14:00-15:30 (East Foyer / Virtual)
**#102 Representative Demonstration Selection for In-Context Learning with Two-Stage Determinantal Point Process**
*Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao and Kang Liu*
Although In-Context Learning has proven effective across a broad array of tasks, its efficiency is noticeably influenced by the selection of demonstrations. Existing methods tend to select different demonstrations for each test instance, which is time-consuming and poses limitations in practical scenarios. Therefore, this study aims to address the challenge of selecting a representative subset of in-context demonstrations that can effectively prompt different test instances in a specific task. We propose that this representative subset should be of high quality and diversity. Our empirical analyses confirm that demonstrations that meet these criteria can indeed bolster model performance. To satisfy these criteria, this paper further introduces a two-stage Determinantal Point Process (DPP) method designed to incorporate both quality and diversity in the process of demonstration selection, thereby obtaining representative in-context demonstrations. Through comprehensive experimentation, we have confirmed the efficacy of our proposed method, paving the way for more practical and effective In-Context Learning.

14:00-15:30 (East Foyer / Virtual)
**#103 TacoPrompt: A Collaborative Multi-Task Prompt Learning Method for Self-Supervised Taxonomy Completion**
*Hongyuan Xu, Ciyi Liu, Yuhang Niu, Yunong Chen, Xiangrui Cai, Yanlong Wen and Xiaojie Yuan*
Automatic taxonomy completion aims to attach the emerging concept to an appropriate pair of hypernym and hyponym in the existing taxonomy. Existing methods suffer from the overfitting to leaf-only problem caused by imbalanced leaf and non-leaf samples when training the newly initialized classification head. Besides, they only leverage subtasks, namely attaching the concept to its hypernym or hyponym, as auxiliary supervision for representation learning yet neglect the effects of subtask results on the final prediction. To address the aforementioned limitations, we propose TacoPrompt, a Collaborative Multi-Task Prompt Learning Method for Self-Supervised Taxonomy Completion. First, we perform triplet semantic matching using the prompt learning paradigm to effectively learn non-leaf attachment ability from imbalanced training samples. Second, we design the result context to relate the final prediction to the subtask results by a contextual approach, enhancing prompt-based multi-task learning. Third, we leverage a two-stage retrieval and re-ranking approach to improve the inference efficiency. Experimental results on three datasets show that TacoPrompt achieves state-of-the-art taxonomy completion performance. Codes are available at https://github.com/cyclexu/TacoPrompt.

14:00-15:30 (East Foyer / Virtual)
**#104 Exploring Discourse Structure in Document-level Machine Translation**
*Xinyu Hu and Xiaojun Wan*
Neural machine translation has achieved great success in the past few years with the help of transformer architectures and large-scale bilingual corpora. However, when the source text gradually grows into an entire document, the performance of current methods for document-level machine translation (DocMT) is less satisfactory. Although the context is beneficial to the translation in general, it is difficult for traditional methods to utilize such long-range information. Previous studies on DocMT have concentrated on extra contents such as multiple surrounding sentences and input instances divided by a fixed length. We suppose that they ignore the structure inside the source text, which leads to under-utilization of the context. In this paper, we present a more sound paragraph-to-paragraph translation mode and explore whether discourse structure can improve DocMT. We introduce several methods from different perspectives, among which our RST-Att model with

a multi-granularity attention mechanism based on the RST parsing tree works best. The experiments show that our method indeed utilizes discourse information and performs better than previous work.

14:00-15:30 (East Foyer / Virtual)
**#105 Location-Aware Visual Question Generation with Lightweight Models**
*Nicholas Collin Suwono, Justin Chen, Tun Min Hung, Ting-Hao Kenneth Huang, I-Bin Liao, Yung-Hui Li, Lun-Wei Ku and Shao-Hua Sun*
This work introduces a novel task, location-aware visual question generation (LocaVQG), which aims to generate engaging questions from data relevant to a particular geographical location. Specifically, we represent such location-aware information with surrounding images and a GPS coordinate. To tackle this task, we present a dataset generation pipeline that leverages GPT-4 to produce diverse and sophisticated questions. Then, we aim to learn a lightweight model that can address the LocaVQG task and fit on an edge device, such as a mobile phone. To this end, we propose a method which can reliably generate engaging questions from location-aware information. Our proposed method outperforms baselines regarding human evaluation (e.g., engagement, grounding, coherence) and automatic evaluation metrics (e.g., BERTScore, ROUGE-2). Moreover, we conduct extensive ablation studies to justify our proposed techniques for both generating the dataset and solving the task.

14:00-15:30 (East Foyer / Virtual)
**#106 Open Information Extraction via Chunks**
*Kuicai Dong, Aixin Sun, Jung-jae Kim and Xiaoli Li*
Open Information Extraction (OIE) aims to extract relational tuples from open-domain sentences. Existing OIE systems split a sentence into tokens and recognize token spans as tuple relations and arguments. We instead propose Sentence as Chunk sequence (SaC) and recognize chunk spans as tuple relations and arguments. We argue that SaC has better properties for OIE than sentence as token sequence, and evaluate four choices of chunks (i.e., CoNLL chunks, OIA simple phrases, noun phrases, and spans from SpanOIE). Also, we propose a simple end-to-end BERT-based model, Chunk-OIE, for sentence chunking and tuple extraction on top of SaC. Chunk-OIE achieves state-of-the-art results on multiple OIE datasets, showing that SaC benefits the OIE task.

14:00-15:30 (East Foyer / Virtual)
**#107 Revisiting Source Context in Nearest Neighbor Machine Translation**
*Xuanhong Li, Peng Li and Po Hu*
Nearest neighbor machine translation ($k$NN-MT), which interpolates target token probabilities with estimates derived from additional examples, has achieved significant improvements and attracted extensive interest in recent years. However, existing research does not explicitly consider the source context when retrieving similar examples, potentially leading to suboptimal performance. To address this, we comprehensively revisit the role of source context and propose a simple and effective method for improving neural machine translation via source context enhancement, demonstrating its crucial role in both retrieving superior examples and determining more suitable interpolation coefficients. Furthermore, we reveal that the probability estimation can be further optimized by incorporating a source-aware distance calibration module. Comprehensive experiments show that our proposed approach can be seamlessly integrated with representative $k$NN-MT baselines, resulting in substantial improvements over these strong baselines across a number of settings and domains. Remarkably, these improvements can reach up to 1.6 BLEU points.

14:00-15:30 (East Foyer / Virtual)
**#108 An Empirical Study of Translation Hypothesis Ensembling with Large Language Models**
*António Farinhas, José G. C. de Souza and Andre Martins*
Large language models (LLMs) are becoming a one-fits-many solution, but they sometimes hallucinate or produce unreliable output. In this paper, we investigate how hypothesis ensembling can improve the quality of the generated text for the specific problem of LLM-based machine translation. We experiment with several techniques for ensembling hypotheses produced by LLMs such as ChatGPT, LLaMA, and Alpaca. We provide a comprehensive study along multiple dimensions, including the method to generate hypotheses (multiple prompts, temperature-based sampling, and beam search) and the strategy to produce the final translation (instruction-based, quality-based reranking, and minimum Bayes risk (MBR) decoding). Our results show that MBR decoding is a very effective method, that translation quality can be improved using a small number of samples, and that instruction tuning has a strong impact on the relation between the diversity of the hypotheses and the sampling temperature.

14:00-15:30 (East Foyer / Virtual)
**#109 KCTS: Knowledge-Constrained Tree Search Decoding with Token-Level Hallucination Detection**
*Sehyun Choi, Tianqing Fang, Zhaowei Wang and Yangqiu Song*
Large Language Models (LLMs) have demonstrated remarkable human-level natural language generation capabilities. However, their potential to generate misinformation, often called the *hallucination* problem, poses a significant risk to their deployment. A common approach to address this issue is to retrieve relevant knowledge and fine-tune the LLM with the knowledge in its input. Unfortunately, this method incurs high training costs and may cause catastrophic forgetting for multi-tasking models. To overcome these limitations, we propose a knowledge-constrained decoding method called KCTS (Knowledge-Constrained Tree Search), which guides a frozen LM to generate text aligned with the reference knowledge at each decoding step using a knowledge classifier score and MCTS (Monte-Carlo Tree Search). To adapt the sequence-level knowledge classifier to token-level guidance, we also propose a novel token-level hallucination detection method called RIPA (Reward Inflection Point Approximation). Our empirical results on knowledge-grounded dialogue and abstractive summarization demonstrate the strength of KCTS as a plug-and-play, model-agnostic decoding method that can effectively reduce hallucinations in natural language generation.

14:00-15:30 (East Foyer / Virtual)
**#110 Self-Influence Guided Data Reweighting for Language Model Pre-training**
*Megh Thakkar, Tolga Bolukbasi, Sriram Ganapathy, Shikhar Vashishth, Sarath Chandar and Partha Talukdar*
Language Models (LMs) pre-trained with selfsupervision on large text corpora have become the default starting point for developing models for various NLP tasks. Once the pre-training corpus has been assembled, all data samples in the corpus are treated with equal importance during LM pre-training. However, due to varying levels of relevance and quality of data, equal importance to all the data samples may not be the optimal choice. While data reweighting has been explored in the context of task-specific supervised learning and LM fine-tuning, model-driven reweighting for pretraining data has not been explored. We fill this important gap and propose PRESENCE, a method for jointly reweighting samples by leveraging self-influence (SI) scores as an indicator of sample importance and pre-training. PRESENCE promotes novelty and stability for model pre-training. Through extensive analysis spanning multiple model sizes, datasets, and tasks, we present PRESENCE as an important first step in the research direction of sample reweighting for pre-training language models.

14:00-15:30 (East Foyer / Virtual)
**#111 Dual-Channel Span for Aspect Sentiment Triplet Extraction**
*Pan Li, Ping Li and Kai Zhang*

Aspect Sentiment Triplet Extraction (ASTE) is one of the compound tasks of fine-grained aspect-based sentiment analysis (ABSA), aiming at extracting the triplets of aspect terms, corresponding opinion terms and the associated sentiment orientation. Recent efforts in exploiting span-level semantic interaction shown superior performance on ASTE task. However, most of the existing span-based approaches suffer from enumerating all possible spans, since it can introduce too much noise in sentiment triplet extraction. To ease this burden, we propose a dual-channel span generation method to coherently constrain the search space of span candidates. Specifically, we leverage the syntactic relations among aspect/opinion terms and the associated part-of-speech characteristics in those terms to generate span candidates, which reduces span enumeration by nearly half. Besides, feature representations are learned from syntactic and part-of-speech correlation among terms, which renders span representation fruitful linguistic information. Extensive experiments on two versions of public datasets demonstrate both the effectiveness of our design and the superiority on ASTE/ATE/OTE tasks.

14:00-15:30 (East Foyer / Virtual)
**#112 Adaptive Policy with Wait-k Model for Simultaneous Translation**
*Libo Zhao, Kai Fan, Wei Luo, Wu Jing, Shushu Wang, Ziqian Zeng and Zhongqiang Huang*
Simultaneous machine translation (SiMT) requires a robust read/write policy in conjunction with a high-quality translation model. Traditional methods rely on either a fixed wait-k policy coupled with a standalone wait-k translation model, or an adaptive policy jointly trained with the translation model. In this study, we propose a more flexible approach by decoupling the adaptive policy model from the translation model. Our motivation stems from the observation that a standalone multi-path wait-k model performs competitively with adaptive policies utilized in state-of-the-art SiMT approaches. Specifically, we introduce DaP, a divergence-based adaptive policy, that makes read/write decisions for any translation model based on the potential divergence in translation distributions resulting from future information. DaP extends a frozen wait-k model with lightweight parameters, and is both memory and computation efficient. Experimental results across various benchmarks demonstrate that our approach offers an improved trade-off between translation accuracy and latency, outperforming strong baselines.

14:00-15:30 (East Foyer / Virtual)
**#113 A Training-Free Debiasing Framework with Counterfactual Reasoning for Conversational Emotion Detection**
*Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong and Ruifeng Xu*
Unintended dataset biases typically exist in existing Emotion Recognition in Conversations (ERC) datasets, including label bias, where models favor the majority class due to imbalanced training data, as well as the speaker and neutral word bias, where models make unfair predictions because of excessive correlations between specific neutral words or speakers and classes. However, previous studies in ERC generally focus on capturing context-sensitive and speaker-sensitive dependencies, ignoring the unintended dataset biases of data, which hampers the generalization and fairness in ERC. To address this issue, we propose a Training-Free Debiasing framework (TFD) that operates during prediction without additional training. To ensure compatibility with various ERC models, it does not balance data or modify the model structure. Instead, TFD extracts biases from the model by generating counterfactual utterances and contexts and mitigates them using simple yet empirically robust element-wise subtraction operations. Extensive experiments on three public datasets demonstrate that TFD effectively improves generalization ability and fairness across different ERC models.

14:00-15:30 (East Foyer / Virtual)
**#114 Lost in Translation, Found in Spans: Identifying Claims in Multilingual Social Media**
*Shubham Mittal, Megha Sundriyal and Preslav Nakov*
Claim span identification (CSI) is an important step in fact-checking pipelines, aiming to identify text segments that contain a check-worthy claim or assertion in a social media post. Despite its importance to journalists and human fact-checkers, it remains a severely understudied problem, and the scarce research on this topic so far has only focused on English. Here we aim to bridge this gap by creating a novel dataset, X-CLAIM, consisting of 7K real-world claims collected from numerous social media platforms in five Indian languages and English. We report strong baselines with state-of-the-art encoder-only language models (e.g., XLM-R) and we demonstrate the benefits of training on multiple languages over alternative cross-lingual transfer methods such as zero-shot transfer, or training on translated data, from a high-resource language such as English. We evaluate generative large language models from the GPT series using prompting methods on the X-CLAIM dataset and we find that they underperform the smaller encoder-only language models for low-resource languages.

14:00-15:30 (East Foyer / Virtual)
**#115 Stance Detection on Social Media with Background Knowledge**
*Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang and Ruifeng Xu*
Identifying users' stances regarding specific targets/topics is a significant route to learning public opinion from social media platforms. Most existing studies of stance detection strive to learn stance information about specific targets from the context, in order to determine the user's stance on the target. However, in real-world scenarios, we usually have a certain understanding of a target when we express our stance on it. In this paper, we investigate stance detection from a novel perspective, where the background knowledge of the targets is taken into account for better stance detection. To be specific, we categorize background knowledge into two categories: episodic knowledge and discourse knowledge, and propose a novel Knowledge-Augmented Stance Detection (KASD) framework. For episodic knowledge, we devise a heuristic retrieval algorithm based on the topic to retrieve the Wikipedia documents relevant to the sample. Further, we construct a prompt for ChatGPT to filter the Wikipedia documents to derive episodic knowledge. For discourse knowledge, we construct a prompt for ChatGPT to paraphrase the hashtags, references, etc., in the sample, thereby injecting discourse knowledge into the sample. Experimental results on four benchmark datasets demonstrate that our KASD achieves state-of-the-art performance in in-target and zero-shot stance detection.

14:00-15:30 (East Foyer / Virtual)
**#116 Text Rendering Strategies for Pixel Language Models**
*Jonas F. Lotz, Elizabeth Salesky, Phillip Rust and Desmond Elliott*
Pixel-based language models process text rendered as images, which allows them to handle any script, making them a promising approach to open vocabulary language modelling. However, recent approaches use text renderers that produce a large set of almost-equivalent input patches, which may prove sub-optimal for downstream tasks, due to redundancy in the input representations. In this paper, we investigate four approaches to rendering text in the PIXEL model (Rust et al., 2023), and find that simple character bigram rendering brings improved performance on sentence-level tasks without compromising performance on token-level or multilingual tasks. This new rendering strategy also makes it possible to train a more compact model with only 22M parameters that performs on par with the original 86M parameter model. Our analyses show that character bigram rendering leads to a consistently better model but with an anisotropic patch embedding space, driven by a patch frequency bias, highlighting the connections between image patch- and tokenization-based language models.

14:00-15:30 (East Foyer / Virtual)
**#117 MT2: Towards a Multi-Task Machine Translation Model with Translation-Specific In-Context Learning**
*Chunyou Li, Mingtong Liu, Hongxiao Zhang, Yufeng Chen, Jinan Xu and Ming Zhou*
Sentence-level translation, document-level translation, translation memory, and terminology constrained translation play an important role in machine translation. Most of the previous work uses separate models or methods to solve these tasks, which is not conducive to knowledge transfer of different tasks and increases the complexity of system construction. In this work, we explore the potential of pre-trained language

model in machine translation tasks and propose a Multi-Task Machine Translation (MT2) model to integrate these translation tasks. We design a novel translation-specific In-Context Learning (ICL) paradigm for model training, in which all of the translation tasks can be modeled as context-learning tasks that integrate contextual information for performance improvement. Specifically, we propose a retrieval and alignment method to obtain a large scale context-enhancement training data, then we train the model in an in-context learning manner. Furthermore, we adopt two context-dependent training strategies to encourage the model to better understand and utilize contextual information for translation. Extensive experiments on translation memory, terminology constrained translation, document-level translation, and few-shot domain-adaptation tasks demonstrate the superior performance of our model, verifying the effectiveness of our proposed approach.

14:00-15:30 (East Foyer / Virtual)
**#118 Gradient-based Gradual Pruning for Language-Specific Multilingual Neural Machine Translation**
*Dan He, Minh-Quang Pham, Thanh-Le Ha and Marco Turchi*
Multilingual neural machine translation (MNMT) offers the convenience of translating between multiple languages with a single model. However, MNMT often suffers from performance degradation in high-resource languages compared to bilingual counterparts. This degradation is commonly attributed to parameter interference, which occurs when parameters are fully shared across all language pairs. In this work, to tackle this issue we propose a gradient-based gradual pruning technique for MNMT. Our approach aims to identify an optimal sub-network for each language pair within the multilingual model by leveraging gradient-based information as pruning criterion and gradually increasing the pruning ratio as schedule. Our approach allows for partial parameter sharing across language pairs to alleviate interference, and each pair preserves its unique parameters to capture language-specific information. Comprehensive experiments on IWSLT and WMT datasets show that our approach yields a notable performance gain on both datasets.

14:00-15:30 (East Foyer / Virtual)
**#119 ScdNER: Span-Based Consistency-Aware Document-Level Named Entity Recognition**
*Ying Wei and Qi Li*
Document-level NER approaches use global information via word-based key-value memory for accurate and consistent predictions. However, such global information on word level can introduce noise when the same word appears in different token sequences and has different labels. This work proposes a two-stage document-level NER model, ScdNER, for more accurate and consistent predictions via adaptive span-level global feature fusion. In the first stage, ScdNER trains a binary classifier to predict if a token sequence is an entity with a probability. Via a span-based key-value memory, the probabilities are further used to obtain the entity's global features with reduced impact of non-entity sequences. The second stage predicts the entity types using a gate mechanism to balance its local and global information, leading to adaptive global feature fusion. Experiments on benchmark datasets from scientific, biomedical, and general domains show the effectiveness of the proposed methods.

14:00-15:30 (East Foyer / Virtual)
**#120 The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models**
*Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson and Richard Johansson*
Large Language Models (LLMs) make natural interfaces to factual knowledge, but their usefulness is limited by their tendency to deliver inconsistent answers to semantically equivalent questions. For example, a model might supply the answer "Edinburgh" to "Anne Redpath passed away in X." and "London" to "Anne Redpath's life ended in X." In this work, we identify potential causes of inconsistency and evaluate the effectiveness of two mitigation strategies: up-scaling and augmenting the LM with a passage retrieval database. Our results on the LLaMA and Atlas models show that both strategies reduce inconsistency but that retrieval augmentation is considerably more efficient. We further consider and disentangle the consistency contributions of different components of Atlas. For all LMs evaluated we find that syntactical form and task artifacts impact consistency. Taken together, our results provide a better understanding of the factors affecting the factual consistency of language models.

14:00-15:30 (East Foyer / Virtual)
**#121 Contextual Interaction for Argument Post Quality Assessment**
*Yiran Wang, Xuanang Chen, Ben He and Le Sun*
Recently, there has been an increased emphasis on assessing the quality of natural language arguments. Existing approaches primarily focus on evaluating the quality of individual argument posts. However, they often fall short when it comes to effectively distinguishing arguments that possess a narrow quality margin. To address this limitation, this paper delves into two alternative methods for modeling the relative quality of different arguments. These approaches include: 1) Supervised contrastive learning that captures the intricate interactions between arguments. By incorporating this approach, we aim to enhance the assessment of argument quality by effectively distinguishing between arguments with subtle differences in quality. 2) Large language models (LLMs) with in-context examples that harness the power of LLMs and enrich them with in-context examples. Through extensive evaluation and analysis on the publicly available IBM-Rank-30k dataset, we demonstrate the superiority of our contrastive argument quality assessment approach over state-of-the-art baselines. On the other hand, while LLMs with in-context examples showcase a commendable ability to identify high-quality argument posts, they exhibit relatively limited efficacy in discerning between argument posts with a narrow quality gap.

14:00-15:30 (East Foyer / Virtual)
**#122 A Deeper (Autoregressive) Approach to Non-Convergent Discourse Parsing**
*Oren Tsur and Yoav Tulpan*
Online social platforms provide a bustling arena for information-sharing and for multi-party discussions. Various frameworks for dialogic discourse parsing were developed and used for the processing of discussions and for predicting the productivity of a dialogue. However, most of these frameworks are not suitable for the analysis of contentious discussions that are commonplace in many online platforms. A novel multi-label scheme for contentious dialog parsing was recently introduced by Zakharov et al. (2021). While the schema is well developed, the computational approach they provide is both naive and inefficient, as a different model (architecture) using a different representation of the input, is trained for each of the 31 tags in the annotation scheme. Moreover, all their models assume full knowledge of label collocations and context, which is unlikely in any realistic setting. In this work, we present a unified model for Non-Convergent Discourse Parsing that does not require any additional input other than the previous dialog utterances. We fine-tuned a RoBERTa backbone, combining embeddings of the utterance, the context and the labels through GRN layers and an asymmetric loss function. Overall, our model achieves results comparable with SOTA, without using label collocation and without training a unique architecture/model for each label. Our proposed architecture makes the labeling feasible at large scale, promoting the development of tools that deepen our understanding of discourse dynamics.

14:00-15:30 (East Foyer / Virtual)
**#123 Predict the Future from the Past? On the Temporal Data Distribution Shift in Financial Sentiment Classifications**
*Yue Guo, Chenxi Hu and Yi Yang*
Temporal data distribution shift is prevalent in the financial text. How can a financial sentiment analysis system be trained in a volatile market environment that can accurately infer sentiment and be robust to temporal data distribution shifts? In this paper, we conduct an empirical study on the financial sentiment analysis system under temporal data distribution shifts using a real-world financial social media dataset that spans

three years. We find that the fine-tuned models suffer from general performance degradation in the presence of temporal distribution shifts. Furthermore, motivated by the unique temporal nature of the financial text, we propose a novel method that combines out-of-distribution detection with time series modeling for temporal financial sentiment analysis. Experimental results show that the proposed method enhances the model's capability to adapt to evolving temporal shifts in a volatile financial market.

14:00-15:30 (East Foyer / Virtual)
**#124 KEPL: Knowledge Enhanced Prompt Learning for Chinese Hypernym-Hyponym Extraction**
*Ningchen Ma, Dong Wang, Hongyun Bao, Lei He and Suncong Zheng*
Modeling hypernym-hyponym ("is-a") relations is very important for many natural language processing (NLP) tasks, such as classification, natural language inference and relation extraction. Existing work on is-a relation extraction is mostly in the English language environment. Due to the flexibility of language expression and the lack of high-quality Chinese annotation datasets, it is still a challenge to accurately identify such relations from Chinese unstructured texts. To tackle this problem, we propose a Knowledge Enhanced Prompt Learning (KEPL) method for Chinese hypernym-hyponym relation extraction. Our model uses the Hearst-like patterns as the prior knowledge. By exploiting a Dynamic Adaptor Architecture to select the matching pattern for the text into prompt, our model embeds patterns and text simultaneously. Additionally, we construct a Chinese hypernym-hyponym relation extraction dataset, which contains three typical scenarios, as baike, news and We-media. The experimental results on the dataset demonstrate the efficiency and effectiveness of our proposed model.

14:00-15:30 (East Foyer / Virtual)
**#125 Controllable Contrastive Generation for Multilingual Biomedical Entity Linking**
*Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu and Yang Xiang*
Multilingual biomedical entity linking (MBEL) aims to map language-specific mentions in the biomedical text to standardized concepts in a multilingual knowledge base (KB) such as Unified Medical Language System (UMLS). In this paper, we propose Con2GEN, a prompt-based controllable contrastive generation framework for MBEL, which summarizes multidimensional information of the UMLS concept mentioned in biomedical text into a natural sentence following a predefined template. Instead of tackling the MBEL problem with a discriminative classifier, we formulate it as a sequence-to-sequence generation task, which better exploits the shared dependencies between source mentions and target entities. Moreover, Con2GEN matches against UMLS concepts in as many languages and types as possible, hence facilitating cross-information disambiguation. Extensive experiments show that our model achieves promising performance improvements compared with several state-of-the-art techniques on the XL-BEL and the Mantra GSC datasets spanning 12 typologically diverse languages.

14:00-15:30 (East Foyer / Virtual)
**#126 Architectural Sweet Spots for Modeling Human Label Variation by the Example of Argument Quality: It's Best to Relate Perspectives!**
*Philipp Heinisch, Matthias Orlikowski, Julia Romberg and Philipp Cimiano*
Many annotation tasks in natural language processing are highly subjective in that there can be different valid and justified perspectives on what is a proper label for a given example. This also applies to the judgment of argument quality, where the assignment of a single ground truth is often questionable. At the same time, there are generally accepted concepts behind argumentation that form a common ground. To best represent the interplay of individual and shared perspectives, we consider a continuum of approaches ranging from models that fully aggregate perspectives into a majority label to "share nothing"-architectures in which each annotator is considered in isolation from all other annotators. In between these extremes, inspired by models used in the field of recommender systems, we investigate the extent to which architectures that predict labels for single annotators but include layers that model the relations between different annotators are beneficial. By means of two tasks of argument quality classification (argument concreteness and validity/novelty of conclusions), we show that recommender architectures increase the averaged annotator-individual F1-scores up to 43% over a majority-label model. Our findings indicate that approaches to subjectivity can benefit from relating individual perspectives.

14:00-15:30 (East Foyer / Virtual)
**#127 Failures Pave the Way: Enhancing Large Language Models through Tuning-free Rule Accumulation**
*Zeyuan Yang, Peng Li and Yang Liu*
Large Language Models (LLMs) have showcased impressive performance. However, due to their inability to capture relationships among samples, these frozen LLMs inevitably keep repeating similar mistakes. In this work, we propose our Tuning-free Rule Accumulation (TRAN) framework, which guides LLMs in improving their performance by learning from previous mistakes. Considering data arrives sequentially, LLMs gradually accumulate rules from incorrect cases, forming a rule collection. These rules are then utilized by the LLMs to avoid making similar mistakes when processing subsequent inputs. Moreover, the rules remain independent of the primary prompts, seamlessly complementing prompt design strategies. Experimentally, we show that TRAN improves over recent baselines by a large margin.

14:00-15:30 (East Foyer / Virtual)
**#128 COVID-19 Vaccine Misinformation in Middle Income Countries**
*Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika J. Wirtz, Traci Hong and Derry Wijaya*
This paper introduces a multilingual dataset of COVID-19 vaccine misinformation, consisting of annotated tweets from three middle-income countries: Brazil, Indonesia, and Nigeria. The expertly curated dataset includes annotations for 5,952 tweets, assessing their relevance to COVID-19 vaccines, presence of misinformation, and the themes of the misinformation. To address challenges posed by domain specificity, the low-resource setting, and data imbalance, we adopt two approaches for developing COVID-19 vaccine misinformation detection models: domain-specific pre-training and text augmentation using a large language model. Our best misinformation detection models demonstrate improvements ranging from 2.7 to 15.9 percentage points in macro F1-score compared to the baseline models. Additionally, we apply our misinformation detection models in a large-scale study of 19 million unlabeled tweets from the three countries between 2020 and 2022, showcasing the practical application of our dataset and models for detecting and analyzing vaccine misinformation in multiple countries and languages. Our analysis indicates that percentage changes in the number of new COVID-19 cases are positively associated with COVID-19 vaccine misinformation rates in a staggered manner for Brazil and Indonesia, and there are significant positive associations between the misinformation rates across the three countries.

14:00-15:30 (East Foyer / Virtual)
**#129 Structural Priming Demonstrates Abstract Grammatical Representations in Multilingual Language Models**
*James Michaelov, Catherine Arnett, Tyler A. Chang and Ben Bergen*
Abstract grammatical knowledge—of parts of speech and grammatical patterns—is key to the capacity for linguistic generalization in humans. But how abstract is grammatical knowledge in large language models? In the human literature, compelling evidence for grammatical abstraction comes from structural priming. A sentence that shares the same grammatical structure as a preceding sentence is processed and produced more readily. Because confounds exist when using stimuli in a single language, evidence of abstraction is even more compelling from crosslingual structural priming, where use of a syntactic structure in one language primes an analogous structure in another language. We measure crosslingual structural priming in large language models, comparing model behavior to human experimental results from eight crosslingual experiments covering six languages, and four monolingual structural priming experiments in three non-English languages. We

find evidence for abstract monolingual and crosslingual grammatical representations in the models that function similarly to those found in humans. These results demonstrate that grammatical representations in multilingual language models are not only similar across languages, but they can causally influence text produced in different languages.

14:00-15:30 (East Foyer / Virtual)
**#130 Noisy Exemplars Make Large Language Models More Robust: A Domain-Agnostic Behavioral Analysis**
*Hongyi Zheng and Abulhair Saparov*
Recent advances in prompt engineering enable large language models (LLMs) to solve multi-hop logical reasoning problems with impressive accuracy. However, there is little existing work investigating the robustness of LLMs with few-shot prompting techniques. Therefore, we introduce a systematic approach to test the robustness of LLMs in multi-hop reasoning tasks via domain-agnostic perturbations. We include perturbations at multiple levels of abstractions (e.g. lexical perturbations such as typos, and semantic perturbations such as the inclusion of intermediate reasoning steps in the questions) to conduct behavioral analysis on the LLMs. Throughout our experiments, we find that models are more sensitive to certain perturbations such as replacing words with their synonyms. We also demonstrate that increasing the proportion of perturbed exemplars in the prompts improves the robustness of few-shot prompting methods.

14:00-15:30 (East Foyer / Virtual)
**#131 ByteSized32: A Corpus and Challenge Task for Generating Task-Specific World Models Expressed as Text Games**
*Ruoyao Wang, Graham Todd, Xingdi Yuan, Ziang Xiao, Marc-Alexandre Côté and Peter Jansen*
In this work we investigate the capacity of language models to generate explicit, interpretable, and interactive world models of scientific and common-sense reasoning tasks. We operationalize this as a task of generating text games, expressed as hundreds of lines of Python code. To facilitate this task, we introduce ByteSized32, a corpus of 32 reasoning-focused text games totalling 20k lines of Python code. We empirically demonstrate that GPT-4 can use these games as templates for single-shot in-context learning, successfully producing runnable games on unseen topics in 28% of cases. When allowed to self-reflect on program errors, game runnability substantially increases to 58%. While evaluating simulation fidelity is labor intensive, we introduce a suite of automated metrics to assess game fidelity, technical validity, adherence to task specifications, and winnability, showing a high-degree of agreement with expert human ratings. We pose this as a challenge task to spur further development at the juncture of world modeling and code generation.

14:00-15:30 (East Foyer / Virtual)
**#132 Bias Neutralization in Non-Parallel Texts: A Cyclic Approach with Auxiliary Guidance**
*Karthic Madanagopal and James Caverlee*
Objectivity is a goal for Wikipedia and many news sites, as well as a guiding principle of many large language models. Indeed, several methods have recently been developed for automatic subjective bias neutralization. These methods, however, typically rely on parallel text for training (i.e. a biased sentence coupled with a non-biased sentence), demonstrate poor transfer to new domains, and can lose important bias-independent context. Toward expanding the reach of bias neutralization, we propose in this paper a new approach called FairBalance. Three of its unique features are: i) a cycle consistent adversarial network enables bias neutralization without the need for parallel text; ii) the model design preserves bias-independent content; and iii) through auxiliary guidance, the model highlights sequences of bias-inducing words, yielding strong results in terms of bias neutralization quality. Extensive experiments demonstrate how FairBalance significantly improves subjective bias neutralization compared to other methods.

14:00-15:30 (East Foyer / Virtual)
**#133 G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment**
*Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu*
The quality of texts generated by natural language generation (NLG) systems is hard to measure automatically. Conventional reference-based metrics, such as BLEU and ROUGE, have been shown to have relatively low correlation with human judgments, especially for tasks that require creativity and diversity. Recent studies suggest using large language models (LLMs) as reference-free metrics for NLG evaluation, which have the benefit of being applicable to new tasks that lack human references. However, these LLM-based evaluators still have lower human correspondence than medium-size neural evaluators. In this work, we present G-Eval, a framework of using large language models with chain-of-thoughts (CoT) and a form-filling paradigm, to assess the quality of NLG outputs. We experiment with two generation tasks, text summarization and dialogue generation. We show that G-Eval with GPT-4 as the backbone model achieves a Spearman correlation of $0.514$ with human on summarization task, outperforming all previous methods by a large margin. We also propose analysis on the behavior of LLM-based evaluators, and highlight the potential concern of LLM-based evaluators having a bias towards the LLM-generated texts.

14:00-15:30 (East Foyer / Virtual)
**#134 Data Similarity is Not Enough to Explain Language Model Performance**
*Gregory Yauney, Emily Reif and David Mimno*
Large language models achieve high performance on many but not all downstream tasks. The interaction between pretraining data and task data is commonly assumed to determine this variance: a task with data that is more similar to a model's pretraining data is assumed to be easier for that model. We test whether distributional and example-specific similarity measures (embedding-, token- and model-based) correlate with language model performance through a large-scale comparison of the Pile and C4 pretraining datasets with downstream benchmarks. Similarity correlates with performance for multilingual datasets, but in other benchmarks, we surprisingly find that similarity metrics are not correlated with accuracy or even each other. This suggests that the relationship between pretraining data and downstream tasks is more complex than often assumed.

14:00-15:30 (East Foyer / Virtual)
**#135 $k$NN-LM Does Not Improve Open-ended Text Generation**
*Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha and Mohit Iyyer*
In this paper, we study the generation quality of interpolation-based retrieval-augmented language models (LMs). These methods, best exemplified by the $k$NN-LM, interpolate the LM's predicted distribution of the next word with a distribution formed from the most relevant retrievals for a given prefix. While the $k$NN-LM and related methods yield impressive decreases in perplexity, we discover that they do not exhibit corresponding improvements in open-ended generation quality, as measured by both automatic evaluation metrics (e.g., MAUVE) and human evaluations. Digging deeper, we find that interpolating with a retrieval distribution actually increases perplexity compared to a baseline LM for the majority of tokens in the WikiText-103 test set, even though the overall perplexity is lower due to a smaller number of tokens for which perplexity dramatically decreases after interpolation. However, when decoding a long sequence at inference time, significant improvements on this smaller subset of tokens are washed out by slightly worse predictions on most tokens. Furthermore, we discover that the entropy of the retrieval distribution increases faster than that of the base LM as the generated sequence becomes longer, which indicates that retrieval is less reliable when using model-generated text as queries (i.e., is subject to exposure bias). We hope that our analysis spurs future work on improved decoding algorithms and interpolation strategies for retrieval-augmented language models.

14:00-15:30 (East Foyer / Virtual)

**#136 Taxonomy Expansion for Named Entity Recognition**
*Karthikeyan K, Yogarshi Vyas, Jie Ma, Giovanni Paolini, Neha Anna John, Shuai Wang, Yassine Benajiba, Vittorio Castelli, Dan Roth and Miguel Ballesteros*
Training a Named Entity Recognition (NER) model often involves fixing a taxonomy of entity types. However, requirements evolve and we might need the NER model to recognize additional entity types. A simple approach is to re-annotate entire dataset with both existing and additional entity types and then train the model on the re-annotated dataset. However, this is an extremely laborious task. To remedy this, we propose a novel approach called Partial Label Model (PLM) that uses only partially annotated datasets. We experiment with 6 diverse datasets and show that PLM consistently performs better than most other approaches (0.5 - 2.5 F1), including in novel settings for taxonomy expansion not considered in prior work. The gap between PLM and all other approaches is especially large in settings where there is limited data available for the additional entity types (as much as 11 F1), thus suggesting a more cost effective approaches to taxonomy expansion.

14:00-15:30 (East Foyer / Virtual)
**#137 Transcending Scaling Laws with 0.1% Extra Compute**
*Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q. Tran, David So, Siamak Shakeri, Xavier Garcia, Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, Denny Zhou, Donald Metzler, Slav Petrov, Neil Houlsby, Quoc V Le and Mostafa Dehghani*
Scaling language models improves performance but comes with significant computational costs. This paper proposes UL2R, a method that substantially improves existing language models and their scaling curves with a relatively tiny amount of extra compute. The key idea is to continue training a state-of-the-art large language model on a few more steps with UL2's mixture-of-denoiser objective. We show that, with almost negligible extra computational costs and no new sources of data, we are able to substantially improve the scaling properties of large language models on downstream metrics. In this paper, we continue training a baseline language model, PaLM, with ULR2, introducing a new set of models at 8B, 62B, and 540B scale which we call U-PaLM. Impressively, at 540B scale, we show an approximately 2x computational savings rate where U-PaLM achieves the same performance as the final PaLM 540B model at around half its computational budget (i.e., saving $\sim$4.4 million TPUv4 hours). We further show that this improved scaling curve leads to "emergent abilities" on challenging BIG-Bench tasks—for instance, U-PaLM does much better on some tasks or demonstrates better quality at much smaller scale (62B as opposed to 540B). Overall, we show that U-PaLM outperforms PaLM on many few-shot setups, including reasoning tasks with chain-of-thought (e.g., GSM8K), multilingual tasks (MGSM, TydiQA), MMLU and challenging BIG-Bench tasks.

14:00-15:30 (East Foyer / Virtual)
**#138 Recurrent Neural Language Models as Probabilistic Finite-state Automata**
*Anej Svete and Ryan Cotterell*
Studying language models (LMs) in terms of well-understood formalisms allows us to precisely characterize their abilities and limitations. Previous work has investigated the expressive power of recurrent neural network (RNN) LMs in terms of their capacity to recognize unweighted formal languages. However, LMs do not describe unweighted formal languages—rather, they define probability distributions over strings. In this work, we study what classes of such probability distributions RNN LMs can represent, which allows us to make more direct statements about their capabilities. We show that simple RNNs are equivalent to a subclass of probabilistic finite-state automata, and can thus model a strict subset of probability distributions expressible by finite-state models. Furthermore, we study the space complexity of representing finite-state LMs with RNNs. We show that, to represent an arbitrary deterministic finite-state LM with $N$ states over an alphabet $\Sigma$, an RNN requires $\Omega\left(N|\Sigma|\right)$ neurons. These results present a first step towards characterizing the classes of distributions RNN LMs can represent and thus help us understand their capabilities and limitations.

14:00-15:30 (East Foyer / Virtual)
**#139 Bridging Background Knowledge Gaps in Translation with Automatic Explicitation**
*HyoJung Han, Jordan Lee Boyd-Graber and Marine Carpuat*
Translations help people understand content written in another language. However, even correct literal translations do not fulfill that goal when people lack the necessary background to understand them. Professional translators incorporate explicitations to explain the missing context by considering cultural differences between source and target audiences. Despite its potential to help users, NLP research on explicitation is limited because of the dearth of adequate evaluation methods. This work introduces techniques for automatically generating explicitations, motivated by WikiExpl: a dataset that we collect from Wikipedia and annotate with human translators. The resulting explicitations are useful as they help answer questions more accurately in a multilingual question answering framework.

14:00-15:30 (East Foyer / Virtual)
**#140 Dynosaur: A Dynamic Growth Paradigm for Instruction-Tuning Data Curation**
*Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han and Kai-Wei Chang*
Instruction tuning has emerged to enhance the capabilities of large language models (LLMs) to comprehend instructions and generate appropriate responses. Existing methods either manually annotate or employ LLM (e.g., GPT-series) to generate data for instruction tuning. However, they often overlook associating instructions with existing annotated datasets. In this paper, we propose Dynosaur, a dynamic growth paradigm for the automatic curation of instruction-tuning data. Based on the metadata of existing datasets, we use LLMs to automatically construct instruction-tuning data by identifying relevant data fields and generating appropriate instructions. By leveraging the existing annotated datasets, Dynosaur offers several advantages: 1) it reduces the API cost for generating instructions (e.g., it costs less than $12 USD by calling GPT-3.5-turbo for generating 800K instruction tuning samples; 2) it provides high-quality data for instruction tuning (e.g., it performs better than Alpaca and Flan on Super-NI and Longform with comparable data sizes); and 3) it supports the continuous improvement of models by generating instruction-tuning data when a new annotated dataset becomes available. We further investigate a continual learning scheme for learning with the ever-growing instruction-tuning dataset, and demonstrate that replaying tasks with diverse instruction embeddings not only helps mitigate forgetting issues but generalizes to unseen tasks better. Code and data are available at https://github.com/WadeYin9712/Dynosaur.

14:00-15:30 (East Foyer / Virtual)
**#141 Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration**
*Daniel Deutsch, George Foster and Markus Freitag*
Kendall's tau is frequently used to meta-evaluate how well machine translation (MT) evaluation metrics score individual translations. Its focus on pairwise score comparisons is intuitive but raises the question of how ties should be handled, a gray area that has motivated different variants in the literature. We demonstrate that, in settings like modern MT meta-evaluation, existing variants have weaknesses arising from their handling of ties, and in some situations can even be gamed. We propose instead to meta-evaluate metrics with a version of pairwise accuracy that gives metrics credit for correctly predicting ties, in combination with a tie calibration procedure that automatically introduces ties into metric scores, enabling fair comparison between metrics that do and do not predict ties. We argue and provide experimental evidence that these modifications lead to fairer ranking-based assessments of metric performance.

14:00-15:30 (East Foyer / Virtual)
**#142 Oolong: Investigating What Makes Transfer Learning Hard with Controlled Studies**

*Zhengxuan Wu, Alex Tamkin and Isabel Papadimitriou*
When we transfer a pretrained language model to a new language, there are many axes of variation that change at once. To disentangle the impact of different factors like syntactic similarity and vocabulary similarity, we propose a set of *controlled transfer studies*: we systematically transform the language of the GLUE benchmark, altering one axis of crosslingual variation at a time, and then measure the resulting drops in a pretrained model's downstream performance. We find that models can largely recover from syntactic-style shifts, but cannot recover from vocabulary misalignment and embedding matrix re-initialization, even with continued pretraining on 15 million tokens. Moreover, good-quality tokenizers in the transfer language do not make vocabulary alignment easier. Our experiments provide insights into the factors of cross-lingual transfer that researchers should most focus on when designing language transfer scenarios.

14:00-15:30 (East Foyer / Virtual)
**#143 Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback**
*Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn and Christopher D Manning*
A trustworthy real-world prediction system should produce well-calibrated confidence scores; that is, its confidence in an answer should be indicative of the likelihood that the answer is correct, enabling deferral to an expert in cases of low-confidence predictions. Recent studies have shown that unsupervised pre-training produces large language models (LMs) whose conditional probabilities are remarkably well-calibrated. However, the most widely-used LMs are fine-tuned with reinforcement learning from human feedback (RLHF-LMs), and some studies have suggested that RLHF-LMs produce conditional probabilities that are very poorly calibrated. In light of this perceived weakness, we conduct a broad evaluation of methods for extracting confidence scores from RLHF-LMs. For RLHF-LMs such as ChatGPT, GPT-4, and Claude, we find that verbalized confidences emitted as output tokens are typically better-calibrated than the model's conditional probabilities on the TriviaQA, SciQ, and TruthfulQA benchmarks, often reducing the expected calibration error by a relative 50%.

14:00-15:30 (East Foyer / Virtual)
**#144 Preserving Privacy Through Dememorization: An Unlearning Technique For Mitigating Memorization Risks In Language Models**
*Aly M. Kassem, Omar Mahmoud and Sherif Saad*
Large Language models (LLMs) are trained on vast amounts of data, including sensitive information that poses a risk to personal privacy if exposed. LLMs have shown the ability to memorize and reproduce portions of their training data when prompted by adversaries. Prior research has focused on addressing this memorization issue and preventing verbatim replication through techniques like knowledge unlearning and data pre-processing. However, these methods have limitations regarding the number of protected samples, limited privacy types, and potentially lower-quality generative models. To tackle this challenge more effectively, we propose "DeMem," a novel unlearning approach that utilizes an efficient reinforcement learning feedback loop via proximal policy optimization. By fine-tuning the language model with a negative similarity score as a reward signal, we incentivize the LLMs to learn a paraphrasing policy to unlearn the pre-training data. Our experiments demonstrate that DeMem surpasses strong baselines and state-of-the-art methods in terms of its ability to generalize and strike a balance between maintaining privacy and LLM performance.

14:00-15:30 (East Foyer / Virtual)
**#145 Unveiling the Implicit Toxicity in Large Language Models**
*Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai and Minlie Huang*
The open-endedness of large language models (LLMs) combined with their impressive capabilities may lead to new safety issues when being exploited for malicious use. While recent studies primarily focus on probing toxic outputs that can be easily detected with existing toxicity classifiers, we show that LLMs can generate diverse implicit toxic outputs that are exceptionally difficult to detect via simply zero-shot prompting. Moreover, we propose a reinforcement learning (RL) based attacking method to further induce the implicit toxicity in LLMs. Specifically, we optimize the language model with a reward that prefers implicit toxic outputs to explicit toxic and non-toxic ones. Experiments on five widely-adopted toxicity classifiers demonstrate that the attack success rate can be significantly improved through RL fine-tuning. For instance, the RL-finetuned LLaMA-13B model achieves an attack success rate of 90.04% on BAD and 62.85% on Davinci003. Our findings suggest that LLMs pose a significant threat in generating undetectable implicit toxic outputs. We further show that fine-tuning toxicity classifiers on the annotated examples from our attacking method can effectively enhance their ability to detect LLM-generated implicit toxic language.

14:00-15:30 (East Foyer / Virtual)
**#146 Learn and Consolidate: Continual Adaptation for Zero-Shot and Multilingual Neural Machine Translation**
*Kaiyu Huang, Peng Li, Junpeng Liu, Maosong Sun and Yang Liu*
Although existing multilingual neural machine translation (MNMT) models have demonstrated remarkable performance to handle multiple translation directions in a single model and achieved zero-shot translation between language pairs unseen in training, they still suffer from relatively poor translation qualities for some language pairs. A practical scenario is that how to continually update MNMT models for both supervised and zero-shot translations when limited new data arrives. To this end, we propose a two-stage approach that encourages original models to acquire language-agnostic multilingual representations from new data, and preserves the model architecture without introducing parameters. Experimental results and further analysis demonstrate that our method can efficiently improve performance of existing MNMT models in translation directions where they are initially weak, and mitigates the degeneration in the original well-performing translation directions, offering flexibility in the real-world scenario.

14:00-15:30 (East Foyer / Virtual)
**#147 Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models**
*Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt and Sandipan Dandapat*
Temporal reasoning represents a vital component of human communication and understanding, yet remains an underexplored area within the context of Large Language Models (LLMs). Despite LLMs demonstrating significant proficiency in a range of tasks, a comprehensive, large-scale analysis of their temporal reasoning capabilities is missing. Our paper addresses this gap, presenting the first extensive benchmarking of LLMs on temporal reasoning tasks. We critically evaluate 8 different LLMs across 6 datasets using 3 distinct prompting strategies. Additionally, we broaden the scope of our evaluation by including in our analysis 2 Code Generation LMs. Beyond broad benchmarking of models and prompts, we also conduct a fine-grained investigation of performance across different categories of temporal tasks. We further analyze the LLMs on varying temporal aspects, offering insights into their proficiency in understanding and predicting the continuity, sequence, and progression of events over time. Our findings reveal a nuanced depiction of the capabilities and limitations of the models within temporal reasoning, offering a comprehensive reference for future research in this pivotal domain.

14:00-15:30 (East Foyer / Virtual)
**#148 Evaluating Large Language Models on Controlled Generation Tasks**
*Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng and Xuezhe Ma*

While recent studies have looked into the abilities of large language models in various benchmark tasks, including question generation, reading comprehension, multilingual and etc, there have been few studies looking into the controllability of large language models on generation tasks. We present an extensive analysis of various benchmarks including a sentence planning benchmark with different granularities. After comparing large language models against state-of-the-start finetuned smaller models, we present a spectrum showing large language models falling behind, are comparable, or exceed the ability of smaller models. We conclude that *large language models struggle at meeting fine-grained hard constraints*.

14:00-15:30 (East Foyer / Virtual)
**#149 Learning Preference Model for LLMs via Automatic Preference Data Generation**
*Shijia Huang, Jianqiao Zhao, Yanyang Li and Liwei Wang*
Despite the advanced capacities of the state-of-the-art large language models (LLMs), they suffer from issues of hallucination, stereotype, etc. Preference models play an important role in LLM alignment, yet training preference models predominantly rely on human-annotated data. This reliance limits their versatility and scalability. In this paper, we propose learning the preference model for LLMs via automatic preference data generation (AutoPM). Our approach involves both In-Breadth Data Generation, which elicits pairwise preference data from LLMs following the helpful-honest-harmless (HHH) criteria, and In-Depth Data Generation, which enriches the dataset with responses spanning a wide quality range. With HHH-guided preference data, our approach simultaneously enables the LLMs to learn human preferences and align with human values. Quantitative assessments on five benchmark datasets demonstrate the reliability and potential of AutoPM, pointing out a more general and scalable way to improve LLM performance.

14:00-15:30 (East Foyer / Virtual)
**#150 CRaSh: Clustering, Removing, and Sharing Enhance Fine-tuning without Full Large Language Model**
*Kaiyan Zhang, Ning Ding, Biqing Qi, Xuekai Zhu, Xinwei Long and Bowen Zhou*
Instruction tuning has recently been recognized as an effective way of aligning Large Language Models (LLMs) to enhance their generalization ability across various tasks. However, when tuning publicly accessible, centralized LLMs with private instruction data, privacy concerns are inevitable. While direct transfer of parameterized modules between models is a plausible approach to address this, its implications and effectiveness need further exploration. This paper focuses on Offsite-Tuning (OFT), a representative technique that transfers transformer blocks between centralized LLMs and downstream emulators. Given the limited understanding of the underlying mechanism of OFT, we perform an empirical analysis on LLMs from the perspectives of representation and functional similarity. Interestingly, our findings reveal a unique modular structure within the layers of LLMs that appears to emerge as the model size expands. Simultaneously, we note subtle but potentially significant changes in representation and intermediate predictions across the layers. Inspired by these observations, we propose CRaSh, involving Clustering, Removing, and Sharing, a training-free strategy to derive improved emulators from LLMs. CRaSh significantly boosts performance of OFT with billions of parameters. Furthermore, we investigate the optimal solutions yielded by fine-tuning with and without full model through the lens of loss landscape. Our findings demonstrate a linear connectivity among these optima falling over the same basin, thereby highlighting the effectiveness of CRaSh and OFT.

14:00-15:30 (East Foyer / Virtual)
**#151 The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models**
*Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan and Shauli Ravfogel*
Large language models (LLMs) have been shown to possess impressive capabilities, while also raising crucial concerns about the faithfulness of their responses. A primary issue arising in this context is the management of (un)answerable queries by LLMs, which often results in hallucinatory behavior due to overconfidence. In this paper, we explore the behavior of LLMs when presented with (un)answerable queries. We ask: do models *represent* the fact that the question is (un)answerable when generating a hallucinatory answer? Our results show strong indications that such models encode the answerability of an input query, with the representation of the first decoded token often being a strong indicator. These findings shed new light on the spatial organization within the latent representations of LLMs, unveiling previously unexplored facets of these models. Moreover, they pave the way for the development of improved decoding techniques with better adherence to factual generation, particularly in scenarios where query (un)answerability is a concern.

14:00-15:30 (East Foyer / Virtual)
**#152 On the Representational Capacity of Recurrent Neural Language Models**
*Franz Nowak, Anej Svete, Li Du and Ryan Cotterell*
This work investigates the computational expressivity of language models (LMs) based on recurrent neural networks (RNNs). Siegelmann and Sontag (1992) famously showed that RNNs with rational weights and hidden states and unbounded computation time are Turing complete. However, LMs define weightings over strings in addition to just (unweighted) language membership and the analysis of the computational power of RNN LMs (RLMs) should reflect this. We extend the Turing completeness result to the probabilistic case, showing how a rationally weighted RLM with unbounded computation time can simulate any deterministic probabilistic Turing machine (PTM) with rationally weighted transitions. Since, in practice, RLMs work in real-time, processing a symbol at every time step, we treat the above result as an upper bound on the expressivity of RLMs. We also provide a lower bound by showing that under the restriction to real-time computation, such models can simulate deterministic real-time rational PTMs.

14:00-15:30 (East Foyer / Virtual)
**#153 Continual Learning for Multilingual Neural Machine Translation via Dual Importance-based Model Division**
*Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su and Degen Huang*
A persistent goal of multilingual neural machine translation (MNMT) is to continually adapt the model to support new language pairs or improve some current language pairs without accessing the previous training data. To achieve this, the existing methods primarily focus on preventing catastrophic forgetting by making compromises between the original and new language pairs, leading to sub-optimal performance on both translation tasks. To mitigate this problem, we propose a dual importance-based model division method to divide the model parameters into two parts and separately model the translation of the original and new tasks. Specifically, we first remove the parameters that are negligible to the original tasks but essential to the new tasks to obtain a pruned model, which is responsible for the original translation tasks. Then we expand the pruned model with external parameters and fine-tune the newly added parameters with new training data. The whole fine-tuned model will be used for the new translation tasks. Experimental results show that our method can efficiently adapt the original model to various new translation tasks while retaining the performance of the original tasks. Further analyses demonstrate that our method consistently outperforms several strong baselines under different incremental translation scenarios.

14:00-15:30 (East Foyer / Virtual)
**#154 Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text**
*Qi Cao, Takeshi Kojima, Yutaka Matsuo and Yusuke Iwasawa*
While Large Language Models (LLMs) have achieved remarkable performance in many tasks, much about their inner workings remains unclear. In this study, we present novel experimental insights into the resilience of LLMs, particularly GPT-4, when subjected to extensive

character-level permutations. To investigate this, we first propose the Scrambled Bench, a suite designed to measure the capacity of LLMs to handle scrambled input, in terms of both recovering scrambled sentences and answering questions given scrambled context. The experimental results indicate that multiple advanced LLMs demonstrate the capability akin to typoglycemia, a phenomenon where humans can understand the meaning of words even when the letters within those words are scrambled, as long as the first and last letters remain in place. More surprisingly, we found that only GPT-4 nearly flawlessly processes inputs with unnatural errors, a task that poses significant challenges for other LLMs and often even for humans. Specifically, GPT-4 can almost perfectly reconstruct the original sentences from scrambled ones, decreasing the edit distance by 95%, even when all letters within each word are entirely scrambled. It is counter-intuitive that LLMs can exhibit such resilience despite severe disruption to input tokenization caused by scrambled text.

14:00-15:30 (East Foyer / Virtual)
**#155 Adapting Offline Speech Translation Models for Streaming with Future-Aware Distillation and Inference**
*Biao Fu, Minpeng Liao, Kai Fan, Zhongqiang Huang, Boxing Chen, Yidong Chen and Xiaodong Shi*
A popular approach to streaming speech translation is to employ a single offline model with a wait-k policy to support different latency requirements, which is simpler than training multiple online models with different latency constraints. However, there is a mismatch problem in using a model trained with complete utterances for streaming inference with partial input. We demonstrate that speech representations extracted at the end of a streaming input are significantly different from those extracted from a complete utterance. To address this issue, we propose a new approach called Future-Aware Streaming Translation (FAST) that adapts an offline ST model for streaming input. FAST includes a Future-Aware Inference (FAI) strategy that incorporates future context through a trainable masked embedding, and a Future-Aware Distillation (FAD) framework that transfers future context from an approximation of full speech to streaming input. Our experiments on the MuST-C EnDe, EnEs, and EnFr benchmarks show that FAST achieves better trade-offs between translation quality and latency than strong baselines. Extensive analyses suggest that our methods effectively alleviate the aforementioned mismatch problem between offline training and online inference.

14:00-15:30 (East Foyer / Virtual)
**#156 Mitigating Over-Generation for Unsupervised Keyphrase Extraction with Heterogeneous Centrality Detection**
*Mingyang Song, Pengyu Xu, Yi Feng, Huafeng Liu and Liping Jing*
Over-generation errors occur when a keyphrase extraction model correctly determines a candidate keyphrase as a keyphrase because it contains a word that frequently appears in the document but at the same time erroneously outputs other candidates as keyphrases because they contain the same word. To mitigate this issue, we propose a new heterogeneous centrality detection approach (CentralityRank), which extracts keyphrases by simultaneously identifying both implicit and explicit centrality within a heterogeneous graph as the importance score of each candidate. More specifically, CentralityRank detects centrality by taking full advantage of the content within the input document to construct graphs that encompass semantic nodes of varying granularity levels, not limited to just phrases. These additional nodes act as intermediaries between candidate keyphrases, enhancing cross-phrase relations. Furthermore, we introduce a novel adaptive boundary-aware regularization that can leverage the position information of candidate keyphrases, thus influencing the importance of candidate keyphrases. Extensive experimental results demonstrate the superiority of CentralityRank over recent state-of-the-art unsupervised keyphrase extraction baselines across three benchmark datasets.

14:00-15:30 (East Foyer / Virtual)
**#157 A Self-training Framework for Automated Medical Report Generation**
*Siyuan Wang, Zheng Liu and Bo Peng*
Medical report generation, focusing on automatically generating accurate clinical findings from medical images, is an important medical artificial intelligence task. It reduces the workload of physicians in writing reports. Many of the current methods depend heavily on labeled datasets that include a large amount of image-report pairs, but such datasets labeled by physicians are hard to acquire in clinical practice. To this end, in this paper, we introduce a self-training framework named REMOTE (i.e., Revisiting sElf-training for Medical repOrT gEneration) to exploit the unlabeled medical images and a reference-free evaluation metric MedCLIPScore to augment a small-scale medical report generation dataset for training accurate medical report generation model. Experiments and analysis conducted on the MIMIC-CXR and IU-Xray benchmark datasets demonstrate that, our REMOTE framework, using 1% labeled training data, achieves competitive performance with previous fully-supervised models that are trained on entire training data.

14:00-15:30 (East Foyer / Virtual)
**#158 Predict and Use: Harnessing Predicted Gaze to Improve Multimodal Sarcasm Detection**
*Divyank Pratap Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna and Pushpak Bhattacharyya*
Sarcasm is a complex linguistic construct with incongruity at its very core. Detecting sarcasm depends on the actual content spoken and tonality, facial expressions, the context of an utterance, and personal traits like language proficiency and cognitive capabilities. In this paper, we propose the utilization of synthetic gaze data to improve the task performance for multimodal sarcasm detection in a conversational setting. We enrich an existing multimodal conversational dataset, i.e., MUStARD++ with gaze features. With the help of human participants, we collect gaze features for 20% of data instances, and we investigate various methods for gaze feature prediction for the rest of the dataset. We perform extrinsic and intrinsic evaluations to assess the quality of the predicted gaze features. We observe a performance gain of up to 6.6% points by adding a new modality, i.e., collected gaze features. When both collected and predicted data are used, we observe a performance gain of 2.3% points on the complete dataset. Interestingly, with only predicted gaze features, too, we observe a gain in performance (1.9% points). We retain and use the feature prediction model, which maximally correlates with collected gaze features. Our model trained on combining collected and synthetic gaze data achieves SoTA performance on the MUStARD++ dataset. To the best of our knowledge, ours is the first predict-and-use model for sarcasm detection. We publicly release the code, gaze data, and our best models for further research.

14:00-15:30 (East Foyer / Virtual)
**#159 Adapting to the Long Tail: A Meta-Analysis of Transfer Learning Research for Language Understanding Tasks**
*Aakanksha Naik, Jill Lehman and Carolyn Rose*
Natural language understanding (NLU) has made massive progress driven by large benchmarks, but benchmarks often leave a long tail of infrequent phenomena underrepresented. We reflect on the question: have transfer learning methods sufficiently addressed the poor performance of benchmark-trained models on the long tail? We conceptualize the long tail using macro-level dimensions (e.g., underrepresented genres, topics, etc.), and perform a qualitative meta-analysis of 100 representative papers on transfer learning research for NLU. Our analysis asks three questions: (i) Which long tail dimensions do transfer learning studies target? (ii) Which properties of adaptation methods help improve performance on the long tail? (iii) Which methodological gaps have greatest negative impact on long tail performance? Our answers highlight major avenues for future research in transfer learning for the long tail. Lastly, using our meta-analysis framework, we perform a case study comparing the performance of various adaptation methods on clinical narratives, which provides interesting insights that may enable us to make progress along these future avenues.

14:00-15:30 (East Foyer / Virtual)
**#160 Compositional Zero-Shot Domain Transfer with Text-to-Text Models**

*Stephanie L Hyland, Fangyu Liu, Qianchu Liu, Shruthi Bannur, Fernando Pérez-García, Naoto Usuyama, Sheng Zhang, Tristan Naumann, Aditya Nori, Hoifung Poon, Javier Alvarez-Valle and Ozan Oktay*
We propose a novel graph-based approach for semantic parsing that resolves two problems observed in the literature: (1) seq2seq models fail on compositional generalization tasks; (2) previous work using phrase structure parsers cannot cover all the semantic parses observed in treebanks. We prove that both MAP inference and latent tag anchoring (required for weakly-supervised learning) are NP-hard problems. We propose two optimization algorithms based on constraint smoothing and conditional gradient to approximately solve these inference problems. Experimentally, our approach delivers state-of-the-art results on GeoQuery, Scan, and Clevr, both for i.i.d. splits and for splits that test for compositional generalization.

14:00-15:30 (East Foyer / Virtual)
**#161 Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off"**
*Yuchen Lian, Arianna Bisazza and Tessa Verhoef*
Artificial learners often behave differently from human learners in the context of neural agent-based simulations of language emergence and change. A common explanation is the lack of appropriate cognitive biases in these learners. However, it has also been proposed that more naturalistic settings of language learning and use could lead to more human-like results. We investigate this latter account focusing on the word-order/case-marking trade-off, a widely attested language universal that has proven particularly hard to simulate. We propose a new Neural-agent Language Learning and Communication framework (NeLLCom) where pairs of speaking and listening agents first learn a miniature language via supervised learning, and then optimize it for communication via reinforcement learning. Following closely the setup of earlier human experiments, we succeed in replicating the trade-off with the new framework without hard-coding specific biases in the agents. We see this as an essential step towards the investigation of language universals with neural learners.

14:00-15:30 (East Foyer / Virtual)
**#162 Hallucinations in Large Multilingual Translation Models**
*Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo and André F.T. Martins*
Hallucinated translations can severely undermine and raise safety issues when machine translation systems are deployed in the wild. Previous research on the topic focused on small bilingual models trained on high-resource languages, leaving a gap in our understanding of hallucinations in multilingual models across diverse translation scenarios. In this work, we fill this gap by conducting a comprehensive analysis – over 100 language pairs across various resource levels and going beyond English-centric directions – on both the M2M neural machine translation (NMT) models and GPT large language models (LLMs). Among several insights, we highlight that models struggle with hallucinations primarily in low-resource directions and when translating out of English, where, critically, they may reveal toxic patterns that can be traced back to the training data. We also find that LLMs produce qualitatively different hallucinations to those of NMT models. Finally, we show that hallucinations are hard to reverse by merely scaling models trained with the same data. However, employing more diverse models, trained on different data or with different procedures, as fallback systems can improve translation quality and virtually eliminate certain pathologies.

14:00-15:30 (East Foyer / Virtual)
**#163 Removing Backdoors in Pre-trained Models by Regularized Continual Pre-training**
*Biru Zhu, Ganqu Cui, Yangyi Chen, Yujia Qin, Lifan Yuan, Chong Fu, Yangdong Deng, Zhiyuan Liu, Maosong Sun and Ming Gu*
Recent researches reveal that pre-trained models (PTMs) are vulnerable to backdoor attacks before the fine-tuning stage. The attackers can implant transferable task-agnostic backdoors in PTMs, and control model outputs on any downstream task, which poses severe security threats to all downstream applications. Existing backdoor-removal defenses focus on task-specific classification models and they are not suitable for defending PTMs against task-agnostic backdoor attacks. To this end, we propose the first task-agnostic backdoor removal method for PTMs. Based on the selective activation phenomenon in backdoored PTMs, we design a simple and effective backdoor eraser, which continually pre-trains the backdoored PTMs with a regularization term in an end-to-end approach. The regularization term removes backdoor functionalities from PTMs while the continual pre-training maintains the normal functionalities of PTMs. We conduct extensive experiments on pre-trained models across different modalities and architectures. The experimental results show that our method can effectively remove backdoors inside PTMs and preserve benign functionalities of PTMs with a few downstream-task-irrelevant auxiliary data, e.g., unlabeled plain texts. The average attack success rate on three downstream datasets is reduced from 99.88% to 8.10% after our defense on the backdoored BERT. The codes are publicly available at https://github.com/thunlp/RECIPE.

14:00-15:30 (East Foyer / Virtual)
**#164 Capturing Fine-Grained Regional Differences in Language Use through Voting Precinct Embeddings**
*Alex Rosenfeld and Lars Hinrichs*
Linguistic variation across a region of interest can be captured by partitioning the region into areas and using social media data to train embeddings that represent language use in those areas. Recent work has focused on larger areas, such as cities or counties, to ensure that enough social media data is available in each area, but larger areas have a limited ability to find fine-grained distinctions, such as intracity differences in language use. We demonstrate that it is possible to embed smaller areas, which can provide higher resolution analyses of language variation. We embed voting precincts, which are tiny, evenly sized political divisions for the administration of elections. The issue with modeling language use in small areas is that the data becomes incredibly sparse, with many areas having scant social media data. We propose a novel embedding approach that alternates training with smoothing, which mitigates these sparsity issues. We focus on linguistic variation across Texas as it is relatively understudied. We developed two novel quantitative evaluations that measure how well the embeddings can be used to capture linguistic variation. The first evaluation measures how well a model can map a dialect given terms specific to that dialect. The second evaluation measures how well a model can map preference of lexical variants. These evaluations show how embedding models could be used directly by sociolinguists and measure how much sociolinguistic information is contained within the embeddings. We complement this second evaluation with a methodology for using embeddings as a kind of genetic code where we identify "genes" that correspond to a sociological variable and connect those "genes" to a linguistic phenomenon thereby connecting sociological phenomena to linguistic ones. Finally, we explore approaches for inferring isoglosses using embeddings.

14:00-15:30 (East Foyer / Virtual)
**#165 Rethinking the Exploitation of Monolingual Data for Low-Resource Neural Machine Translation**
*Jianhui Pang, Baosong Yang, Derek Fai Wong, Yu Wan, Dayiheng Liu, Lidia Sam Chao and Jun Xie*
The utilization of monolingual data has been shown to be a promising strategy for addressing low-resource machine translation problems. Previous studies have demonstrated the effectiveness of techniques such as Back-Translation and self-supervised objectives, including Masked Language Modeling, Causal Language Modeling, and Denoise Autoencoding, in improving the performance of machine translation models. However, the manner in which these methods contribute to the success of machine translation tasks and how they can be effectively combined remains an under-researched area. In this study, we carry out a systematic investigation of the effects of these techniques on linguistic properties through the use of probing tasks, including source language comprehension, bilingual word alignment, and translation fluency. We further evaluate the impact of Pre-Training, Back-Translation, and Multi-Task Learning on bitexts of varying sizes. Our findings inform the design of more effective pipelines for leveraging monolingual data in extremely low-resource and low-resource machine translation tasks.

Experiment results show consistent performance gains in seven translation directions, which provide further support for our conclusions and understanding of the role of monolingual data in machine translation.

## Findings 2

14:00-15:30 (East Foyer / Virtual)

---

14:00-15:30 (East Foyer / Virtual)
**ReFSQL: A Retrieval-Augmentation Framework for Text-to-SQL Generation**
*Kun Zhang, Xiexiong Lin, Yuanzhuo Wang, Xin Zhang, Fei Sun, Cen Jianhe, Hexiang Tan, Xuhui Jiang and Huawei Shen*
Text-to-SQL is the task that aims at translating natural language questions into SQL queries. Existing methods directly align the natural language with SQL Language and train one encoder-decoder-based model to fit all questions. However, they underestimate the inherent structural characteristics of SQL, as well as the gap between specific structure knowledge and general knowledge. This leads to structure errors in the generated SQL. To address the above challenges, we propose a retrieval-argument framework, namely ReFSQL. It contains two parts, structure-enhanced retriever and the generator. Structure-enhanced retriever is designed to identify samples with comparable specific knowledge in an unsupervised way. Subsequently, we incorporate the retrieved samples' SQL into the input, enabling the model to acquire prior knowledge of similar SQL grammar. To further bridge the gap between specific and general knowledge, we present a mahalanobis contrastive learning method, which facilitates the transfer of the sample toward the specific knowledge distribution constructed by the retrieved samples. Experimental results on five datasets verify the effectiveness of our approach in improving the accuracy and robustness of Text-to-SQL generation. Our framework has achieved improved performance when combined with many other backbone models (including the 11B flan-T5) and also achieved state-of-the-art performance when compared to existing methods that employ the fine-tuning approach.

14:00-15:30 (East Foyer / Virtual)
**Injecting structural hints: Using language models to study inductive biases in language learning**
*Isabel Papadimitriou and Dan Jurafsky*
Both humans and transformer language models are able to learn language without explicit structural supervision. What cognitive inductive biases make this learning possible? Here, we examine the effect of different inductive learning biases by actively controlling the inductive biases of artificial learners: we structurally bias models by pretraining on synthetic formally-structured data, and evaluate these structural biases by fine-tuning on three typologically-distant human languages: English, Japanese, and Basque. We investigate the effect on downstream language perplexity of three types of inductive bias: 1) recursive, hierarchical processing 2) unrestricted token-token dependencies that can't be modeled by context-free grammars, and 3) a Zipfian power-law vocabulary distribution. We show that complex, non-context-free interactions between tokens form the best inductive biases. Our study leverages the capabilities of transformer models to run controlled language learning experiments that are not possible to run on humans, and surfaces hypotheses about the structures that facilitate language learning in both humans and machines.

14:00-15:30 (East Foyer / Virtual)
**From Simple to Complex: A Progressive Framework for Document-level Informative Argument Extraction**
*Quzhe Huang, Yanxi Zhang and Dongyan Zhao*
Document-level Event Argument Extraction (EAE) requires the model to extract arguments of multiple events from a single document. Considering the underlying dependencies between these events, recent efforts leverage the idea of "memory", where the results of already predicted events are cached and can be retrieved to help the prediction of upcoming events. These methods extract events according to their appearance order in the document, however, the event that appears in the first sentence does not mean that it is the easiest to extract. Existing methods might introduce noise to the extraction of upcoming events if they rely on an incorrect prediction of previous events. In order to provide more reliable memory, we propose a simple-to-complex progressive framework for document-level EAE. Specifically, we first calculate the difficulty of each event and then, we conduct the extraction following a simple-to-complex order. In this way, the memory will store the most certain results, and the model could use these reliable sources to help the prediction of more difficult events. Experiments on WikiEvents show that our model outperforms SOTA by 1.4% in F1, indicating the proposed simple-to-complex framework is useful in the EAE task.

14:00-15:30 (East Foyer / Virtual)
**Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy**
*Zhihong Shao, Yeyun Gong, yelong shen, Minlie Huang, Nan Duan and Weizhu Chen*
Retrieval-augmented generation has raise extensive attention as it is promising to address the limitations of large language models including outdated knowledge and hallucinations. However, retrievers struggle to capture relevance, especially for queries with complex information needs. Recent work has proposed to improve relevance modeling by having large language models actively involved in retrieval, i.e., to guide retrieval with generation. In this paper, we show that strong performance can be achieved by a method we call Iter-RetGen, which synergizes retrieval and generation in an iterative manner: a model's response to a task input shows what might be needed to finish the task, and thus can serve as an informative context for retrieving more relevant knowledge which in turn helps generate a better response in another iteration. Compared with recent work which interleaves retrieval with generation when completing a single output, Iter-RetGen processes all retrieved knowledge as a whole and largely preserves the flexibility in generation without structural constraints. We evaluate Iter-RetGen on multi-hop question answering, fact verification, and commonsense reasoning, and show that it can flexibly leverage parametric knowledge and non-parametric knowledge, and is superior to or competitive with state-of-the-art retrieval-augmented baselines while causing fewer overheads of retrieval and generation. We can further improve performance via generation-augmented retrieval adaptation.

14:00-15:30 (East Foyer / Virtual)
**X-SNS: Cross-Lingual Transfer Prediction through Sub-Network Similarity**
*Taejun Yun, Jinhyeon Kim, Deokyeong Kang, Seonghoon Lim, Jihoon Kim and Taeuk Kim*
Cross-lingual transfer (XLT) is an emergent ability of multilingual language models that preserves their performance on a task to a significant extent when evaluated in languages that were not included in the fine-tuning process. While English, due to its widespread usage, is typically regarded as the primary language for model adaption in various tasks, recent studies have revealed that the efficacy of XLT can be amplified by selecting the most appropriate source languages based on specific conditions. In this work, we propose the utilization of sub-network similarity between two languages as a proxy for predicting the compatibility of the languages in the context of XLT. Our approach is model-oriented, better reflecting the inner workings of foundation models. In addition, it requires only a moderate amount of raw text from candidate languages, distinguishing it from the majority of previous methods that rely on external resources. In experiments, we demonstrate that our method is more effective than baselines across diverse tasks. Specifically, it shows proficiency in ranking candidates for zero-shot XLT, achieving an improvement of 4.6% on average in terms of NDCG@3. We also provide extensive analyses that confirm the utility of sub-networks for XLT prediction.

14:00-15:30 (East Foyer / Virtual)
**Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification**
*Chia-Yu Hung, Zhiqiang Hu, Yujia Hu and Roy Ka-Wei Lee*
Authorship verification (AV) is a fundamental task in natural language processing (NLP) and computational linguistics, with applications in forensic analysis, plagiarism detection, and identification of deceptive content. Existing AV techniques, including traditional stylometric and deep learning approaches, face limitations in terms of data requirements and lack of explainability. To address these limitations, this paper proposes PromptAV, a novel technique that leverages Large-Language Models (LLMs) for AV by providing step-by-step stylometric explanation prompts. PromptAV outperforms state-of-the-art baselines, operates effectively with limited training data, and enhances interpretability through intuitive explanations, showcasing its potential as an effective and interpretable solution for the AV task.

14:00-15:30 (East Foyer / Virtual)
**Uniform Complexity for Text Generation**
*Joseph Marvin Imperial and Harish Tayyar Madabushi*
Large language models (LLMs) have shown promising results in a wide array of generative NLP tasks, such as summarization and machine translation. In the context of narrative generation, however, existing models still do not capture factors that contribute to producing consistent text. For instance, it is logical that a piece of text or a story should be uniformly readable throughout and that this form of complexity should be controllable. As such, if the complexity of an input text prompt is rated first-grade reading level in the Flesch Reading Ease test, then the generated text continuing the plot should also be within this range of complexity. With this in mind, we introduce Uniform Complexity for Text Generation (UCTG), a new benchmark test which raises the challenge of making generative models observe uniform linguistic properties with respect to prompts. We experiment with over 150+ linguistically and cognitively motivated features for evaluating text complexity in humans and generative models. From our results, we find that models such as GPT-2 struggle to preserve the complexity of input prompts used in its generations, even if finetuned with professionally written texts.

14:00-15:30 (East Foyer / Virtual)
**HuatuoGPT, Towards Taming Language Model to Be a Doctor**
*Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Hardy Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang and Haizhou Li*
In this paper, we present HuatuoGPT, a Large Language Model (LLM) for medical consultation. The core recipe of HuatuoGPT is to leverage both distilled data from **ChatGPT** and real-world data from **doctors** in the supervised fine-tuning stage. This is not only because purely using **ChatGPT**-distilled data might cause 'model collapse', but also because real-world data from **doctors** would be complementary to **ChatGPT**-distilled data. The responses from ChatGPT are usually detailed, well-presented, fluent, and instruction-followed, but it cannot perform like a doctor in many aspects, e.g. for interactive diagnosis. Therefore, the extra doctors' data could tame a distilled language model to perform like doctors. To synergize the strengths of both data sources, we introduce RLMF (Reinforcement Learning from Mixed Feedback) where a reward model is trained to align the language model with the merits that both sources (ChatGPT and doctors) bring. Experimental results (in GPT-4 evaluation, human evaluation, and medical benchmark datasets) demonstrate that HuatuoGPT achieves state-of-the-art results in performing medical consultation among open-source LLMs. It is worth noting that by using additional real-world data and RLMF, the distilled language model (i.e., HuatuoGPT) outperforms its teacher model (i.e., ChatGPT) in most cases.

14:00-15:30 (East Foyer / Virtual)
**Is ChatGPT a Good Multi-Party Conversation Solver?**
*Chao-Hong Tan, Jia-Chen Gu and Zhen-Hua Ling*
Large Language Models (LLMs) have emerged as influential instruments within the realm of natural language processing; nevertheless, their capacity to handle multi-party conversations (MPCs) – a scenario marked by the presence of multiple interlocutors involved in intricate information exchanges – remains uncharted. In this paper, we delve into the potential of generative LLMs such as ChatGPT and GPT-4 within the context of MPCs. An empirical analysis is conducted to assess the zero-shot learning capabilities of ChatGPT and GPT-4 by subjecting them to evaluation across three MPC datasets that encompass five representative tasks. The findings reveal that ChatGPT's performance on a number of evaluated MPC tasks leaves much to be desired, whilst GPT-4's results portend a promising future. Additionally, we endeavor to bolster performance through the incorporation of MPC structures, encompassing both speaker and addressee architecture. This study provides an exhaustive evaluation and analysis of applying generative LLMs to MPCs, casting a light upon the conception and creation of increasingly effective and robust MPC agents. Concurrently, this work underscores the challenges implicit in the utilization of LLMs for MPCs, such as deciphering graphical information flows and generating stylistically consistent responses.

14:00-15:30 (East Foyer / Virtual)
**Open-source Large Language Models are Strong Zero-shot Query Likelihood Models for Document Ranking**
*Shengyao Zhuang, Bing Liu, Bevan Koopman and Guido Zuccon*
In the field of information retrieval, Query Likelihood Models (QLMs) rank documents based on the probability of generating the query given the content of a document. Recently, advanced large language models (LLMs) have emerged as effective QLMs, showcasing promising ranking capabilities. This paper focuses on investigating the genuine zero-shot ranking effectiveness of recent LLMs, which are solely pre-trained on unstructured text data without supervised instruction fine-tuning. Our findings reveal the robust zero-shot ranking ability of such LLMs, highlighting that additional instruction fine-tuning may hinder effectiveness unless a question generation task is present in the fine-tuning dataset. Furthermore, we introduce a novel state-of-the-art ranking system that integrates LLM-based QLMs with a hybrid zero-shot retriever, demonstrating exceptional effectiveness in both zero-shot and few-shot scenarios. We make our codebase publicly available at https://github.com/ielab/llm-qlm.

14:00-15:30 (East Foyer / Virtual)
**WordNet Is All You Need: A Surprisingly Effective Unsupervised Method for Graded Lexical Entailment**
*Joseph Renner, Pascal Denis and Rémi Gilleron*
We propose a simple unsupervised approach which exclusively relies on WordNet (Miller,1995) for predicting graded lexical entailment (GLE) in English. Inspired by the seminal work of Resnik (1995), our method models GLE as the sum of two information-theoretic scores: a symmetric semantic similarity score and an asymmetric specificity loss score, both exploiting the hierarchical synset structure of WordNet. Our approach also includes a simple disambiguation mechanism to handle polysemy in a given word pair. Despite its simplicity, our method achieves performance above the state of the art (Spearman $\rho = 0.75$) on HyperLex (Vulic et al., 2017), the largest GLE dataset, outperforming all previous methods, including specialized word embeddings approaches that use WordNet as weak supervision.

14:00-15:30 (East Foyer / Virtual)
**Battle of the Large Language Models: Dolly vs LLaMA vs Vicuna vs Guanaco vs Bard vs ChatGPT - A Text-to-SQL Parsing Comparison**
*Shuo Sun, Yuchen Zhang, Jiahuan Yan, Yuze Gao, Donovan Ong, Bin Chen and Jian Su*
The success of ChatGPT has ignited an AI race, with researchers striving to develop new large language models (LLMs) that can match or

surpass the language understanding and generation abilities of commercial ones. In recent times, a number of models have emerged, claiming performance near that of GPT-3.5 or GPT-4 through various instruction-tuning methods. As practitioners of Text-to-SQL parsing, we are grateful for their valuable contributions to open-source research. However, it is important to approach these claims with a sense of scrutiny and ascertain the actual effectiveness of these models. Therefore, we pit six popular large language models against each other, systematically evaluating their Text-to-SQL parsing capability on nine benchmark datasets with five different prompting strategies, covering both zero-shot and few-shot scenarios. Regrettably, the open-sourced models fell significantly short of the performance achieved by closed-source models like GPT-3.5, highlighting the need for further work to bridge the performance gap between these models.

14:00-15:30 (East Foyer / Virtual)
**Beyond Candidates : Adaptive Dialogue Agent Utilizing Persona and Knowledge**
*Jungwoo Lim, Myunghoon Kang, Jinsung Kim, Jeongwook Kim, Yuna Hur and Heuiseok Lim*
To build ultimate dialogue agents, previous studies suggest models that ground both persona and knowledge. However, applying the dialogue system directly to the usual conversation is still limited because the system requires a complete sentence-formed persona and knowledge candidate sets from the given dataset. In contrast to the dialogue setting in the dataset, humans utilize semantic concepts in their minds rather than a set of pre-defined candidate sentences. Following this manner of human dialogue, we suggest an adaptive dialogue system that is applicable to situations where complete sentence-formed candidates are not given. Our model generates consistent and relevant persona descriptions and identifies relevant knowledge for engaging and knowledgeable responses, even with fragmentary information. We show that our model outperforms previous baselines that utilize persona and knowledge candidate sentences and conduct the human evaluation on the machine-generated responses. In addition, we conduct ablation studies to demonstrate the effectiveness of each component of our model. Furthermore, we apply our model to other dialogue datasets that only ground knowledge or persona to showcase its adaptability. Our code is available at https://github.com/dlawjddn803/BeCand.

14:00-15:30 (East Foyer / Virtual)
**Discovering Highly Influential Shortcut Reasoning: An Automated Template-Free Approach**
*Daichi Haraguchi, Kiyoaki Shirai, Naoya Inoue and Natthawut Kertkeidkachorn*
Shortcut reasoning is an irrational process of inference, which degrades the robustness of an NLP model. While a number of previous work has tackled the identification of shortcut reasoning, there are still two major limitations: (i) a method for quantifying the severity of the discovered shortcut reasoning is not provided; (ii) certain types of shortcut reasoning may be missed. To address these issues, we propose a novel method for identifying shortcut reasoning. The proposed method quantifies the severity of the shortcut reasoning by leveraging out-of-distribution data and does not make any assumptions about the type of tokens triggering the shortcut reasoning. Our experiments on Natural Language Inference and Sentiment Analysis demonstrate that our framework successfully discovers known and unknown shortcut reasoning in the previous work.

14:00-15:30 (East Foyer / Virtual)
**Multimodal Automated Fact-Checking: A Survey**
*Mubashara Akhtar, Michael Sejr Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl and Andreas Vlachos*
Misinformation is often conveyed in multiple modalities, e.g. a miscaptioned image. Multimodal misinformation is perceived as more credible by humans, and spreads faster than its text-only counterparts. While an increasing body of research investigates automated fact-checking (AFC), previous surveys mostly focus on text. In this survey, we conceptualise a framework for AFC including subtasks unique to multimodal misinformation. Furthermore, we discuss related terms used in different communities and map them to our framework. We focus on four modalities prevalent in real-world fact-checking: text, image, audio, and video. We survey benchmarks and models, and discuss limitations and promising directions for future research

14:00-15:30 (East Foyer / Virtual)
**Grounded and well-rounded: a methodological approach to the study of cross-modal and cross-lingual grounding**
*Timothee Mickus, Elaine Zosa and Denis Paperno*
Grounding has been argued to be a crucial component towards the development of more complete and truly semantically competent artificial intelligence systems. Literature has divided into two camps: While some argue that grounding allows for qualitatively different generalizations, others believe it can be compensated by mono-modal data quantity. Limited empirical evidence has emerged for or against either position, which we argue is due to the methodological challenges that come with studying grounding and its effects on NLP systems. In this paper, we establish a methodological framework for studying what the effects are—if any—of providing models with richer input sources than text-only. The crux of it lies in the construction of comparable samples of populations of models trained on different input modalities, so that we can tease apart the qualitative effects of different input sources from quantifiable model performances. Experiments using this framework reveal qualitative differences in model behavior between cross-modally grounded, cross-lingually grounded, and ungrounded models, which we measure both at a global dataset level as well as for specific word representations, depending on how concrete their semantics is.

14:00-15:30 (East Foyer / Virtual)
**Miracle: Towards Personalized Dialogue Generation with Latent-Space Multiple Personal Attribute Control**
*Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen and Jixiong Chen*
Personalized dialogue systems aim to endow the chatbot agent with more anthropomorphic traits for human-like interactions. Previous approaches have explored explicitly user profile modeling using text descriptions, implicit derivation of user embeddings, or utilizing handicraft prompts for ChatGPT-like models. However, textual personas are limited in describing multi-faceted attributes (*e.g., language style, inner character nuances*), implicit embedding suffers from personality sparsity, and handicraft prompts lack fine-grained and stable controllability. Hence, these approaches may struggle with complex personalized dialogue generation tasks that require generating controllable responses with multiple personal attributes. To this end, we propose **MIRACLE**, a novel personalized dialogue generation method through **M**ul**I**ple Pe**R**sonal **A**ttributes **C**ontrol within **L**atent-Space **E**nergy-based Models. ttributes **C**ontrol within **L**atent-Space **E**nergy-based Models. Specifically, our approach first disentangles complex personality into multi-faceted attributes. Subsequently, we employ a conditional variational auto-encoder to align with the dense personalized responses within a latent joint attribute space. We have also tailored a dedicated energy function and customized the ordinary differential equations sampling method to offer flexible attribute composition and precise attribute control. Extensive experiments demonstrate that MIRACLE outperforms several strong baselines in terms of personality controllability and response generation quality. Our dataset and code are available at `https://github.com/LZY-the-boys/MIRACLE`

14:00-15:30 (East Foyer / Virtual)
**A Comprehensive Evaluation of Tool-Assisted Generation Strategies**
*Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohnet and Mor Geva*
A growing area of research investigates augmenting language models with tools (e.g., search engines, calculators) to overcome their shortcomings (e.g., missing or incorrect knowledge, incorrect logical inferences). Various few-shot tool-usage strategies have been proposed. However, there is no systematic and fair comparison across different strategies, or between these strategies and strong baselines that do not leverage tools. We conduct an extensive empirical analysis, finding that (1) across various datasets, example difficulty levels, and models,

strong no-tool baselines are competitive to tool-assisted strategies, implying that effectively using tools with in-context demonstrations is a difficult unsolved problem; (2) for knowledge-retrieval tasks, strategies that *refine* incorrect outputs with tools outperform strategies that retrieve relevant information *ahead of* or *during generation*; (3) tool-assisted strategies are expensive in the number of tokens they require to work—incurring additional costs by orders of magnitude—which does not translate into significant improvement in performance. Overall, our findings suggest that few-shot tool integration is still an open challenge, emphasizing the need for comprehensive evaluations of future strategies to accurately assess their *benefits* and *costs*.

14:00-15:30 (East Foyer / Virtual)
**QADYNAMICS: Training Dynamics-Driven Synthetic QA Diagnostic for Zero-Shot Commonsense Question Answering**
*Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu and Yangqiu Song*
Zero-shot commonsense Question-Answering (QA) requires models to reason about general situations beyond specific benchmarks. State-of-the-art approaches fine-tune language models on QA pairs constructed from CommonSense Knowledge Bases (CSKBs) to equip the models with more commonsense knowledge in a QA context. However, current QA synthesis protocols may introduce noise from the CSKBs and generate ungrammatical questions and false negative options, which impede the model's ability to generalize. To address these issues, we propose QADYNAMICS, a training dynamics-driven framework for QA diagnostics and refinement. Our approach analyzes the training dynamics of each QA pair at both the question level and option level, discarding machine-detectable artifacts by removing uninformative QA pairs and mislabeled or false-negative options. Extensive experiments demonstrate the effectiveness of our approach, which outperforms all baselines while using only 33% of the synthetic data, even including LLMs such as ChatGPT. Moreover, expert evaluations confirm that our framework significantly improves the quality of QA synthesis. Our code and model checkpoints are available at https://github.com/HKUST-KnowComp/QaDynamics.

14:00-15:30 (East Foyer / Virtual)
**Enhancing Reasoning Capabilities by Instruction Learning and Chain-of-Thoughts for Implicit Discourse Relation Recognition**
*Yuxiang Lu, Yu Hong, Zhipang Wang and Guodong Zhou*
The aim of implicit discourse relation recognition is to comprehend the sense of connection between two arguments. In this work, we present a classification method that is solely based on generative models. Our proposed approach employs a combination of instruction templates and in-context learning to refine the generative model for effectively addressing the implicit discourse relation recognition task. Furthermore, we utilize Chain-of-Thoughts to partition the inference process into a sequence of three successive stages. This strategy enables us to fully utilize the autoregressive generative model's potential for knowledge acquisition and inference, ultimately leading to enhanced performance on this natural language understanding task. The results of our experiments, evaluated on benchmark datasets PDTB 2.0, PDTB 3.0, and the CoNLL16 shared task, demonstrate superior performance compared to previous state-of-the-art models.

14:00-15:30 (East Foyer / Virtual)
**Arabic Mini-ClimateGPT : A Climate Change and Sustainability Tailored Arabic LLM**
*Sahal Shaji Mullappilly, Abdelrahman M Shaker, Omkar Chakradhar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan and Fahad Khan*
Climate change is one of the most significant challenges we face together as a society. Creating awareness and educating policy makers the wide-ranging impact of climate change is an essential step towards a sustainable future. Recently, Large Language Models (LLMs) like ChatGPT and Bard have shown impressive conversational abilities and excel in a wide variety of NLP tasks. While these models are close-source, recently alternative open-source LLMs such as Stanford Alpaca and Vicuna have shown promising results. However, these open-source models are not specifically tailored for climate related domain specific information and also struggle to generate meaningful responses in other languages such as, Arabic. To this end, we propose a light-weight Arabic Mini-ClimateGPT that is built on an open-source LLM and is specifically fine-tuned on a conversational-style instruction tuning curated Arabic dataset Clima500-Instruct with over 500k instructions about climate change and sustainability. Further, our model also utilizes a vector embedding based retrieval mechanism during inference. We validate our proposed model through quantitative and qualitative evaluations on climate-related queries. Our model surpasses the baseline LLM in 88.3% of cases during ChatGPT-based evaluation. Furthermore, our human expert evaluation reveals an 81.6% preference for our model's responses over multiple popular open-source models. Our open-source demos, models and curated instruction sets are available here : https://github.com/mbzuai-oryx/ClimateGPT

14:00-15:30 (East Foyer / Virtual)
**Execution-Based Evaluation for Open-Domain Code Generation**
*Zhiruo Wang, Shuyan Zhou, Daniel Fried and Graham Neubig*
To extend the scope of coding queries to more realistic settings, we propose ODEX, the first Open-Domain EXecution-based natural language (NL) to Python code generation dataset. ODEX has 945 NL-Code pairs spanning 79 diverse libraries, along with 1,707 human-written test cases for execution. Our NL-Code pairs are harvested from StackOverflow forums to encourage natural and practical coding queries. Moreover, ODEX supports four natural languages as intents, in English, Spanish, Japanese, and Russian. ODEX unveils intriguing behavioral differences among top-performing code language models (LM). While CODEX achieves better overall results, CODEGEN improves effectively via scaling – CODEGEN 6.1B performs comparably with CODEX 12B. Both models show substantial gaps between open and closed domains, but CODEGEN gaps tend to decrease with model size while CODEX gaps increase. We release ODEX to facilitate research into open-domain problems for the code generation community.

14:00-15:30 (East Foyer / Virtual)
**TESTA: Temporal-Spatial Token Aggregation for Long-form Video-Language Understanding**
*Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun and Lu Hou*
Large-scale video-language pre-training has made remarkable strides in advancing video-language understanding tasks. However, the heavy computational burden of video encoding remains a formidable efficiency bottleneck, particularly for long-form videos. These videos contain massive visual tokens due to their inherent 3D properties and spatiotemporal redundancy, making it challenging to capture complex temporal and spatial relationships. To tackle this issue, we propose an efficient method called TEmporal-Spatial Token Aggregation (TESTA). TESTA condenses video semantics by adaptively aggregating similar frames, as well as similar patches within each frame. TESTA can reduce the number of visual tokens by 75% and thus accelerate video encoding. Building upon TESTA, we introduce a pre-trained video-language model equipped with a divided space-time token aggregation module in each video encoder block. We evaluate our model on five datasets for paragraph-to-video retrieval and long-form VideoQA tasks. Experimental results show that TESTA improves computing efficiency by 1.7 times, and achieves significant performance gains from its scalability in processing longer input frames, e.g., +13.7 R@1 on QuerYD and +6.5 R@1 on Condensed Movie.

14:00-15:30 (East Foyer / Virtual)
**InfoDiffusion: Information Entropy Aware Diffusion Process for Non-Autoregressive Text Generation**
*Renzhi Wang, Jing Li and Piji Li*
Diffusion models have garnered considerable interest in the field of text generation. Several studies have explored text diffusion models with

different structures and applied them to various tasks, including named entity recognition and summarization. However, there exists a notable disparity between the "easy-first" text generation process of current diffusion models and the "keyword-first" natural text generation process of humans, which has received limited attention. To bridge this gap, we propose InfoDiffusion, a non-autoregressive text diffusion model. Our approach introduces a "keyinfo-first" generation strategy and incorporates a noise schedule based on the amount of text information. In addition, InfoDiffusion combines self-conditioning with a newly proposed partially noising model structure. Experimental results show that InfoDiffusion outperforms the baseline model in terms of generation quality and diversity, as well as exhibiting higher sampling efficiency.

14:00-15:30 (East Foyer / Virtual)
**Probing LLMs for hate speech detection: strengths and vulnerabilities**
*Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee and Punyajoy Saha*
Recently efforts have been made by social media platforms as well as researchers to detect hateful or toxic language using large language models. However, none of these works aim to use explanation, additional context and victim community information in the detection process. We utilise different prompt variation, input information and evaluate large language models in zero shot setting (without adding any in-context examples). We select two large language models (GPT-3.5 and text-davinci) and three datasets - HateXplain, implicit hate and ToxicSpans. We find that on average including the target information in the pipeline improves the model performance substantially ($\sim 20-30\%$) over the baseline across the datasets. There is also a considerable effect of adding the rationales/explanations into the pipeline ($\sim 10-20\%$) over the baseline across the datasets. In addition, we further provide a typology of the error cases where these large language models fail to (i) classify and (ii) explain the reason for the decisions they take. Such vulnerable points automatically constitute 'jailbreak' prompts for these models and industry scale safeguard techniques need to be developed to make the models robust against such prompts.

14:00-15:30 (East Foyer / Virtual)
**Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs**
*Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis, Christian Kaestner and Tongshuang Wu*
Current model testing work has mostly focused on creating test cases. Identifying what to test is a step that is largely ignored and poorly supported. We propose Weaver, an interactive tool that supports requirements elicitation for guiding model testing. Weaver uses large language models to generate knowledge bases and recommends concepts from them interactively, allowing testers to elicit requirements for further testing. Weaver provides rich external knowledge to testers and encourages testers to systematically explore diverse concepts beyond their own biases. In a user study, we show that both NLP experts and non-experts identified more, as well as more diverse concepts worth testing when using Weaver. Collectively, they found more than 200 failing test cases for stance detection with zero-shot ChatGPT. Our case studies further show that Weaver can help practitioners test models in real-world settings, where developers define more nuanced application scenarios (e.g., code understanding and transcript summarization) using LLMs.

14:00-15:30 (East Foyer / Virtual)
**Emptying the Ocean with a Spoon: Should We Edit Models?**
*Yuval Pinter and Michael Elhadad*
We call into question the recently popularized method of direct model editing as a means of correcting factual errors in LLM generations. We contrast model editing with three similar but distinct approaches that pursue better defined objectives: (1) retrieval-based architectures, which decouple factual memory from inference and linguistic capabilities embodied in LLMs; (2) concept erasure methods, which aim at preventing systemic bias in generated text; and (3) attribution methods, which aim at grounding generations into identified textual sources. We argue that direct model editing cannot be trusted as a systematic remedy for the disadvantages inherent to LLMs, and while it has proven potential in improving model explainability, it opens risks by reinforcing the notion that models can be trusted for factuality. We call for cautious promotion and application of model editing as part of the LLM deployment process, and for responsibly limiting the use cases of LLMs to those not relying on editing as a critical component.

14:00-15:30 (East Foyer / Virtual)
**Towards Concept-Aware Large Language Models**
*Chen Shani, Jilles Vreeken and Dafna Shahaf*
Concepts play a pivotal role in various human cognitive functions, including learning, reasoning and communication. However, there is very little work on endowing machines with the ability to form and reason with concepts. In particular, state-of-the-art large language models (LLMs) work at the level of tokens, not concepts. In this work, we analyze how well contemporary LLMs capture human concepts and their structure. We then discuss ways to develop concept-aware LLMs, taking place at different stages of the pipeline. We sketch a method for pretraining LLMs using concepts, and also explore the simpler approach that uses the output of existing LLMs. Despite its simplicity, our proof-of-concept is shown to better match human intuition, as well as improve the robustness of predictions. These preliminary results underscore the promise of concept-aware LLMs.

14:00-15:30 (East Foyer / Virtual)
**The Iron(ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony and Sarcasm Generation**
*Tyler Loakman, Aaron Maladry and Chenghua Lin*
Human evaluation in often considered to be the gold standard method of evaluating a Natural Language Generation system. However, whilst its importance is accepted by the community at large, the quality of its execution is often brought into question. In this position paper, we argue that the generation of more esoteric forms of language - humour, irony and sarcasm - constitutes a subdomain where the characteristics of selected evaluator panels are of utmost importance, and every effort should be made to report demographic characteristics wherever possible, in the interest of transparency and replicability. We support these claims with an overview of each language form and an analysis of examples in terms of how their interpretation is affected by different participant variables. We additionally perform a critical survey of recent works in NLG to assess how well evaluation procedures are reported in this subdomain, and note a severe lack of open reporting of evaluator demographic information, and a significant reliance on crowdsourcing platforms for recruitment.

14:00-15:30 (East Foyer / Virtual)
**1-PAGER: One Pass Answer Generation and Evidence Retrieval**
*Palak Jain, Livio Baldini Soares and Tom Kwiatkowski*
We present 1-Pager the first system that answers a question and retrieves evidence using a single Transformer-based model and decoding process. 1-Pager incrementally partitions the retrieval corpus using constrained decoding to select a document and answer string, and we show that this is competitive with comparable retrieve-and-read alternatives according to both retrieval and answer accuracy metrics. 1-Pager also outperforms the equivalent 'closed-book' question answering model, by grounding predictions in an evidence corpus. While 1-Pager is not yet on-par with more expensive systems that read many more documents before generating an answer, we argue that it provides an important step toward attributed generation by folding retrieval into the sequence-to-sequence paradigm that is currently dominant in NLP. We also show that the search paths used to partition the corpus are easy to read and understand, paving a way forward for interpretable neural retrieval.

14:00-15:30 (East Foyer / Virtual)
**Generalizing Few-Shot Named Entity Recognizers to Unseen Domains with Type-Related Features**
*Zihan Wang, Ziqi Zhao, Zhumin Chen, Pengjie Ren, Maarten de Rijke and Zhaochun Ren*
Few-shot named entity recognition (NER) has shown remarkable progress in identifying entities in low-resource domains. However, few-shot NER methods still struggle with out-of-domain (OOD) examples due to their reliance on manual labeling for the target domain. To address this limitation, recent studies enable generalization to an unseen target domain with only a few labeled examples using data augmentation techniques. Two important challenges remain: First, augmentation is limited to the training data, resulting in minimal overlap between the generated data and OOD examples. Second, knowledge transfer is implicit and insufficient, severely hindering model generalizability and the integration of knowledge from the source domain. In this paper, we propose a framework, prompt learning with type-related features (PLTR), to address these challenges. To identify useful knowledge in the source domain and enhance knowledge transfer, PLTR automatically extracts entity type-related features (TRFs) based on mutual information criteria. To bridge the gap between training and OOD data, PLTR generates a unique prompt for each unseen example by selecting relevant TRFs. We show that PLTR achieves significant performance improvements on in-domain and cross-domain datasets. The use of PLTR facilitates model adaptation and increases representation similarities between the source and unseen domains.

14:00-15:30 (East Foyer / Virtual)
**WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia**
*Sina Semnani, Violet Yao, Heidi Chenyu Zhang and Monica Lam*
This paper presents the first few-shot LLM-based chatbot that almost never hallucinates and has high conversationality and low latency. WikiChat is grounded on the English Wikipedia, the largest curated free-text corpus. WikiChat generates a response from an LLM, retains only the grounded facts, and combines them with additional information it retrieves from the corpus to form factual and engaging responses. We distill WikiChat based on GPT-4 into a 7B-parameter LLaMA model with minimal loss of quality, to significantly improve its latency, cost and privacy, and facilitate research and deployment. Using a novel hybrid human-and-LLM evaluation methodology, we show that our best system achieves 97.3% factual accuracy in simulated conversations. It significantly outperforms all retrieval-based and LLM-based baselines, and by 3.9%, 38.6% and 51.0% on head, tail and recent knowledge compared to GPT-4. Compared to previous state-of-the-art retrieval-based chatbots, WikiChat is also significantly more informative and engaging, just like an LLM. WikiChat achieves 97.9% factual accuracy in conversations with human users about recent topics, 55.0% better than GPT-4, while receiving significantly higher user ratings and more favorable comments.

14:00-15:30 (East Foyer / Virtual)
**A Multi-Modal Multilingual Benchmark for Document Image Classification**
*Yoshinari Fujinuma, Siddharth Varia, Nishant Sankaran, Srikar Appalaraju, Bonan Min and Yogarshi Vyas*
Document image classification is different from plain-text document classification and consists of classifying a document by understanding the content and structure of documents such as forms, emails, and other such documents. We show that the only existing dataset for this task (Lewis et al., 2006) has several limitations and we introduce two newly curated multilingual datasets WIKI-DOC and MULTIEURLEX-DOC that overcome these limitations. We further undertake a comprehensive study of popular visually-rich document understanding or Document AI models in previously untested setting in document image classification such as 1) multi-label classification, and 2) zero-shot cross-lingual transfer setup. Experimental results show limitations of multilingual Document AI models on cross-lingual transfer across typologically distant languages. Our datasets and findings open the door for future research into improving Document AI models.

14:00-15:30 (East Foyer / Virtual)
**A Computational Interface to Translate Strategic Intent from Unstructured Language in a Low-Data Setting**
*Pradyumna Tambwekar, Lakshita Dodeja, Nathan Vaska, Wei Xu and Matthew Gombolay*
Many real-world tasks involve a mixed-initiative setup, wherein humans and AI systems collaboratively perform a task. While significant work has been conducted towards enabling humans to specify, through language, exactly how an agent should complete a task (i.e., low-level specification), prior work lacks on interpreting the high-level strategic intent of the human commanders. Parsing strategic intent from language will allow autonomous systems to independently operate according to the user's plan without frequent guidance or instruction. In this paper, we build a computational interface capable of translating unstructured language strategies into actionable intent in the form of goals and constraints. Leveraging a game environment, we collect a dataset of over 1000 examples, mapping language strategies to the corresponding goals and constraints, and show that our model, trained on this dataset, significantly outperforms human interpreters in inferring strategic intent (i.e., goals and constraints) from language ($p < 0.05$). Furthermore, we show that our model (125M parameters) significantly outperforms ChatGPT for this task ($p < 0.05$) in a low-data setting.

14:00-15:30 (East Foyer / Virtual)
**Salespeople vs SalesBot: Exploring the Role of Educational Value in Conversational Recommender Systems**
*Lidiya Murakhovs'ka, Philippe Laban, Tian Xie, Caiming Xiong and Chien-Sheng Wu*
Making big purchases requires consumers to research or consult a salesperson to gain domain expertise. However, existing conversational recommender systems (CRS) often overlook users' lack of background knowledge, focusing solely on gathering preferences. In this work, we define a new problem space for conversational agents that aim to provide both product recommendations and educational value through mixed-type mixed-initiative dialog. We introduce SalesOps, a framework that facilitates the simulation and evaluation of such systems by leveraging recent advancements in large language models (LLMs). We build SalesBot and ShopperBot, a pair of LLM-powered agents that can simulate either side of the framework. A comprehensive human study compares SalesBot against professional salespeople, revealing that although SalesBot approaches professional performance in terms of fluency and informativeness, it lags behind in recommendation quality. We emphasize the distinct limitations both face in providing truthful information, highlighting the challenges of ensuring faithfulness in the CRS context. We release our code and make all data available.

14:00-15:30 (East Foyer / Virtual)
**Conditioning on Dialog Acts improves Empathy Style Transfer**
*Renyi Qu, Lyle Ungar and João Sedoc*
We explore the role of dialog acts in style transfer, specifically empathy style transfer – rewriting a sentence to make it more empathetic without changing its meaning. Specifically, we use two novel few-shot prompting strategies: target prompting, which only uses examples of the target style (unlike traditional prompting with source/target pairs), and dialog-act-conditioned prompting, which first estimates the dialog act of the source sentence and then makes it more empathetic using few-shot examples of the same dialog act. Our study yields two key findings: (1) Target prompting typically improves empathy more effectively while maintaining the same level of semantic similarity; (2) Dialog acts matter. Dialog-act-conditioned prompting enhances empathy while preserving both semantics and the dialog-act type. Different dialog acts benefit differently from different prompting methods, highlighting the need for further investigation of the role of dialog acts in style transfer.

14:00-15:30 (East Foyer / Virtual)
**Automated Few-Shot Classification with Instruction-Finetuned Language Models**

*Rami Aly, Xingjian Shi, Kaixiang Lin, Aston Zhang and Andrew Gordon Wilson*
A particularly successful class of approaches for few-shot learning combines language models with prompts - hand-crafted task descriptions that complement data samples. However, designing prompts by hand for each task commonly requires domain knowledge and substantial guesswork. We observe, in the context of classification tasks, that instruction finetuned language models are remarkably robust towards some dimensions of a prompt's design. We subsequently propose a simple method to eliminate the need for handcrafted prompts, named AuT-Few. This approach consists of (i) a prompt retrieval module that selects suitable task instructions from the instruction-tuning knowledge base, and (ii) the generation of two distinct, semantically meaningful, class descriptions and a selection mechanism via cross-validation. Over 12 datasets, spanning 8 classification tasks, we show that AuT-Few outperforms current state-of-the-art few-shot learning methods. Moreover, AuT-Few is the best ranking method across datasets on the RAFT few-shot benchmark. Notably, these results are achieved without task-specific handcrafted prompts on unseen tasks.

14:00-15:30 (East Foyer / Virtual)
**GLGR: Question-aware Global-to-Local Graph Reasoning for Multi-party Dialogue Reading Comprehension**
*Yanling Li, Bowei Zou, Yifan Fan, Xibo Li, Ai Ti Aw and Yu Hong*
Graph reasoning contributes to the integration of discretely-distributed attentive information (clues) for Multi-party Dialogue Reading Comprehension (MDRC). This is attributed primarily to multi-hop reasoning over global conversational structures. However, existing approaches barely apply questions for anti-noise graph reasoning. More seriously, the local semantic structures in utterances are neglected, although they are beneficial for bridging across semantically-related clues. In this paper, we propose a question-aware global-to-local graph reasoning approach. It expands the canonical Interlocutor-Utterance graph by introducing a question node, enabling comprehensive global graph reasoning. More importantly, it constructs a semantic-role graph for each utterance, and accordingly performs local graph reasoning conditioned on the semantic relations. We design a two-stage encoder network to implement the progressive reasoning from the global graph to local. The experiments on the benchmark datasets Molweni and FriendsQA show that our approach yields significant improvements, compared to BERT and ELECTRA baselines. It achieves 73.6% and 77.2% F1-scores on Molweni and FriendsQA, respectively, outperforming state-of-the-art methods that employ different pretrained language models as backbones.

14:00-15:30 (East Foyer / Virtual)
**Measure Children's Mindreading Ability with Machine Reading**
*Yuliang Yan, Xiaohua Wang, Xiang Zhou, Xiaoqing Zheng and Xuanjing Huang*
Recently, much research in psychology has benefited from the advances in machine learning techniques. Some recent studies showed that it is possible to build automated scoring models for children's mindreading. These models were trained on a set of manually-labeled question-response pairs, which were collected by asking children to answer one or two questions after a short story is told or a video clip is played. However, existing models did not take the features of the stories and video clips into account when scoring, which obviously will reduce the accuracy of the scoring models. Furthermore, considering that different psychological tests may contain the same questions, this approach cannot be extended to other related psychological test datasets. In this study, we proposed a multi-modal learning framework to leverage the features extracted from the stories and videos related to the questions being asked during the children's mindreading evaluation. Experimental results show that the scores produced by the proposed models agree well with those graded by human experts, highlighting the potential of the proposed network architecture for practical automated children's mindreading scoring systems.

14:00-15:30 (East Foyer / Virtual)
**Verb Conjugation in Transformers Is Determined by Linear Encodings of Subject Number**
*Sophie Hao and Tal Linzen*
Deep architectures such as Transformers are sometimes criticized for having uninterpretable "black-box" representations. We use causal intervention analysis to show that, in fact, some linguistic features are represented in a linear, interpretable format. Specifically, we show that BERT's ability to conjugate verbs relies on a linear encoding of subject number that can be manipulated with predictable effects on conjugation accuracy. This encoding is found in the subject position at the first layer and the verb position at the last layer, but distributed across positions at middle layers, particularly when there are multiple cues to subject number.

14:00-15:30 (East Foyer / Virtual)
**Speaking Style Conversion in the Waveform Domain Using Discrete Self-Supervised Units**
*Gallil Maimon and Yossi Adi*
We introduce DISSC, a novel, lightweight method that converts the rhythm, pitch contour and timbre of a recording to a target speaker in a textless manner. Unlike DISSC, most voice conversion (VC) methods focus primarily on timbre, and ignore people's unique speaking style (prosody). The proposed approach uses a pretrained, self-supervised model for encoding speech to discrete units, which makes it simple, effective, and fast to train. All conversion modules are only trained on reconstruction like tasks, thus suitable for any-to-many VC with no paired data. We introduce a suite of quantitative and qualitative evaluation metrics for this setup, and empirically demonstrate that DISSC significantly outperforms the evaluated baselines. Code and samples are available at https://pages.cs.huji.ac.il/adiyoss-lab/dissc/.

14:00-15:30 (East Foyer / Virtual)
**SELFOOD: Self-Supervised Out-Of-Distribution Detection via Learning to Rank**
*Dheeraj Mekala, Adithya Samavedhi, Chengyu Dong and Jingbo Shang*
Deep neural classifiers trained with cross-entropy loss (CE loss) often suffer from poor calibration, necessitating the task of out-of-distribution (OOD) detection. Traditional supervised OOD detection methods require expensive manual annotation of in-distribution and OOD samples. To address the annotation bottleneck, we introduce SELFOOD, a self-supervised OOD detection method that requires only in-distribution samples as supervision. We cast OOD detection as an inter-document intra-label (IDIL) ranking problem and train the classifier with our pairwise ranking loss, referred to as IDIL loss. Specifically, given a set of in-distribution documents and their labels, for each label, we train the classifier to rank the softmax scores of documents belonging to that label to be higher than the scores of documents that belong to other labels. Unlike CE loss, our IDIL loss function reaches zero when the desired confidence ranking is achieved and gradients are backpropagated to decrease probabilities associated with incorrect labels rather than continuously increasing the probability of the correct label. Extensive experiments with several classifiers on multiple classification datasets demonstrate the effectiveness of our method in both coarse- and fine-grained settings.

14:00-15:30 (East Foyer / Virtual)
**Interpreting Answers to Yes-No Questions in User-Generated Content**
*Shivam Mathur, Keun Hee Park, Dhivya Chinnappa, Saketh Kotamraju and Eduardo Blanco*
Interpreting answers to yes-no questions in social media is difficult. Yes and no keywords are uncommon, and the few answers that include them are rarely to be interpreted what the keywords suggest. In this paper, we present a new corpus of 4,442 yes-no question-answer pairs from Twitter. We discuss linguistic characteristics of answers whose interpretation is yes or no, as well as answers whose interpretation is unknown. We show that large language models are far from solving this problem, even after fine-tuning and blending other corpora for the same problem but outside social media.

14:00-15:30 (East Foyer / Virtual)
**Fusing Temporal Graphs into Transformers for Time-Sensitive Question Answering**
*Xin Su, Phillip Howard, Nagib Hakim and Steven Bethard*
Answering time-sensitive questions from long documents requires temporal reasoning over the times in questions and documents. An important open question is whether large language models can perform such reasoning solely using a provided text document, or whether they can benefit from additional temporal information extracted using other systems. We address this research question by applying existing temporal information extraction systems to construct temporal graphs of events, times, and temporal relations in questions and documents. We then investigate different approaches for fusing these graphs into Transformer models. Experimental results show that our proposed approach for fusing temporal graphs into input text substantially enhances the temporal reasoning capabilities of Transformer models with or without fine-tuning. Additionally, our proposed method outperforms various graph convolution-based approaches and establishes a new state-of-the-art performance on SituatedQA and three splits of TimeQA.

14:00-15:30 (East Foyer / Virtual)
**Identifying Conspiracy Theories News based on Event Relation Graph**
*Yuanyuan Lei and Ruihong Huang*
Conspiracy theories, as a type of misinformation, are narratives that explains an event or situation in an irrational or malicious manner. While most previous work examined conspiracy theory in social media short texts, limited attention was put on such misinformation in long news documents. In this paper, we aim to identify whether a news article contains conspiracy theories. We observe that a conspiracy story can be made up by mixing uncorrelated events together, or by presenting an unusual distribution of relations between events. Achieving a contextualized understanding of events in a story is essential for detecting conspiracy theories. Thus, we propose to incorporate an event relation graph for each article, in which events are nodes, and four common types of event relations, coreference, temporal, causal, and subevent relations, are considered as edges. Then, we integrate the event relation graph into conspiracy theory identification in two ways: an event-aware language model is developed to augment the basic language model with the knowledge of events and event relations via soft labels; further, a heterogeneous graph attention network is designed to derive a graph embedding based on hard labels. Experiments on a large benchmark dataset show that our approach based on event relation graph improves both precision and recall of conspiracy theory identification, and generalizes well for new unseen media sources.

14:00-15:30 (East Foyer / Virtual)
**Debiasing Multimodal Models via Causal Information Minimization**
*Vaidehi Patil, Adyasha Maharana and Mohit Bansal*
Most existing debiasing methods for multimodal models, including causal intervention and inference methods, utilize approximate heuristics to represent the biases, such as shallow features from early stages of training or unimodal features for multimodal tasks like VQA, etc., which may not be accurate. In this paper, we study bias arising from confounders in a causal graph for multimodal data, and examine a novel approach that leverages causally-motivated information minimization to learn the confounder representations. Robust predictive features contain diverse information that helps a model generalize to out-of-distribution data. Hence, minimizing the information content of features obtained from a pretrained biased model helps learn the simplest predictive features that capture the underlying data distribution. We treat these features as confounder representations and use them via methods motivated by causal theory to remove bias from models. We find that the learned confounder representations indeed capture dataset biases and the proposed debiasing methods improve out-of-distribution (OOD) performance on multiple multimodal datasets without sacrificing in-distribution performance. Additionally, we introduce a novel metric to quantify the sufficiency of spurious features in models' predictions that further demonstrates the effectiveness of our proposed methods.

14:00-15:30 (East Foyer / Virtual)
**Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting**
*Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O'Connell, Agustina Saenz, Pranav Rajpurkar and Michael Moor*
Automatically generated reports from medical images promise to improve the workflow of radiologists. Existing methods consider an image-to-report modeling task by directly generating a fully-fledged report from an image. However, this conflates the content of the report (e.g., findings and their attributes) with its style (e.g., format and choice of words), which can lead to clinically inaccurate reports. To address this, we propose a two-step approach for radiology report generation. First, we extract the content from an image; then, we verbalize the extracted content into a report that matches the style of a specific radiologist. For this, we leverage RadGraph—a graph representation of reports—together with large language models (LLMs). In our quantitative evaluations, we find that our approach leads to beneficial performance. Our human evaluation with clinical raters highlights that the AI-generated reports are indistinguishably tailored to the style of individual radiologist despite leveraging only a few examples as context.

14:00-15:30 (East Foyer / Virtual)
**InfoCL: Alleviating Catastrophic Forgetting in Continual Text Classification from An Information Theoretic Perspective**
*Yifan Song, Peiyi Wang, Weimin Xiong, Dawei Zhu, Tianyu Liu, Zhifang Sui and Sujian Li*
Continual learning (CL) aims to constantly learn new knowledge over time while avoiding catastrophic forgetting on old tasks. We focus on continual text classification under the class-incremental setting. Recent CL studies have identified the severe performance decrease on analogous classes as a key factor for catastrophic forgetting. In this paper, through an in-depth exploration of the representation learning process in CL, we discover that the compression effect of the information bottleneck leads to confusion on analogous classes. To enable the model learn more sufficient representations, we propose a novel replay-based continual text classification method, InfoCL. Our approach utilizes fast-slow and current-past contrastive learning to perform mutual information maximization and better recover the previously learned representations. In addition, InfoCL incorporates an adversarial memory augmentation strategy to alleviate the overfitting problem of replay. Experimental results demonstrate that InfoCL effectively mitigates forgetting and achieves state-of-the-art performance on three text classification tasks.

14:00-15:30 (East Foyer / Virtual)
**Pretraining Language Models with Text-Attributed Heterogeneous Graphs**
*Tao Zou, Le Yu, Yifei Huang, Leilei Sun and Bowen Du*
In many real-world scenarios (e.g., academic networks, social platforms), different types of entities are not only associated with texts but also connected by various relationships, which can be abstracted as Text-Attributed Heterogeneous Graphs (TAHGs). Current pretraining tasks for Language Models (LMs) primarily focus on separately learning the textual information of each entity and overlook the crucial aspect of capturing topological connections among entities in TAHGs. In this paper, we present a new pretraining framework for LMs that explicitly considers the topological and heterogeneous information in TAHGs. Firstly, we define a context graph as neighborhoods of a target node within specific orders and propose a topology-aware pretraining task to predict nodes involved in the context graph by jointly optimizing an LM and an auxiliary heterogeneous graph neural network. Secondly, based on the observation that some nodes are text-rich while others have little text, we devise a text augmentation strategy to enrich textless nodes with their neighbors' texts for handling the imbalance issue. We conduct link prediction and node classification tasks on three datasets from various domains. Experimental results demonstrate the supe-

riority of our approach over existing methods and the rationality of each design. Our code is available at https://github.com/Hope-Rita/THLM.

14:00-15:30 (East Foyer / Virtual)
**ReadPrompt: A Readable Prompting Method for Reliable Knowledge Probing**
*Zezhong Wang, Luyao Ye, Hongru Wang, Wai-Chung Kwan, David Ho and Kam-Fai Wong*
Knowledge probing is a task to assess the knowledge encoded within pre-trained language models (PLMs) by having the PLM complete prompts such as "Italy is located in __,". The model's prediction precision serves as a lower bound for the amount of knowledge it contains. Subsequent works explore training a series of vectors as prompts to guide PLMs towards more accurate predictions. However, these methods compromise the readability of the prompts. We cannot directly understand these prompts from their literal meaning, making it difficult to verify whether they are correct. Consequently, the credibility of probing results derived from these prompts is diminished. To address the issue, we propose a novel method called ReadPrompt, which aims to identify meaningful sentences to serve as prompts. Experiments show that ReadPrompt achieves state-of-the-art performance on the current knowledge probing benchmark. Moreover, since the prompt is readable, we discovered a misalignment between constructed prompts and knowledge, which is also present in current prompting methods verified by an attack experiment. We claim that the probing outcomes of the current prompting methods are unreliable that overestimate the knowledge contained within PLMs.

14:00-15:30 (East Foyer / Virtual)
**Adaptive Textual Label Noise Learning based on Pre-trained Models**
*Shaohuan Cheng, Wenyu Chen, Fu Mingsheng, Xuanting Xie and Hong Qu*
The label noise in real-world scenarios is unpredictable and can even be a mixture of different types of noise. To meet this challenge, we develop an adaptive textual label noise learning framework based on pre-trained models, which consists of an adaptive warm-up stage and a hybrid training stage. Specifically, an early stopping method, relying solely on the training set, is designed to dynamically terminate the warm-up process based on the model's fit level to different noise scenarios. The hybrid training stage incorporates several generalization strategies to gradually correct mislabeled instances, thereby making better use of noisy data. Experiments on multiple datasets demonstrate that our approach performs comparably or even surpasses the state-of-the-art methods in various noise scenarios, including scenarios with the mixture of multiple types of noise.

14:00-15:30 (East Foyer / Virtual)
**From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification**
*Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan and Xueqi Cheng*
Retrieval-enhanced methods have become a primary approach in fact verification (FV); it requires reasoning over multiple retrieved pieces of evidence to verify the integrity of a claim. To retrieve evidence, existing work often employs off-the-shelf retrieval models whose design is based on the probability ranking principle. We argue that, rather than relevance, for FV we need to focus on the utility that a claim verifier derives from the retrieved evidence. We introduce the **feedback-based evidence retriever**( $FER$) that optimizes the evidence retrieval process by incorporating feedback from the claim verifier. As a feedback signal we use the divergence in utility between how effectively the verifier utilizes the retrieved evidence and the ground-truth evidence to produce the final claim label. Empirical studies demonstrate the superiority of FER over prevailing baselines.

14:00-15:30 (East Foyer / Virtual)
**Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance**
*Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche and Tanja Käser*
Large Language Models (LLMs) are increasingly utilized in educational tasks such as providing writing suggestions to students. Despite their potential, LLMs are known to harbor inherent biases which may negatively impact learners. Previous studies have investigated bias in models and data representations separately, neglecting the potential impact of LLM bias on human writing. In this paper, we investigate how bias transfers through an AI writing support pipeline. We conduct a large-scale user study with 231 students writing business case peer reviews in German. Students are divided into five groups with different levels of writing support: one in-classroom group with recommender system feature-based suggestions and four groups recruited from Prolific – a control group with no assistance, two groups with suggestions from fine-tuned GPT-2 and GPT-3 models, and one group with suggestions from pre-trained GPT-3.5. Using GenBit gender bias analysis and Word Embedding Association Tests (WEAT), we evaluate the gender bias at various stages of the pipeline: in reviews written by students, in suggestions generated by the models, and in model embeddings directly. Our results demonstrate that there is no significant difference in gender bias between the resulting peer reviews of groups with and without LLM suggestions. Our research is therefore optimistic about the use of AI writing support in the classroom, showcasing a context where bias in LLMs does not transfer to students' responses.

14:00-15:30 (East Foyer / Virtual)
**Watermarking PLMs on Classification Tasks by Combining Contrastive Learning with Weight Perturbation**
*Chenxi Gu, Xiaoqing Zheng, Jianhan Xu, Muling Wu, Cenyuan Zhang, Chengsong Huang, Hua Cai and Xuanjing Huang*
Large pre-trained language models (PLMs) have achieved remarkable success, making them highly valuable intellectual property due to their expensive training costs. Consequently, model watermarking, a method developed to protect the intellectual property of neural models, has emerged as a crucial yet underexplored technique. The problem of watermarking PLMs has remained unsolved since the parameters of PLMs will be updated when fine-tuned on downstream datasets, and then embedded watermarks could be removed easily due to the catastrophic forgetting phenomenon. This study investigates the feasibility of watermarking PLMs by embedding backdoors that can be triggered by specific inputs. We employ contrastive learning during the watermarking phase, allowing the representations of specific inputs to be isolated from others and mapped to a particular label after fine-tuning. Moreover, we demonstrate that by combining weight perturbation with the proposed method, watermarks can be embedded in a flatter region of the loss landscape, thereby increasing their robustness to watermark removal. Extensive experiments on multiple datasets demonstrate that the embedded watermarks can be robustly extracted without any knowledge about downstream tasks, and with a high success rate.

14:00-15:30 (East Foyer / Virtual)
**The Interpreter Understands Your Meaning: End-to-end Spoken Language Understanding Aided by Speech Translation**
*Mutian He and Philip N. Garner*
End-to-end spoken language understanding (SLU) remains elusive even with current large pretrained language models on text and speech, especially in multilingual cases. Machine translation has been established as a powerful pretraining objective on text as it enables the model to capture high-level semantics of the input utterance and associations between different languages, which is desired for speech models that work on lower-level acoustic frames. Motivated particularly by the task of cross-lingual SLU, we demonstrate that the task of speech translation (ST) is a good means of pretraining speech models for end-to-end SLU on both intra- and cross-lingual scenarios. By introducing ST, our models reach higher performance over baselines on monolingual and multilingual intent classification as well as spoken question answering using SLURP, MINDS-14, and NMSQA benchmarks. To verify the effectiveness of our methods, we also create new benchmark datasets from both synthetic and real sources, for speech summarization and low-resource/zero-shot transfer from English to French or Spanish. We further show the value of preserving knowledge for the ST pretraining task for better downstream performance, possibly using Bayesian

transfer regularizers.

14:00-15:30 (East Foyer / Virtual)
**Cultural Compass: Predicting Transfer Learning Success in Offensive Language Detection with Cultural Features**
*Li Zhou, Antonia Karamolegkou, Wenyu Chen and Daniel Hershcovich*
The increasing ubiquity of language technology necessitates a shift towards considering cultural diversity in the machine learning realm, particularly for subjective tasks that rely heavily on cultural nuances, such as Offensive Language Detection (OLD). Current understanding underscores that these tasks are substantially influenced by cultural values, however, a notable gap exists in determining if cultural features can accurately predict the success of cross-cultural transfer learning for such subjective tasks. Addressing this, our study delves into the intersection of cultural features and transfer learning effectiveness. The findings reveal that cultural value surveys indeed possess a predictive power for cross-cultural transfer learning success in OLD tasks, and that it can be further improved using offensive word distance. Based on these results, we advocate for the integration of cultural information into datasets. Additionally, we recommend leveraging data sources rich in cultural information, such as surveys, to enhance cultural adaptability. Our research signifies a step forward in the quest for more inclusive, culturally sensitive language technologies.

14:00-15:30 (East Foyer / Virtual)
**Evaluating Parameter-Efficient Finetuning Approaches for Pre-trained Models on the Financial Domain**
*Isabella Olariu, Cedric Lothritz, Jacques Klein, Tegawendé F. Bissyandé, Siwen Guo and Shohreh Haddadan*
Large-scale language models with millions, billions, or trillions of trainable parameters are becoming increasingly popular. However, they risk becoming rapidly over-parameterized and the adaptation cost of fully fine-tuning them increases significantly. Storing them becomes progressively impractical as it requires keeping a separate copy of all the fine-tuned weights for each task. By freezing all pre-trained weights during fine-tuning, parameter-efficient tuning approaches have become an appealing alternative to traditional fine-tuning. The performance of these approaches has been evaluated on common NLP tasks of the GLUE benchmark and shown to match full fine-tuning performance, however, their impact is less researched in domain-specific fields such as finance. This work compares the performance of a set of financial BERT-like models to their fully fine-tuned counterparts by leveraging different parameter-efficient tuning methods. We see that results are comparable to traditional fine-tuning while gaining in time and resource efficiency.

### Industry 2

14:00-15:30 (East Foyer / Virtual)

---

14:00-15:30 (East Foyer / Virtual)
**Angel: Enterprise Search System for the Non-Profit Industry**
*Saiful Haq, Ashutosh Sharma and Pushpak Bhattacharyya*
Non-profit industry need a system for accurately matching fund-seekers (e.g., AMERICAN NATIONAL RED CROSS) with fund-givers (e.g., BILL AND MELINDA GATES FOUNDATION) aligned in cause (e.g., cancer) and target beneficiary group (e.g., children). In this paper, we create an enterprise search system "ANGEL" for the non-profit industry that takes a fund-giver's mission description as input and returns a ranked list of fund-seekers as output, and vice-versa. ANGEL employs ColBERT, a neural information retrieval model, which we enhance by exploiting the two techniques of (a) Syntax-aware local attention (SLA) to combine syntactic information in the mission description with multi-head self-attention and (b) Dense Pseudo Relevance Feedback (DPRF) for augmentation of short mission descriptions. We create a mapping dictionary "non-profit-dict" to curate a "non-profit-search database" containing information on 594K fund-givers and 194K fund-seekers from IRS-990 filings for the non-profit industry search engines . We also curate a "non-profit-evaluation" dataset containing scored matching between 463 fund-givers and 100 fund-seekers. The research is in collaboration with a philanthropic startup that identifies itself as an "AI matching platform, fundraising assistant, and philanthropy search base." Domain experts at the philanthropic startup annotate the non-profit evaluation dataset and continuously evaluate the performance of ANGEL. ANGEL achieves an improvement of 0.14 MAP@10 and 0.16 MRR@10 over the state-of-the-art baseline on the non-profit evaluation dataset. To the best of our knowledge, ours is the first effort at building an enterprise search engine based on neural information retrieval for the non-profit industry.

## Coffee Break

15:30-16:00 - Location: West Foyer

## Session 4: Oral & Poster - 16:00-17:30

### Interpretability, Interactivity, and Analysis of Models for NLP 1

16:00-17:30 (East Ballroom)

---

16:00-16:15 (East Ballroom)
**Dissecting Recall of Factual Associations in Auto-Regressive Language Models**
*Mor Geva, Jasmijn Bastings, Katja Filippova and Amir Globerson*
Transformer-based language models (LMs) are known to capture factual knowledge in their parameters. While previous work looked into where factual associations are stored, only little is known about how they are retrieved internally during inference. We investigate this question through the lens of information flow. Given a subject-relation query, we study how the model aggregates information about the subject and relation to predict the correct attribute. With interventions on attention edges, we first identify two critical points where information propagates to the prediction: one from the relation positions followed by another from the subject positions. Next, by analyzing the information at these points, we unveil a three-step internal mechanism for attribute extraction. First, the representation at the last-subject position goes through an enrichment process, driven by the early MLP sublayers, to encode many subject-related attributes. Second, information from the relation propagates to the prediction. Third, the prediction representation "queries" the enriched subject to extract the attribute. Perhaps surprisingly, this extraction is typically done via attention heads, which often encode subject-attribute mappings in their parameters. Overall, our findings introduce a comprehensive view of how factual associations are stored and extracted internally in LMs, facilitating future research on knowledge localization and editing.

16:15-16:30 (East Ballroom)
**Interpreting Embedding Spaces by Conceptualization**

*Adi Simhi and Shaul Markovitch*
One of the main methods for computational interpretation of a text is mapping it into a vector in some embedding space. Such vectors can then be used for a variety of textual processing tasks. Recently, most embedding spaces are a product of training large language models (LLMs). One major drawback of this type of representation is their incomprehensibility to humans. Understanding the embedding space is crucial for several important needs, including the need to debug the embedding method and compare it to alternatives, and the need to detect biases hidden in the model. In this paper, we present a novel method of understanding embeddings by transforming a latent embedding space into a comprehensible conceptual space. We present an algorithm for deriving a conceptual space with dynamic on-demand granularity. We devise a new evaluation method, using either human rater or LLM-based raters, to show that the conceptualized vectors indeed represent the semantics of the original latent ones. We show the use of our method for various tasks, including comparing the semantics of alternative models and tracing the layers of the LLM. The code is available online https://github.com/adiSimhi/Interpreting-Embedding-Spaces-by-Conceptualization.

16:30-16:45 (East Ballroom)
**Norm of Word Embedding Encodes Information Gain**
*Momose Oyama, Sho Yokoi and Hidetoshi Shimodaira*
Distributed representations of words encode lexical semantic information, but what type of information is encoded and how? Focusing on the skip-gram with negative-sampling method, we found that the squared norm of static word embedding encodes the information gain conveyed by the word; the information gain is defined by the Kullback-Leibler divergence of the co-occurrence distribution of the word to the unigram distribution. Our findings are explained by the theoretical framework of the exponential family of probability distributions and confirmed through precise experiments that remove spurious correlations arising from word frequency. This theory also extends to contextualized word embeddings in language models or any neural networks with the softmax output layer. We also demonstrate that both the KL divergence and the squared norm of embedding provide a useful metric of the informativeness of a word in tasks such as keyword extraction, proper-noun discrimination, and hypernym discrimination.

16:45-17:00 (East Ballroom)
**Assessing Step-by-Step Reasoning against Lexical Negation: A Case Study on Syllogism**
*Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi and Hiroaki Funayama*
Large language models (LLMs) take advantage of step-by-step reasoning instructions, e.g., chain-of-thought (CoT) prompting. Building on this, their ability to perform CoT-style reasoning robustly is of interest from a probing perspective. In this study, we inspect the step-by-step reasoning ability of LLMs with a focus on negation, which is a core linguistic phenomenon that is difficult to process. In particular, we introduce several controlled settings (e.g., reasoning in case of fictional entities) to evaluate the logical reasoning abilities of the models. We observed that dozens of modern LLMs were not robust against lexical negation (e.g., plausible→implausible) when performing CoT-style reasoning, and the results highlight unique limitations in each LLM family.

17:00-17:15 (East Ballroom)
**Can LLMs Facilitate Interpretation of Pre-trained Language Models?**
*Basel Mousi, Nadir Durrani and Fahim Dalvi*
Work done to uncover the knowledge encoded within pre-trained language models rely on annotated corpora or human-in-the-loop methods. However, these approaches are limited in terms of scalability and the scope of interpretation. We propose using a large language model, ChatGPT, as an annotator to enable fine-grained interpretation analysis of pre-trained language models. We discover latent concepts within pre-trained language models by applying agglomerative hierarchical clustering over contextualized representations and then annotate these concepts using ChatGPT. Our findings demonstrate that ChatGPT produces accurate and semantically richer annotations compared to human-annotated concepts. Additionally, we showcase how GPT-based annotations empower interpretation analysis methodologies of which we demonstrate two: probing frameworks and neuron interpretation. To facilitate further exploration and experimentation in the field, we make available a substantial ConceptNet dataset (TCN) comprising 39,000 annotated concepts.

17:15-17:30 (East Ballroom)
**Can You Follow Me? Testing Situational Understanding for ChatGPT**
*Chenghao Yang and Allyson Ettinger*
Understanding sentence meanings and updating information states appropriately across time—what we call "situational understanding" (SU)—is a critical ability for human-like AI agents. SU is essential in particular for chat models, such as ChatGPT, to enable consistent, coherent, and effective dialogue between humans and AI. Previous works have identified certain SU limitations in non-chatbot Large Language models (LLMs), but the extent and causes of these limitations are not well understood, and capabilities of current chat-based models in this domain have not been explored. In this work we tackle these questions, proposing a novel synthetic environment for SU testing which allows us to do controlled and systematic testing of SU in chat-oriented models, through assessment of models' ability to track and enumerate environment states. Our environment also allows for close analysis of dynamics of model performance, to better understand underlying causes for performance patterns. We apply our test to ChatGPT, the state-of-the-art chatbot, and find that despite the fundamental simplicity of the task, the model's performance reflects an inability to retain correct environment states across time. Our follow-up analyses suggest that performance degradation is largely because ChatGPT has non-persistent in-context memory (although it can access the full dialogue history) and it is susceptible to hallucinated updates—including updates that artificially inflate accuracies. Our findings suggest overall that ChatGPT is not currently equipped for robust tracking of situation states, and that trust in the impressive dialogue performance of ChatGPT comes with risks. We release the codebase for reproducing our test environment, as well as all prompts and API responses from ChatGPT, at https://github.com/yangalan123/SituationalTesting.

## Language Grounding to Vision, Robotics and Beyond

16:00-17:30 (Central 1 Ballroom)

16:00-16:15 (Central 1 Ballroom)
**Models See Hallucinations: Evaluating the Factuality in Video Captioning**
*Hui Liu and Xiaojun Wan*
Video captioning aims to describe events in a video with natural language. In recent years, many works have focused on improving captioning models' performance. However, like other text generation tasks, it risks introducing factual errors not supported by the input video. Factual errors can seriously affect the quality of the generated text, sometimes making it completely unusable. Although factual consistency has received much research attention in text-to-text tasks (e.g., summarization), it is less studied in vision-based text generation. In this work, we conduct the first human evaluation of the factuality in video captioning and annotate two factuality datasets. We find that 56% of the model-generated sentences have factual errors, indicating it is a severe problem in this field, but existing evaluation metrics show little correlation with human factuality annotation. We further propose a weakly-supervised, model-based factuality metric FactVC, which outperforms

previous metrics on factuality evaluation of video captioning.

16:15-16:30 (Central 1 Ballroom)
**Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions**
*Bill Noble and Nikolai Ilinykh*
Human speakers can generate descriptions of perceptual concepts, abstracted from the instance-level. Moreover, such descriptions can be used by other speakers to learn provisional representations of those concepts. Learning and using abstract perceptual concepts is under-investigated in the language-and-vision field. The problem is also highly relevant to the field of representation learning in multi-modal NLP. In this paper, we introduce a framework for testing category-level perceptual grounding in multi-modal language models. In particular, we train separate neural networks to **generate** and **interpret** descriptions of visual categories. We measure the *communicative success* of the two models with the zero-shot classification performance of the interpretation model, which we argue is an indicator of perceptual grounding. Using this framework, we compare the performance of *prototype*- and *exemplar*-based representations. Finally, we show that communicative success exposes performance issues in the generation model, not captured by traditional intrinsic NLG evaluation metrics, and argue that these issues stem from a failure to properly ground language in vision at the category level.

16:30-16:45 (Central 1 Ballroom)
**Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms**
*Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi and Youngjae Yu*
Commonsense norms are defeasible by context: reading books is usually great, but not when driving a car. While contexts can be explicitly described in language, in embodied scenarios, contexts are often provided visually. This type of visually grounded reasoning about defeasible commonsense norms is generally easy for humans, but (as we show) poses a challenge for machines, as it necessitates both visual understanding and reasoning about commonsense norms. We construct a new multimodal benchmark for studying commonsense norms: NormLens. NormLens consists of 10K human judgments accompanied by free-form explanations covering 2K multimodal situations, and serves as a probe to address two questions: (1) to what extent can models align with average human judgment? and (2) how well can models explain their predicted judgments? We find that state-of-the-art model judgments and explanations are not well-aligned with human annotation. Additionally, we present a simple yet effective approach to better align models with humans by distilling social commonsense knowledge from large language models. The data and code will be released.

16:45-17:00 (Central 1 Ballroom)
**Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models**
*Joan Nwatu, Oana Ignat and Rada Mihalcea*
Despite the impressive performance of current AI models reported across various tasks, performance reports often do not include evaluations of how these models perform on the specific groups that will be impacted by these technologies. Among the minority groups under-represented in AI, data from low-income households are often overlooked in data collection and model evaluation. We evaluate the performance of a state-of-the-art vision-language model (CLIP) on a geo-diverse dataset containing household images associated with different income values (DollarStreet) and show that performance inequality exists among households of different income levels. Our results indicate that performance for the poorer groups is consistently lower than the wealthier groups across various topics and countries. We highlight insights that can help mitigate these issues and propose actionable steps for economic-level inclusive AI development.

17:00-17:15 (Central 1 Ballroom)
**3DRP-Net: 3D Relative Position-aware Network for 3D Visual Grounding**
*Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin and Zhou Zhao*
3D visual grounding aims to localize the target object in a 3D point cloud by a free-form language description. Typically, the sentences describing the target object tend to provide information about its relative relation between other objects and its position within the whole scene. In this work, we propose a relation-aware one-stage framework, named 3D Relative Position-aware Network (3DRP-Net), which can effectively capture the relative spatial relationships between objects and enhance object attributes. Specifically, 1) we propose a 3D Relative Position Multi-head Attention (3DRP-MA) module to analyze relative relations from different directions in the context of object pairs, which helps the model to focus on the specific object relations mentioned in the sentence. 2) We designed a soft-labeling strategy to alleviate the spatial ambiguity caused by redundant points, which further stabilizes and enhances the learning process through a constant and discriminative distribution. Extensive experiments conducted on three benchmarks (i.e., ScanRefer and Nr3D/Sr3D) demonstrate that our method outperforms all the state-of-the-art methods in general.

17:15-17:30 (Central 1 Ballroom)
**Localizing Active Objects from Egocentric Vision with Symbolic World Knowledge**
*Te-Lin Wu, Yu Zhou and Nanyun Peng*
The ability to actively ground task instructions from an egocentric view is crucial for AI agents to accomplish tasks or assist humans virtually. One important step towards this goal is to localize and track key active objects that undergo major state change as a consequence of human actions/interactions to the environment without being told exactly what/where to ground (e.g., localizing and tracking the ‘sponge‘ in video from the instruction "Dip the sponge into the bucket."). While existing works approach this problem from a pure vision perspective, we investigate to which extent the textual modality (i.e., task instructions) and their interaction with visual modality can be beneficial. Specifically, we propose to improve phrase grounding models’ ability on localizing the active objects by: (1) learning the role of ‘objects undergoing change‘ and extracting them accurately from the instructions, (2) leveraging pre- and post-conditions of the objects during actions, and (3) recognizing the objects more robustly with descriptional knowledge. We leverage large language models (LLMs) to extract the aforementioned action-object knowledge, and design a per-object aggregation masking technique to effectively perform joint inference on object phrases and symbolic knowledge. We evaluate our framework on Ego4D and Epic-Kitchens datasets. Extensive experiments demonstrate the effectiveness of our proposed framework, which leads to>54% improvements in all standard metrics on the TREK-150-OPE-Det localization + tracking task, >7% improvements in all standard metrics on the TREK-150-OPE tracking task, and >3% improvements in average precision (AP) on the Ego4D SCOD task.

## Language Modeling and Analysis of Language Models 1

16:00-17:30 (Central 3 Ballroom)

16:00-16:15 (Central 3 Ballroom)
**FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation**
*Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer and Hannaneh Hajishirzi*

Evaluating the factuality of long-form text generated by large language models (LMs) is non-trivial because (1) generations often contain a mixture of supported and unsupported pieces of information, making binary judgments of quality inadequate, and (2) human evaluation is time-consuming and costly. In this paper, we introduce FACTSCORE, a new evaluation that breaks a generation into a series of atomic facts and computes the percentage of atomic facts supported by a reliable knowledge source. We conduct an extensive human evaluation to obtain FACTSCOREs of people biographies generated by several state-of-the-art commercial LMs—InstructGPT, ChatGPT, and the retrieval-augmented PerplexityAI—and report new analysis demonstrating the need for such a fine-grained score (e.g., ChatGPT only achieves 58%). Since human evaluation is costly, we also introduce an automated model that estimates FACTSCORE using retrieval and a strong language model, with less than a 2% error rate. Finally, we use this automated metric to evaluate 6,500 generations from a new set of 13 recent LMs that would have cost $26K if evaluated by humans, with various findings: GPT-4 and ChatGPT are more factual than public models, and Vicuna and Alpaca are some of the best public models. FACTSCORE is available for public use via 'pip install factscore'.

16:15-16:30 (Central 3 Ballroom)
**ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing**
*Nam Quoc Nguyen, Thang Chau Phan, Duc-Vu Nguyen and Kiet Van Nguyen*
English and Chinese, known as resource-rich languages, have witnessed the strong development of transformer-based language models for natural language processing tasks. Although Vietnam has approximately 100M people speaking Vietnamese, several pre-trained models, e.g., PhoBERT, ViBERT, and vELECTRA, performed well on general Vietnamese NLP tasks, including POS tagging and named entity recognition. These pre-trained language models are still limited to Vietnamese social media tasks. In this paper, we present the first monolingual pre-trained language model for Vietnamese social media texts, ViSoBERT, which is pre-trained on a large-scale corpus of high-quality and diverse Vietnamese social media texts using XLM-R architecture. Moreover, we explored our pre-trained model on five important natural language downstream tasks on Vietnamese social media texts: emotion recognition, hate speech detection, sentiment analysis, spam reviews detection, and hate speech spans detection. Our experiments demonstrate that ViSoBERT, with far fewer parameters, surpasses the previous state-of-the-art models on multiple Vietnamese social media tasks. Our ViSoBERT model is available only for research purposes. Disclaimer: This paper contains actual comments on social networks that might be construed as abusive, offensive, or obscene.

16:30-16:45 (Central 3 Ballroom)
**Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models**
*Kaitlyn Zhou, Dan Jurafsky and Tatsunori Hashimoto*
The increased deployment of LMs for real-world tasks involving knowledge and facts makes it important to understand model epistemology: what LMs think they know, and how their attitudes toward that knowledge are affected by language use in their inputs. Here, we study an aspect of model epistemology: how epistemic markers of certainty, uncertainty, or evidentiality like "I'm sure it's", "I think it's", or "Wikipedia says it's" affect models, and whether they contribute to model failures. We develop a typology of epistemic markers and inject 50 markers into prompts for question answering. We find that LMs are highly sensitive to epistemic markers in prompts, with accuracies varying more than 80%. Surprisingly, we find that expressions of high certainty result in a 7% decrease in accuracy as compared to low certainty expressions; similarly, factive verbs hurt performance, while evidentials benefit performance. Our analysis of a popular pretraining dataset shows that these markers of uncertainty are associated with answers on question-answering websites, while markers of certainty are associated with questions. These associations may suggest that the behavior of LMs is based on mimicking observed language use, rather than truly reflecting epistemic uncertainty.

16:45-17:00 (Central 3 Ballroom)
**CodeT5+: Open Code Large Language Models for Code Understanding and Generation**
*Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D. Q. Bui, Junnan Li and Steven Hoi*
Large language models (LLMs) pretrained on vast source code have achieved prominent progress in code intelligence. However, existing code LLMs have two main limitations. First, they often adopt a specific architecture (encoder-only or decoder-only) or rely on a unified encoder-decoder network for different downstream tasks, lacking the flexibility to operate in the optimal architecture for a specific task. Secondly, they often employ a limited set of pretraining objectives which might not be relevant to some tasks and hence result in substantial performance degrade. To address these limitations, we propose "CodeT5+", a family of encoder-decoder LLMs for code in which component modules can be flexibly combined to suit a wide range of code tasks. Such flexibility is enabled by our proposed mixture of pretraining objectives, which cover span denoising, contrastive learning, text-code matching, and causal LM pretraining tasks, on both unimodal and bimodal multilingual code corpora. Furthermore, we propose to initialize CodeT5+ with frozen off-the-shelf LLMs without training from scratch to efficiently scale up our models, and explore instruction-tuning to align with natural language instructions. We extensively evaluate CodeT5+ on over 20 code-related benchmarks in different settings, including zero-shot, finetuning, and instruction-tuning. We observe state-of-the-art (SoTA) performance on various code-related tasks, and our instruction-tuned CodeT5+ 16B achieves new SoTA results of 35.0% pass@1 and 54.5% pass@10 on the HumanEval code generation task against other open code LLMs, even surpassing the OpenAI code-cushman-001 model.

17:00-17:15 (Central 3 Ballroom)
**MAGNIFICo: Evaluating the In-Context Learning Ability of Large Language Models to Generalize to Novel Interpretations**
*Arkil Patel, Satwik Bhattamishra, Siva Reddy and Dzmitry Bahdanau*
Humans possess a remarkable ability to assign novel interpretations to linguistic expressions, enabling them to learn new words and understand community-specific connotations. However, Large Language Models (LLMs) have a knowledge cutoff and are costly to finetune repeatedly. Therefore, it is crucial for LLMs to learn novel interpretations in-context. In this paper, we systematically analyse the ability of LLMs to acquire novel interpretations using in-context learning. To facilitate our study, we introduce MAGNIFICo, an evaluation suite implemented within a text-to-SQL semantic parsing framework that incorporates diverse tokens and prompt settings to simulate real-world complexity. Experimental results on MAGNIFICo demonstrate that LLMs exhibit a surprisingly robust capacity for comprehending novel interpretations from natural language descriptions as well as from discussions within long conversations. Nevertheless, our findings also highlight the need for further improvements, particularly when interpreting unfamiliar words or when composing multiple novel interpretations simultaneously in the same example. Additionally, our analysis uncovers the semantic predispositions in LLMs and reveals the impact of recency bias for information presented in long contexts.

17:15-17:30 (Central 3 Ballroom)
**CombLM: Adapting Black-Box Language Models through Small Fine-Tuned Models**
*Aitor Ormazabal, Mikel Artetxe and Eneko Agirre*
Methods for adapting language models (LMs) to new tasks and domains have traditionally assumed white-box access to the model, and work by modifying its parameters. However, this is incompatible with a recent trend in the field, where the highest quality models are only available as black-boxes through inference APIs. Even when the model weights are available, the computational cost of fine-tuning large LMs can be prohibitive for most practitioners. In this work, we present a lightweight method for adapting large LMs to new domains and tasks, assuming no access to their weights or intermediate activations. Our approach fine-tunes a small white-box LM and combines it with the large black-box LM at the probability level through a small network, learned on a small validation set. We validate our approach by adapting a large LM (OPT-30B) to several domains and a downstream task (machine translation), observing improved performance in all cases, of up to 9%, while

using a domain expert 23x smaller.

## Information Retrieval and Text Mining

16:00-17:30 (West 1 Ballroom)

16:00-16:15 (West 1 Ballroom)
**Hybrid Inverted Index Is a Robust Accelerator for Dense Retrieval**
*Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou and Jing Yao*
Inverted file structure is a common technique for accelerating dense retrieval. It clusters documents based on their embeddings; during searching, it probes nearby clusters w.r.t. an input query and only evaluates documents within them by subsequent codecs, thus avoiding the expensive cost from exhaustive traversal. However, the clustering is always lossy, which results in the miss of relevant documents in the probed clusters and hence degrades retrieval quality. In contrast, lexical matching, such as overlaps of salient terms, tend to be strong features for identifying relevant documents. In this work, we present the Hybrid Inverted Index ($HI^2$), where the embedding clusters and salient terms work collaboratively to accelerate dense retrieval. To make best of both effectiveness and efficiency, we devise a cluster selector and a term selector, to construct compact inverted lists and efficiently searching through them. Moreover, we leverage simple unsupervised algorithms as well as end-to-end knowledge distillation to learn these two modules, with the latter further boosting the effectiveness. Based on comprehensive experiments on popular retrieval benchmarks, we verify that clusters and terms indeed complement each other, enabling $HI^2$ to achieve lossless retrieval quality with competitive efficiency across a variety of index settings.

16:15-16:30 (West 1 Ballroom)
**Robust Prompt Optimization for Large Language Models Against Distribution Shifts**
*Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang and Tat-Seng Chua*
Large Language Model (LLM) has demonstrated significant ability in various Natural Language Processing tasks. However, their effectiveness is highly dependent on the phrasing of the task prompt, leading to research on automatic prompt optimization using labeled task data. We reveal that these prompt optimization techniques are vulnerable to distribution shifts such as subpopulation shifts, which are common for LLMs in real-world scenarios such as customer reviews analysis. In this light, we propose a new problem of robust prompt optimization for LLMs against distribution shifts, which requires the prompt optimized over the labeled source group can simultaneously generalize to an unlabeled target group. To solve this problem, we propose Generalized Prompt Optimization framework , which incorporates the unlabeled data from the target group into prompt optimization. Extensive experimental results demonstrate the effectiveness of the proposed framework with significant performance improvement on the target group and comparable performance on the source group.

16:30-16:45 (West 1 Ballroom)
**WSDMS: Debunk Fake News via Weakly Supervised Detection of Misinforming Sentences with Contextualized Social Wisdom**
*Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin and Zhiwei Yang*
Fake news debunking primarily focuses on determining the truthfulness of news articles, which oversimplifies the issue as fake news often combines elements of both truth and falsehood. Thus, it becomes crucial to identify specific instances of misinformation within the articles. In this research, we investigate a novel task in the field of fake news debunking, which involves detecting sentence-level misinformation. One of the major challenges in this task is the absence of a training dataset with sentence-level annotations regarding veracity. Inspired by the Multiple Instance Learning (MIL) approach, we propose a model called Weakly Supervised Detection of Misinforming Sentences (WSDMS). This model only requires bag-level labels for training but is capable of inferring both sentence-level misinformation and article-level veracity, aided by relevant social media conversations that are attentively contextualized with news sentences. We evaluate WSDMS on three real-world benchmarks and demonstrate that it outperforms existing state-of-the-art baselines in debunking fake news at both the sentence and article levels.

16:45-17:00 (West 1 Ballroom)
**Learning to Describe for Predicting Zero-shot Drug-Drug Interactions**
*Fangqi Zhu, Yongqi Zhang, Lei Chen, Bing Qin and Ruifeng Xu*
Adverse drug-drug interactions (DDIs) can compromise the effectiveness of concurrent drug administration, posing a significant challenge in healthcare. As the development of new drugs continues, the potential for unknown adverse effects resulting from DDIs becomes a growing concern. Traditional computational methods for DDI prediction may fail to capture interactions for new drugs due to the lack of knowledge. In this paper, we introduce a new problem setup as zero-shot DDI prediction that deals with the case of new drugs. Leveraging textual information from online databases like DrugBank and PubChem, we propose an innovative approach TextDDI with a language model-based DDI predictor and a reinforcement learning (RL)-based information selector, enabling the selection of concise and pertinent text for accurate DDI prediction on new drugs. Empirical results show the benefits of the proposed approach on several settings including zero-shot and few-shot DDI prediction, and the selected texts are semantically relevant. Our code and data are available at https://github.com/zhufq00/DDIs-Prediction.

17:00-17:15 (West 1 Ballroom)
**Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents**
*Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin and Zhaochun Ren*
Large Language Models (LLMs) have demonstrated remarkable zero-shot generalization across various language-related tasks, including search engines. However, existing work utilizes the generative ability of LLMs for Information Retrieval (IR) rather than direct passage ranking. The discrepancy between the pre-training objectives of LLMs and the ranking objective poses another challenge. In this paper, we first investigate generative LLMs such as ChatGPT and GPT-4 for relevance ranking in IR. Surprisingly, our experiments reveal that properly instructed LLMs can deliver competitive, even superior results to state-of-the-art supervised methods on popular IR benchmarks. Furthermore, to address concerns about data contamination of LLMs, we collect a new test set called NovelEval, based on the latest knowledge and aiming to verify the model's ability to rank unknown knowledge. Finally, to improve efficiency in real-world applications, we delve into the potential for distilling the ranking capabilities of ChatGPT into small specialized models using a permutation distillation scheme. Our evaluation results turn out that a distilled 440M model outperforms a 3B supervised model on the BEIR benchmark. The code to reproduce our results is available at www.github.com/sunnweiwei/RankGPT.

17:15-17:30 (West 1 Ballroom)
**Goal-Driven Explainable Clustering via Language Descriptions**
*Zihan Wang, Jingbo Shang and Ruiqi Zhong*
Unsupervised clustering is widely used to explore large corpora, but existing formulations neither consider the users' goals nor explain clus-

ters' meanings. We propose a new task formulation, "Goal-Driven Clustering with Explanations" (GoalEx), which represents both the goal and the explanations as free-form language descriptions. For example, to categorize the errors made by a summarization system, the input to GoalEx is a corpus of annotator-written comments for system-generated summaries and a goal description "cluster the comments based on why the annotators think the summary is imperfect."; the outputs are text clusters each with an explanation ("this cluster mentions that the summary misses important context information."), which relates to the goal and accurately explains which comments should (not) belong to a cluster. To tackle GoalEx, we prompt a language model with "[corpus subset] + [goal] + Brainstorm a list of explanations each representing a cluster."; then we classify whether each sample belongs to a cluster based on its explanation; finally, we use integer linear programming to select a subset of candidate clusters to cover most samples while minimizing overlaps. Under both automatic and human evaluation on corpora with or without labels, our method produces more accurate and goal-related explanations than prior methods.

## Linguistic Theories, Cognitive Modeling and Psycholinguistics

16:00-17:30 (West 2 Ballroom)

16:00-16:15 (West 2 Ballroom)
**Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives**
*Mario Giulianelli, Sarenne Wallbridge and Raquel Fernández*
We present information value, a measure which quantifies the predictability of an utterance relative to a set of plausible alternatives. We introduce a method to obtain interpretable estimates of information value using neural text generators, and exploit their psychometric predictive power to investigate the dimensions of predictability that drive human comprehension behaviour. Information value is a stronger predictor of utterance acceptability in written and spoken dialogue than aggregates of token-level surprisal and it is complementary to surprisal for predicting eye-tracked reading times.

16:15-16:30 (West 2 Ballroom)
**Addressing Linguistic Bias through a Contrastive Analysis of Academic Writing in the NLP Domain**
*Robert Ridley, Zhen Wu, Jianbing Zhang, Shujian Huang and Xinyu Dai*
It has been well documented that a reviewer's opinion of the nativeness of expression in an academic paper affects the likelihood of it being accepted for publication. Previous works have also shone a light on the stress and anxiety authors who are non-native English speakers experience when attempting to publish in international venues. We explore how this might be a concern in the field of Natural Language Processing (NLP) through conducting a comprehensive statistical analysis of NLP paper abstracts, identifying how authors of different linguistic backgrounds differ in the lexical, morphological, syntactic and cohesive aspects of their writing. Through our analysis, we identify that there are a number of characteristics that are highly variable across the different corpora examined in this paper. This indicates potential for the presence of linguistic bias. Therefore, we outline a set of recommendations to publishers of academic journals and conferences regarding their guidelines and resources for prospective authors in order to help enhance inclusivity and fairness.

16:30-16:45 (West 2 Ballroom)
**The neural dynamics of word recognition and integration**
*Jon Gauthier and Roger P. Levy*
Listeners recognize and integrate words in rapid and noisy everyday speech by combining expectations about upcoming content with incremental sensory evidence. We present a computational model of word recognition which formalizes this perceptual process in Bayesian decision theory. We fit this model to explain scalp EEG signals recorded as subjects passively listened to a fictional story, revealing both the dynamics of the online auditory word recognition process and the neural correlates of the recognition and integration of words. The model reveals distinct neural processing of words depending on whether or not they can be quickly recognized. While all words trigger a neural response characteristic of probabilistic integration — voltage modulations predicted by a word's surprisal in context — these modulations are amplified for words which require more than roughly 150 ms of input to be recognized. We observe no difference in the latency of these neural responses according to words' recognition times. Our results support a two-part model of speech comprehension, combining an eager and rapid process of word recognition with a temporally independent process of word integration. However, we also developed alternative models of the scalp EEG signal not incorporating word recognition dynamics which showed similar performance improvements. We discuss potential future modeling steps which may help to separate these hypotheses.

16:45-17:00 (West 2 Ballroom)
**Quantifying the redundancy between prosody and text**
*Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox and Tamar I Regev*
Prosody—the suprasegmental component of speech, including pitch, loudness, and tempo—carries critical aspects of meaning. However, the relationship between the information conveyed by prosody vs. by the words themselves remains poorly understood. We use large language models (LLMs) to estimate how much information is redundant between prosody and the words themselves. Using a large spoken corpus of English audiobooks, we extract prosodic features aligned to individual words and test how well they can be predicted from LLM embeddings, compared to non-contextual word embeddings. We find a high degree of redundancy between the information carried by the words and prosodic information across several prosodic features, including intensity, duration, pauses, and pitch contours. Furthermore, a word's prosodic information is redundant with both the word itself and the context preceding as well as following it. Still, we observe that prosodic features can not be fully predicted from text, suggesting that prosody carries information above and beyond the words. Along with this paper, we release a general-purpose data processing pipeline for quantifying the relationship between linguistic information and extra-linguistic features.

17:00-17:15 (West 2 Ballroom)
**Interpreting and Exploiting Functional Specialization in Multi-Head Attention under Multi-task Learning**
*Chong Li, Shaonan Wang, Yunhao Zhang, Jiajun Zhang and Chengqing Zong*
Transformer-based models, even though achieving super-human performance on several downstream tasks, are often regarded as a black box and used as a whole. It is still unclear what mechanisms they have learned, especially their core module: multi-head attention. Inspired by functional specialization in the human brain, which helps to efficiently handle multiple tasks, this work attempts to figure out whether the multi-head attention module will evolve similar function separation under multi-tasking training. If it is, can this mechanism further improve the model performance? To investigate these questions, we introduce an interpreting method to quantify the degree of functional specialization in multi-head attention. We further propose a simple multi-task training method to increase functional specialization and mitigate negative information transfer in multi-task learning. Experimental results on seven pre-trained transformer models have demonstrated that multi-head attention does evolve functional specialization phenomenon after multi-task training which is affected by the similarity of tasks. Moreover, the multi-task training strategy based on functional specialization boosts performance in both multi-task learning and transfer learning without adding any parameters.

17:15-17:30 (West 2 Ballroom)
**Assessing the influence of attractor-verb distance on grammatical agreement in humans and language models**
*Christos Nikolaos Zacharopoulos, Théo Desbordes and Mathias Sablé-Meyer*
Subject-verb agreement in the presence of an attractor noun located between the main noun and the verb elicits complex behavior: judgments of grammaticality are modulated by the grammatical features of the attractor. For example, in the sentence *"The girl near the boys likes climbing"*, the attractor (*boys*) disagrees in grammatical number with the verb (*likes*), creating a locally implausible transition probability. Here, we parametrically modulate the distance between the attractor and the verb while keeping the length of the sentence equal. We evaluate the performance of both humans and two artificial neural network models: both make more mistakes when the attractor is closer to the verb, but neural networks get close to the chance level while humans are mostly able to overcome the attractor interference. Additionally, we report a linear effect of attractor distance on reaction times. We hypothesize that a possible reason for the proximity effect is the calculation of transition probabilities between adjacent words. Nevertheless, classical models of attraction such as the cue-based model might suffice to explain this phenomenon, thus paving the way for new research. Data and analyses available at https://osf.io/d4g6k

## Dialogue and Interactive Systems 2

16:00-17:30 (West 3 Ballroom)

---

16:00-16:15 (West 3 Ballroom)
**Just Adjust One Prompt: Enhancing In-Context Dialogue Scoring via Constructing the Optimal Subgraph of Demonstrations and Prompts**
*Jiashu Pu, Ling Cheng, Lu Fan, Tangjie Lv and Rongsheng Zhang*
The use of modern Large Language Models (LLMs) as chatbots still has some problems such as hallucinations and lack of empathy. Identifying these issues can help improve chatbot performance. The community has been continually iterating on reference-free dialogue evaluation methods based on large language models (LLMs) that can be readily applied. However, many of these LLM-based metrics require selecting specific datasets and developing specialized training tasks for different evaluation dimensions (e.g., coherence, informative). The developing step can be time-consuming and may need to be repeated for new evaluation dimensions. To enable efficient and flexible adaptation to diverse needs of dialogue evaluation, we propose a dimension-agnostic scoring method that leverages the in-context learning (ICL) capability of LLMs to learn from human scoring to the fullest extent. Our method has three key features. To begin with, rather than manual prompt crafting, we propose automatically generating prompts, allowing the LLM to observe human labels and summarize the most suitable prompt. Additionally, since the LLM has a token limit and ICL is sensitive to demonstration variations, we train a selector to finely customize demonstrations and prompts for each dialogue input. Finally, during inference, we propose to request the LLM multiple times with a subgraph of demonstrations and prompts that are diverse and suitable to maximize ICL from various human scoring. We validate the efficacy of our method on five datasets, even with a small amount of annotated data, our method outperforms all strong baselines. Code is available at https://github.com/iamlxb3/EMNLP2023-ADOROR.

16:15-16:30 (West 3 Ballroom)
**An "Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives"**
*Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc and Sharath Chandra Guntuku*
Mental health conversational agents (a.k.a. chatbots) are widely studied for their potential to offer accessible support to those experiencing mental health challenges. Previous surveys on the topic primarily consider papers published in either computer science or medicine, leading to a divide in understanding and hindering the sharing of beneficial knowledge between both domains. To bridge this gap, we conduct a comprehensive literature review using the PRISMA framework, reviewing 534 papers published in both computer science and medicine. Our systematic review reveals 136 key papers on building mental health-related conversational agents with diverse characteristics of modeling and experimental design techniques. We find that computer science papers focus on LLM techniques and evaluating response quality using automated metrics with little attention to the application while medical papers use rule-based conversational agents and outcome metrics to measure the health outcomes of participants. Based on our findings on transparency, ethics, and cultural heterogeneity in this review, we provide a few recommendations to help bridge the disciplinary divide and enable the cross-disciplinary development of mental health conversational agents.

16:30-16:45 (West 3 Ballroom)
**From Multilingual Complexity to Emotional Clarity: Leveraging Commonsense to Unveil Emotions in Code-Mixed Dialogues**
*Shivani Kumar, Ramaneswaran S, Md Shad Akhtar and Tanmoy Chakraborty*
Understanding emotions during conversation is a fundamental aspect of human communication, driving NLP research for Emotion Recognition in Conversation (ERC). While considerable research has focused on discerning emotions of individual speakers in monolingual dialogues, understanding the emotional dynamics in code-mixed conversations has received relatively less attention. This motivates our undertaking of ERC for code-mixed conversations in this study. Recognizing that emotional intelligence encompasses a comprehension of worldly knowledge, we propose an innovative approach that integrates commonsense information with dialogue context to facilitate a deeper understanding of emotions. To achieve this, we devise an efficient pipeline that extracts relevant commonsense from existing knowledge graphs based on the code-mixed input. Subsequently, we develop an advanced fusion technique that seamlessly combines the acquired commonsense information with the dialogue representation obtained from a dedicated dialogue understanding module. Our comprehensive experimentation showcases the substantial performance improvement obtained through the systematic incorporation of commonsense in ERC. Both quantitative assessments and qualitative analyses further corroborate the validity of our hypothesis, reaffirming the pivotal role of commonsense integration in enhancing ERC.

16:45-17:00 (West 3 Ballroom)
**e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts**
*Kshitij Mishra, Priyanshu Priya, Manisha Burja and Asif Ekbal*
The shortage of therapists for mental health patients emphasizes the importance of globally accessible dialogue systems alleviating their issues. To have effective interpersonal psychotherapy, these systems must exhibit politeness and empathy when needed. However, these factors may vary as per the user's gender, age, persona, and sentiment. Hence, in order to establish trust and provide a personalized cordial experience, it is essential that generated responses should be tailored to individual profiles and attributes. Focusing on this objective, we propose e-THERAPIST, a novel polite interpersonal psychotherapy dialogue system to address issues like depression, anxiety, schizophrenia, etc. We begin by curating a unique conversational dataset for psychotherapy, called PsyCon. It is annotated at two levels: (i) dialogue-level - including user's profile information (gender, age, persona) and therapist's psychotherapeutic approach; and (ii) utterance-level - encompassing user's sentiment and therapist's politeness, and interpersonal behaviour. Then, we devise a novel reward model to adapt correct polite interpersonal behaviour and use it to train e-THERAPIST on PsyCon employing NLPO loss. Our extensive empirical analysis validates the effectiveness

of each component of the proposed e-THERAPIST demonstrating its potential impact in psychotherapy settings.

17:00-17:15 (West 3 Ballroom)
**ReSee: Responding through Seeing Fine-grained Visual Knowledge in Open-domain Dialogue**
*Haoqin Tu, Yitong Li, Fei Mi and Zhongliang Yang*
Incorporating visual knowledge into text-only dialogue systems has become a potential direction to imitate the way humans think, imagine, and communicate. However, existing multimodal dialogue systems are either confined by the scale and quality of available datasets or the coarse concept of visual knowledge. To address these issues, we provide a new paradigm of constructing multimodal dialogues as well as two datasets extended from text-only dialogues under such paradigm (ReSee-WoW, ReSee-DD). We propose to explicitly split the visual knowledge into finer granularity ("turn-level" and "entity-level"). To further boost the accuracy and diversity of augmented visual information, we retrieve them from the Internet or a large image dataset. To demonstrate the superiority and universality of the provided visual knowledge, we propose a simple but effective framework ReSee to add visual representation into vanilla dialogue models by modality concatenations. We also conduct extensive experiments and ablations w.r.t. different model configurations and visual knowledge settings. Empirical, encouraging results not only demonstrate the effectiveness of introducing visual knowledge at both entity and turn level but also verify the proposed model ReSee outperforms several state-of-the-art methods on automatic and human evaluations. By leveraging text and vision knowledge, ReSee can produce informative responses with real-world visual concepts. Our code is available at https://github.com/ImKeTT/ReSee.

17:15-17:30 (West 3 Ballroom)
**PK-ICR: Persona-Knowledge Interactive Multi-Context Retrieval for Grounded Dialogue**
*Minsik Oh, Joosung Lee, Jiwei Li and Guoyin Wang*
Identifying relevant persona or knowledge for conversational systems is critical to grounded dialogue response generation. However, each grounding has been mostly researched in isolation with more practical multi-context dialogue tasks introduced in recent works. We define Persona and Knowledge Dual Context Identification as the task to identify persona and knowledge jointly for a given dialogue, which could be of elevated importance in complex multi-context dialogue settings. We develop a novel grounding retrieval method that utilizes all contexts of dialogue simultaneously. Our method requires less computational power via utilizing neural QA retrieval models. We further introduce our novel null-positive rank test which measures ranking performance on semantically dissimilar samples (i.e. hard negatives) in relation to data augmentation.

## Demo session 3

16:00-17:30 (East Foyer / Virtual)

16:00-17:30 (East Foyer / Virtual)
**RobustQA: A Framework for Adversarial Text Generation Analysis on Question Answering Systems**
*Yasaman Boreshban, Seyed Morteza Mirbostani, Seyedeh Fatemeh Ahmadi, Gita Shojaee, Fatemeh Kamani, Gholamreza Ghassem-Sani and Seyed Abolghasem Mirroshandel*
Question answering (QA) systems have reached human-level accuracy; however, these systems are not robust enough and are vulnerable to adversarial examples. Recently, adversarial attacks have been widely investigated in text classification. However, there have been few research efforts on this topic in QA. In this article, we have modified the attack algorithms widely used in text classification to fit those algorithms for QA systems. We have evaluated the impact of various attack methods on QA systems at character, word, and sentence levels. Furthermore, we have developed a new framework, named RobustQA, as the first open-source toolkit for investigating textual adversarial attacks in QA systems. RobustQA consists of seven modules: Tokenizer, Victim Model, Goals, Metrics, Attacker, Attack Selector, and Evaluator. It currently supports six different attack algorithms. Furthermore, the framework simplifies the development of new attack algorithms in QA. The source code and documentation of RobustQA are available at https://github.com/mirbostani/RobustQA.

16:00-17:30 (East Foyer / Virtual)
**Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback**
*Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi and Thien Nguyen*
A key technology for large language models (LLMs) involves instruction tuning that helps align the models' responses with human expectations to realize impressive learning abilities. Two major approaches for instruction tuning characterize supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), which are applied to produce the best commercial LLMs. To improve the accessibility of LLMs, various instruction-tuned open-source LLMs have also been introduced recently. However, existing open-source LLMs have only been instruction-tuned for English and a few popular languages, thus hindering their accessibility to many other languages in the world. In addition, SFT has been used as the only approach to instruction-tune open-source LLMs for multiple languages. This has left a significant gap for fine-tuned LLMs based on RLHF in diverse languages and raised important questions on how RLHF can boost the performance of multilingual instruction tuning. To overcome this issue, we present Okapi, the first system with instruction-tuned LLMs based on RLHF for multiple languages. Okapi introduces instruction and response-ranked data in 26 diverse languages to facilitate the experiments and development of future multilingual LLM research. We also present benchmark datasets to enable the evaluation of generative LLMs in multiple languages. Our experiments demonstrate the advantages of RLHF for multilingual instruction over SFT for different base models and datasets. Our framework with created resources, fine-tuned LLMs, interaction scripts are released at https://github.com/nlp-uoregon/Okapi. A demo video to show our framework can also be found at: https://youtu.be/QFV2fkPwvi0.

16:00-17:30 (East Foyer / Virtual)
**SAGEViz: SchemA GEneration and Visualization**
*Sugam Devare, Mahnaz Koupaee, Gautham Gunapati, Sayontan Ghosh, Sai Vallurupalli, Yash Kumar Lal, Francis Ferraro, Nathanael Chambers, Greg Durrett, Raymond Mooney, Katrin Erk and Niranjan Balasubramanian*
Schema induction involves creating a graph representation depicting how events unfold in a scenario. We present SAGEViz, an intuitive and modular tool that utilizes human-AI collaboration to create and update complex schema graphs efficiently, where multiple annotators (humans and models) can work simultaneously on a schema graph from any domain. The tool consists of two components: (1) a curation component powered by plug-and-play event language models to create and expand event sequences while human annotators validate and enrich the sequences to build complex hierarchical schemas, and (2) an easy-to-use visualization component to visualize schemas at varying levels of hierarchy. Using supervised and few-shot approaches, our event language models can continually predict relevant events starting from a seed event. We conduct a user study and show that users need less effort in terms of interaction steps with SAGEViz to generate schemas of better quality. We also include a video demonstrating the system.

16:00-17:30 (East Foyer / Virtual)
**Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation**

*David Heineman, Yao Dou and Wei Xu*
Fine-grained, span-level human evaluation has emerged as a reliable and robust method for evaluating text generation tasks such as summarization, simplification, machine translation and news generation, and the derived annotations have been useful for training automatic metrics and improving language models. However, existing annotation tools implemented for these evaluation frameworks lack the adaptability to be extended to different domains or languages, or modify annotation settings according to user needs; and, the absence of a unified annotated data format inhibits the research in multi-task learning. In this paper, we introduce Thresh, a unified, customizable and deployable platform for fine-grained evaluation. With a single YAML configuration file, users can build and test an annotation interface for any framework within minutes – all in one web browser window. To facilitate collaboration and sharing, Thresh provides a community hub that hosts a collection of fine-grained frameworks and corresponding annotations made and collected by the community, covering a wide range of NLP tasks. For deployment, Thresh offers multiple options for any scale of annotation projects from small manual inspections to large crowdsourcing ones. Additionally, we introduce a Python library to streamline the entire process from typology design and deployment to annotation processing. Thresh is publicly accessible at https://thresh.tools.

16:00-17:30 (East Foyer / Virtual)
**InsightPilot: An LLM-Empowered Automated Data Exploration System**
*Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han and Dongmei Zhang*
Exploring data is crucial in data analysis, as it helps users understand and interpret the data more effectively. However, performing effective data exploration requires in-depth knowledge of the dataset, the user intent and expertise in data analysis techniques. Not being familiar with either can create obstacles that make the process time-consuming and overwhelming. To address this issue, we introduce InsightPilot, an LLM (Large Language Model)-based, automated data exploration system designed to simplify the data exploration process. InsightPilot features a set of carefully designed analysis actions that streamline the data exploration process. Given a natural language question, InsightPilot collaborates with the LLM to issue a sequence of analysis actions, explore the data and generate insights. We demonstrate the effectiveness of InsightPilot in a user study and a case study, showing how it can help users gain valuable insights from their datasets.

16:00-17:30 (East Foyer / Virtual)
**SynJax: Structured Probability Distributions for JAX**
*Miloš Stanojević and Laurent Sartran*
The development of deep learning software libraries enabled significant progress in the field by allowing users to focus on modeling, while letting the library to take care of the tedious and time-consuming task of optimizing execution for modern hardware accelerators. However, this has benefited only particular types of deep learning models, such as Transformers, whose primitives map easily to the vectorized computation. The models that explicitly account for structured objects, such as trees and segmentations, did not benefit equally because they require custom algorithms that are difficult to implement in a vectorized form. SynJax directly addresses this problem by providing an efficient vectorized implementation of inference algorithms for structured distributions covering alignment, tagging, segmentation, constituency trees and spanning trees. This is done by exploiting the connection between algorithms for automatic differentiation and probabilistic inference. With SynJax we can build large-scale differentiable models that explicitly model structure in the data. The code is available at https://github.com/google-deepmind/synjax

16:00-17:30 (East Foyer / Virtual)
**RESIN-EDITOR: A Schema-guided Hierarchical Event Graph Visualizer and Editor**
*Khanh Duy Nguyen, Zixuan Zhang, Reece Suchocki, Sha Li, Martha Palmer, Susan Windisch Brown, Jiawei Han and Heng Ji*
In this paper, we present RESIN-EDITOR, an interactive event graph visualizer and editor designed for analyzing complex events. Our RESIN-EDITOR system allows users to render and freely edit hierarchical event graphs extracted from multimedia and multi-document news clusters with guidance from human-curated event schemas. RESIN-EDITOR's unique features include hierarchical graph visualization, comprehensive source tracing, and interactive user editing, which significantly outperforms existing Information Extraction (IE) visualization tools in both IE result analysis and general model improvements. In our evaluation of RESIN-EDITOR, we demonstrate ways in which our tool is effective in understanding complex events and enhancing system performances. The source code, a video demonstration, and a live website for RESIN-EDITOR have been made publicly available.

## Poster session 3

16:00-17:30 (East Foyer / Virtual)

---

16:00-17:30 (East Foyer / Virtual)
**#1 AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages**
*Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino D. M. A. Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Ibrahim Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku and Stephen Arthur*
Africa is home to over 2,000 languages from over six language families and has the highest linguistic diversity among all continents. This includes 75 languages with at least one million speakers each. Yet, there is little NLP research conducted on African languages. Crucial in enabling such research is the availability of high-quality annotated datasets. In this paper, we introduce AfriSenti, a sentiment analysis benchmark that contains a total of >110,000 tweets in 14 African languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yoruba) from four language families. The tweets were annotated by native speakers and used in the AfriSenti-SemEval shared task (with over 200 participants, see website: https://afrisenti-semeval.github.io). We describe the data collection methodology, annotation process, and the challenges we dealt with when curating each dataset. We further report baseline experiments conducted on the AfriSenti datasets and discuss their usefulness.

16:00-17:30 (East Foyer / Virtual)
**#2 Towards Noise-Tolerant Speech-Referring Video Object Segmentation: Bridging Speech and Text**
*Xiang Li, Jinglu Wang, Xiaohao Xu, Muqiao Yang, Fan Yang, Yizhou Zhao, Rita Singh and Bhiksha Raj*
Linguistic communication is prevalent in Human-Computer Interaction (HCI). Speech (spoken language) serves as a convenient yet potentially ambiguous form due to noise and accents, exposing a gap compared to text. In this study, we investigate the prominent HCI task, Referring Video Object Segmentation (R-VOS), which aims to segment and track objects using linguistic references. While text input is well-investigated, speech input is under-explored. Our objective is to bridge the gap between speech and text, enabling the adaptation of existing text-input R-VOS models to accommodate noisy speech input effectively. Specifically, we propose a method to align the semantic spaces between speech and text by incorporating two key modules: 1) Noise-Aware Semantic Adjustment (NSA) for clear semantics extraction

from noisy speech; and 2) Semantic Jitter Suppression (SJS) enabling R-VOS models to tolerate noisy queries. Comprehensive experiments conducted on the challenging AVOS benchmarks reveal that our proposed method outperforms state-of-the-art approaches.

16:00-17:30 (East Foyer / Virtual)
**#3 Continually Improving Extractive QA via Human Feedback**
*Ge Gao, Hung-Ting Chen, Yoav Artzi and Eunsol Choi*
We study continually improving an extractive question answering (QA) system via human user feedback. We design and deploy an iterative approach, where information-seeking users ask questions, receive model-predicted answers, and provide feedback. We conduct experiments involving thousands of user interactions under diverse setups to broaden the understanding of learning from feedback over time. Our experiments show effective improvement from user feedback of extractive QA models over time across different data regimes, including significant potential for domain adaptation.

16:00-17:30 (East Foyer / Virtual)
**#4 Generative Spoken Language Model based on continuous word-sized audio tokens**
*Robin Jonathan Algayres, Yossi Adi, Tu Anh Nguyen, Jade Copet, Gabriel Synnaeve, Benoît Sagot and Emmanuel Dupoux*
In NLP, text language models based on words or subwords are known to outperform their character-based counterparts. Yet, in the speech community, the standard input of spoken LMs are 20ms or 40ms-long discrete units (shorter than a phoneme). Taking inspiration from word-based LM, we introduce a Generative Spoken Language Model (GSLM) based on word-size continuous-valued audio tokens that can generate diverse and expressive language output. This is obtained by replacing lookup table for lexical types with a Lexical Embedding function, the cross entropy loss by a contrastive loss, and multinomial sampling by k-NN sampling. The resulting model is the first generative language model based on word-size continuous tokens. Its performance is on par with discrete unit GSLMs regarding generation quality as measured by automatic metrics and subjective human judgements. Moreover, it is five times more memory efficient thanks to its large 200ms units. In addition, the embeddings before and after the Lexical Embedder are phonetically and semantically interpretable.

16:00-17:30 (East Foyer / Virtual)
**#5 SimCSE++: Improving Contrastive Learning for Sentence Embeddings from Two Perspectives**
*Jiahao Xu, Wei Shao, Lihui Chen and Lemao Liu*
This paper improves contrastive learning for sentence embeddings from two perspectives: handling dropout noise and addressing feature corruption. Specifically, for the first perspective, we identify that the dropout noise from negative pairs affects the model's performance. Therefore, we propose a simple yet effective method to deal with such type of noise. Secondly, we pinpoint the rank bottleneck of current solutions to feature corruption and propose a dimension-wise contrastive learning objective to address this issue. Both proposed methods are generic and can be applied to any contrastive learning based models for sentence embeddings. Experimental results on standard benchmarks demonstrate that combining both proposed methods leads to a gain of 1.8 points compared to the strong baseline SimCSE configured with BERT base. Furthermore, applying the proposed method to DiffCSE, another strong contrastive learning based baseline, results in a gain of 1.4 points.

16:00-17:30 (East Foyer / Virtual)
**#6 HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation**
*David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault and Marta R. Costa-jussà*
Hallucinations in machine translation are translations that contain information completely unrelated to the input. Omissions are translations that do not include some of the input information. While both cases tend to be catastrophic errors undermining user trust, annotated data with these types of pathologies is extremely scarce and is limited to a few high-resource languages. In this work, we release an annotated dataset for the hallucination and omission phenomena covering 18 translation directions with varying resource levels and scripts. Our annotation covers different levels of partial and full hallucinations as well as omissions both at the sentence and at the word level. Additionally, we revisit previous methods for hallucination and omission detection, show that conclusions made based on a single language pair largely do not hold for a large-scale evaluation, and establish new solid baselines.

16:00-17:30 (East Foyer / Virtual)
**#7 A Suite of Generative Tasks for Multi-Level Multimodal Webpage Understanding**
*Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni and Mandy Guo*
Webpages have been a rich, scalable resource for vision-language and language only tasks. Yet only pieces of webpages are kept in existing datasets: image-caption pairs, long text articles, or raw HTML, never all in one place. Webpage tasks have resultingly received little attention and structured image-text data left underused. To study multimodal webpage understanding, we introduce the Wikipedia Webpage suite (WikiWeb2M) containing 2M pages with all of the associated image, text, and structure data. We verify its utility on three generative tasks: page description generation, section summarization, and contextual image captioning. We design a novel attention mechanism Prefix Global, which selects the most relevant image and text content as global tokens to attend to the rest of the webpage for context. By using page structure to separate such tokens, it performs better than full attention with lower computational complexity. Extensive experiments show that the new data in WikiWeb2M improves task performance compared to prior work.

16:00-17:30 (East Foyer / Virtual)
**#8 Evaluating the Rationale Understanding of Critical Reasoning in Logical Reading Comprehension**
*Akira Kawabata and Saku Sugawara*
To precisely evaluate a language model's capability for logical reading comprehension, we present a dataset for testing the understanding of the rationale behind critical reasoning. For questions taken from an existing multiple-choice logical reading comprehension dataset, we crowdsource rationale texts that explain why we should select or eliminate answer options, resulting in 3,003 multiple-choice subquestions that are associated with 943 main questions. Experiments on our dataset show that recent large language models (e.g., InstructGPT) struggle to answer the subquestions even if they are able to answer the main questions correctly. We find that the models perform particularly poorly in answering subquestions written for the incorrect options of the main questions, implying that the models have a limited capability for explaining why incorrect alternatives should be eliminated. These results suggest that our dataset encourages further investigation into the critical reasoning ability of language models while focusing on the elimination process of relevant alternatives.

16:00-17:30 (East Foyer / Virtual)
**#9 STAIR: Learning Sparse Text and Image Representation in Grounded Tokens**
*Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander T Toshev, Yantao Zheng, Jonathon Shlens, Ruoming Pang and Yinfei Yang*
Image and text retrieval is one of the foundational tasks in the vision and language domain with multiple real-world applications. State-of-the-art contrastive approaches, e.g. CLIP, ALIGN, represent images and texts as dense embeddings and calculate the similarity in the dense embedding space as the matching score. On the other hand, sparse semantic features like bag-of-words models are more interpretable, but

believed to suffer from inferior accuracy than dense representations. In this work, we show that it is possible to build a sparse semantic representation that is as powerful as, or even better than, dense presentations. We extend the CLIP model and build a sparse text and image representation (STAIR), where the image and text are mapped to a sparse token space. Each token in the space is a (sub-)word in the vocabulary, which is not only interpretable but also easy to integrate with existing information retrieval systems. STAIR model significantly outperforms a CLIP model with +4.9% and +4.3% absolute Recall@1 improvement on COCO-5k text→image and image→text retrieval respectively. It also achieved better performance on both of ImageNet zero-shot and linear probing compared to CLIP.

16:00-17:30 (East Foyer / Virtual)
**#10 DPP-TTS: Diversifying prosodic features of speech via determinantal point processes**
*Seongho Joo, Hyukhun Koh and Kyomin Jung*
With the rapid advancement in deep generative models, recent neural Text-To-Speech(TTS) models have succeeded in synthesizing human-like speech. There have been some efforts to generate speech with various prosody beyond monotonous prosody patterns. However, previous works have several limitations. First, typical TTS models depend on the scaled sampling temperature for boosting the diversity of prosody. Speech samples generated at high sampling temperatures often lack perceptual prosodic diversity, which can adversely affect the naturalness of the speech. Second, the diversity among samples is neglected since the sampling procedure often focuses on a single speech sample rather than multiple ones. In this paper, we propose DPP-TTS: a text-to-speech model based on Determinantal Point Processes (DPPs) with a prosody diversifying module. Our TTS model is capable of generating speech samples that simultaneously consider perceptual diversity in each sample and among multiple samples. We demonstrate that DPP-TTS generates speech samples with more diversified prosody than baselines in the side-by-side comparison test considering the naturalness of speech at the same time.

16:00-17:30 (East Foyer / Virtual)
**#11 Improving Chinese Pop Song and Hokkien Gezi Opera Singing Voice Synthesis by Enhancing Local Modeling**
*Peng Bai, Yue Zhou, Meizhen Zheng, Wujin Sun and Xiaodong Shi*
Singing Voice Synthesis (SVS) strives to synthesize pleasing vocals based on music scores and lyrics. The current acoustic models based on Transformer usually process the entire sequence globally and use a simple L1 loss. However, this approach overlooks the significance of local modeling within the sequence and the local optimization of the hard-to-synthesize parts in the predicted mel-spectrogram. Consequently, the synthesized audio exhibits local incongruities (*e.g.*, local pronunciation jitter or local noise). To address this problem, we propose two methods to enhance local modeling in the acoustic model. First, we devise a nearest neighbor local attention, where each phoneme token focuses only on the adjacent phoneme tokens located before and after it. Second, we propose a phoneme-level local adaptive weights loss function that enables the model to focus more on the hard-to-synthesize parts of the mel-spectrogram. We have verified the universality of our methods on public Chinese pop song and Hokkien Gezi Opera datasets. Extensive experiments have demonstrated the effectiveness of our methods, resulting in significant improvements in both objective and subjective evaluations when compared to the strong baselines. Our code and demonstration samples are available at https://github.com/baipeng1/SVSELM.

16:00-17:30 (East Foyer / Virtual)
**#12 Multilingual Holistic Bias: Extending Descriptors and Patterns to Unveil Demographic Biases in Languages at Scale**
*Marta R. Costa-jussà, Pierre Andrews, Eric Michael Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Edward Licht and Carleigh Wood*
We introduce a multilingual extension of the HolisticBias dataset, the largest English template-based taxonomy of textual people references: Multilingual HolisticBias. This extension consists of 20,459 sentences in 50 languages distributed across 13 demographic axes. Source sentences are built from combinations of 118 demographic descriptors and three patterns, excluding nonsensical combinations. Multilingual translations include alternatives for gendered languages that cover gendered translations when there is ambiguity in English. Our dataset is intended to uncover demographic imbalances and be the tool to quantify mitigations towards them. Our initial findings show that translation quality for EN-to-XX translations is an average of almost 8 spBLEU better when evaluating with the masculine human reference compared to feminine. In the opposite direction, XX-to-EN, we compare the robustness of the model when the source input only differs in gender (masculine or feminine) and masculine translations are an average of almost 4 spBLEU better than feminine. When embedding sentences to a joint multilingual sentence representations space, we find that for most languages masculine translations are significantly closer to the English neutral sentences when embedded.

16:00-17:30 (East Foyer / Virtual)
**#13 NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports**
*Mael Jullien, Marco Valentino, Hannah Ruth Frost, Paul O'Regan, Dónal Landers and Andre Freitas*
How can we interpret and retrieve medical evidence to support clinical decisions? Clinical trial reports (CTR) amassed over the years contain indispensable information for the development of personalized medicine. However, it is practically infeasible to manually inspect over 400,000+ clinical trial reports in order to find the best evidence for experimental treatments. Natural Language Inference (NLI) offers a potential solution to this problem, by allowing the scalable computation of textual entailment. However, existing NLI models perform poorly on biomedical corpora, and previously published datasets fail to capture the full complexity of inference over CTRs. In this work, we present a novel resource to advance research on NLI for reasoning on CTRs. The resource includes two main tasks. Firstly, to determine the inference relation between a natural language statement, and a CTR. Secondly, to retrieve supporting facts to justify the predicted relation. We provide NLI4CT, a corpus of 2400 statements and CTRs, annotated for these tasks. Baselines on this corpus expose the limitations of existing NLI approaches, with 6 state-of-the-art NLI models achieving a maximum F1 score of 0.627. To the best of our knowledge, we are the first to design a task that covers the interpretation of full CTRs. To encourage further work on this challenging dataset, we make the corpus, competition leaderboard, and website, available on CodaLab, and code to replicate the baseline experiments on GitHub.

16:00-17:30 (East Foyer / Virtual)
**#14 Towards Unsupervised Recognition of Token-level Semantic Differences in Related Documents**
*Jannis Vamvas and Rico Sennrich*
Automatically highlighting words that cause semantic differences between two documents could be useful for a wide range of applications. We formulate recognizing semantic differences (RSD) as a token-level regression task and study three unsupervised approaches that rely on a masked language model. To assess the approaches, we begin with basic English sentences and gradually move to more complex, cross-lingual document pairs. Our results show that an approach based on word alignment and sentence-level contrastive learning has a robust correlation to gold labels. However, all unsupervised approaches still leave a large margin of improvement.

16:00-17:30 (East Foyer / Virtual)
**#15 IfQA: A Dataset for Open-domain Question Answering under Counterfactual Presuppositions**
*Wenhao Yu, Meng Jiang, Peter Clark and Ashish Sabharwal*
Although counterfactual reasoning is a fundamental aspect of intelligence, the lack of large-scale counterfactual open-domain question-answering (QA) benchmarks makes it difficult to evaluate and improve models on this ability. To address this void, we introduce the first such dataset, named IfQA, where each question is based on a counterfactual presupposition via an "if" clause. Such questions require models to

go beyond retrieving direct factual knowledge from the Web: they must identify the right information to retrieve and reason about an imagined situation that may even go against the facts built into their parameters. The IfQA dataset contains 3,800 questions that were annotated by crowdworkers on relevant Wikipedia passages. Empirical analysis reveals that the IfQA dataset is highly challenging for existing open-domain QA methods, including supervised retrieve-then-read pipeline methods (F1 score 44.5), as well as recent few-shot approaches such as chain-of-thought prompting with ChatGPT (F1 score 57.2). We hope the unique challenges posed by IfQA will push open-domain QA research on both retrieval and reasoning fronts, while also helping endow counterfactual reasoning abilities to today's language understanding models.

16:00-17:30 (East Foyer / Virtual)
**#16 Not all quantifiers are equal: Probing Transformer-based language models' understanding of generalised quantifiers**
*Tharindu Madusanka, Iqra Zahid, Hao Li, Ian Pratt-Hartmann and Riza Batista-Navarro*
How do different generalised quantifiers affect the behaviour of transformer-based language models (TLMs)? The recent popularity of TLMs and the central role generalised quantifiers have traditionally played in linguistics and logic bring this question into particular focus. The current research investigating this subject has not utilised a task defined purely in a logical sense, and thus, has not captured the underlying logical significance of generalised quantifiers. Consequently, they have not answered the aforementioned question faithfully or adequately. Therefore, we investigate how different generalised quantifiers affect TLMs by employing a textual entailment problem defined in a purely logical sense, namely, model-checking with natural language. Our approach permits the automatic construction of datasets with respect to which we can assess the ability of TLMs to learn the meanings of generalised quantifiers. Our investigation reveals that TLMs generally can comprehend the logical semantics of the most common generalised quantifiers, but that distinct quantifiers influence TLMs in varying ways.

16:00-17:30 (East Foyer / Virtual)
**#17 SCENE: Self-Labeled Counterfactuals for Extrapolating to Negative Examples**
*Deqing Fu, Ameya Godbole and Robin Jia*
Detecting negatives (such as non-entailment relationships, unanswerable questions, and false claims) is an important and challenging aspect of many natural language understanding tasks. Though manually collecting challenging negative examples can help models detect them, it is both costly and domain-specific. In this work, we propose Self-labeled Counterfactuals for Extrapolating to Negative Examples (SCENE), an automatic method for synthesizing training data that greatly improves models' ability to detect challenging negative examples. In contrast with standard data augmentation, which synthesizes new examples for existing labels, SCENE can synthesize negative examples zero-shot from only positive ones. Given a positive example, SCENE perturbs it with a mask infilling model, then determines whether the resulting example is negative based on a self-training heuristic. With access to only answerable training examples, SCENE can close 69.6% of the performance gap on SQuAD 2.0, a dataset where half of the evaluation examples are unanswerable, compared to a model trained on SQuAD 2.0. Our method also extends to boolean question answering and recognizing textual entailment, and improves generalization from SQuAD to ACE-whQA, an out-of-domain extractive QA benchmark.

16:00-17:30 (East Foyer / Virtual)
**#18 CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients**
*Jaehyung Seo, Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park and Heuiseok Lim*
Korean morphological variations present unique opportunities and challenges in natural language processing (NLP), necessitating an advanced understanding of morpheme-based sentence construction. The complexity of morphological variations allows for diverse sentence forms based on the syntactic-semantic integration of functional morphemes (i.e., affixes) to lexical morphemes (i.e., roots). With this in mind, we propose a method - CHEF, replicating the morphological transformations inherent in sentences based on lexical and functional morpheme combinations through generative data augmentation. CHEF operates using a morpheme blender and a label discriminator, thereby enhancing the diversity of Korean sentence forms by capturing the properties of agglutination while maintaining label consistency. We conduct experiments on Korean multiple classification datasets, improving model performance in full- and few-shot settings. Our proposed method boosts performance beyond the preceding data augmentation methods without incurring external data usage. We demonstrate that our approach achieves comparable results yielded by augmentation techniques that use large language models (LLMs).

16:00-17:30 (East Foyer / Virtual)
**#19 AD-NLP: A Benchmark for Anomaly Detection in Natural Language Processing**
*Matei Bejan, Andrei Manolache and Marius Popescu*
Deep learning models have reignited the interest in Anomaly Detection research in recent years. Methods for Anomaly Detection in text have shown strong empirical results on ad-hoc anomaly setups that are usually made by downsampling some classes of a labeled dataset. This can lead to reproducibility issues and models that are biased toward detecting particular anomalies while failing to recognize them in more sophisticated scenarios. In the present work, we provide a unified benchmark for detecting various types of anomalies, focusing on problems that can be naturally formulated as Anomaly Detection in text, ranging from syntax to stylistics. In this way, we are hoping to facilitate research in Text Anomaly Detection. We also evaluate and analyze two strong shallow baselines, as well as two of the current state-of-the-art neural approaches, providing insights into the knowledge the neural models are learning when performing the anomaly detection task. We provide the code for evaluation, downloading, and preprocessing the dataset at https://github.com/mateibejan1/ad-nlp/.

16:00-17:30 (East Foyer / Virtual)
**#20 ORCHID: A Chinese Debate Corpus for Target-Independent Stance Detection and Argumentative Dialogue Summarization**
*Xiutian Zhao, Ke Wang and Wei Peng*
Dialogue agents have been receiving increasing attention for years, and this trend has been further boosted by the recent progress of large language models (LLMs). Stance detection and dialogue summarization are two core tasks of dialogue agents in application scenarios that involve argumentative dialogues. However, research on these tasks is limited by the insufficiency of public datasets, especially for non-English languages. To address this language resource gap in Chinese, we present ORCHID (Oral Chinese Debate), the first Chinese dataset for benchmarking target-independent stance detection and debate summarization. Our dataset consists of 1,218 real-world debates that were conducted in Chinese on 476 unique topics, containing 2,436 stance-specific summaries and 14,133 fully annotated utterances. Besides providing a versatile testbed for future research, we also conduct an empirical study on the dataset and propose an integrated task. The results show the challenging nature of the dataset and suggest a potential of incorporating stance detection in summarization for argumentative dialogue.

16:00-17:30 (East Foyer / Virtual)
**#21 DetGPT: Detect What You Need via Reasoning**
*Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong and Tong Zhang*
In recent years, the field of computer vision has seen significant advancements thanks to the development of large language models (LLMs). These models have enabled more effective and sophisticated interactions between humans and machines, paving the way for novel techniques that blur the lines between human and machine intelligence. In this paper, we introduce a new paradigm for object detection that we call reasoning-based object detection. Unlike conventional object detection methods that rely on specific object names, our approach enables users to interact with the system using natural language instructions, allowing for a higher level of interactivity. Our proposed method, called

DetGPT, leverages state-of-the-art multi-modal models and open-vocabulary object detectors to perform reasoning within the context of the user's instructions and the visual scene. This enables DetGPT to automatically locate the object of interest based on the user's expressed desires, even if the object is not explicitly mentioned. For instance, if a user expresses a desire for a cold beverage, DetGPT can analyze the image, identify a fridge, and use its knowledge of typical fridge contents to locate the beverage. This flexibility makes our system applicable across a wide range of fields, from robotics and automation to autonomous driving. Overall, our proposed paradigm and DetGPT demonstrate the potential for more sophisticated and intuitive interactions between humans and machines. We hope that our proposed paradigm and approach will provide inspiration to the community and open the door to more interactive and versatile object detection systems.

16:00-17:30 (East Foyer / Virtual)
**#22 Are Embedded Potatoes Still Vegetables? On the Limitations of WordNet Embeddings for Lexical Semantics**
*Xuyou Cheng, Michael Sejr Schlichtkrull and Guy Emerson*
Knowledge Base Embedding (KBE) models have been widely used to encode structured information from knowledge bases, including WordNet. However, the existing literature has predominantly focused on link prediction as the evaluation task, often neglecting exploration of the models' semantic capabilities. In this paper, we investigate the potential disconnect between the performance of KBE models of WordNet on link prediction and their ability to encode semantic information, highlighting the limitations of current evaluation protocols. Our findings reveal that some top-performing KBE models on the WN18RR benchmark exhibit subpar results on two semantic tasks and two downstream tasks. These results demonstrate the inadequacy of link prediction benchmarks for evaluating the semantic capabilities of KBE models, suggesting the need for a more targeted assessment approach.

16:00-17:30 (East Foyer / Virtual)
**#23 SLOG: A Structural Generalization Benchmark for Semantic Parsing**
*Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao and Najoung Kim*
The goal of compositional generalization benchmarks is to evaluate how well models generalize to new complex linguistic expressions. Existing benchmarks often focus on lexical generalization, the interpretation of novel lexical items in syntactic structures familiar from training; structural generalization tasks, where a model needs to interpret syntactic structures that are themselves unfamiliar from training, are often underrepresented, resulting in overly optimistic perceptions of how well models can generalize. We introduce SLOG, a semantic parsing dataset that extends COGS (Kim and Linzen, 2020) with 17 structural generalization cases. In our experiments, the generalization accuracy of Transformer models, including pretrained ones, only reaches 40.6%, while a structure-aware parser only achieves 70.8%. These results are far from the near-perfect accuracy existing models achieve on COGS, demonstrating the role of SLOG in foregrounding the large discrepancy between models' lexical and structural generalization capacities.

16:00-17:30 (East Foyer / Virtual)
**#24 ZEROTOP: Zero-Shot Task-Oriented Semantic Parsing using Large Language Models**
*Dheeraj Mekala, Jason Andrew Wolfe and Subhro Roy*
We explore the use of large language models (LLMs) for zero-shot semantic parsing. Semantic parsing involves mapping natural language utterances to task-specific meaning representations. LLMs are generally trained on publicly available text and code and cannot be expected to directly generalize to domain-specific parsing tasks in a zero-shot setting. In this work, we propose ZEROTOP, a zero-shot task-oriented parsing method that decomposes semantic parsing problem into a set of abstractive and extractive question-answering (QA) problems. For each utterance, we prompt the LLM with questions corresponding to its top-level intent and a set of slots and use the LLM generations to construct the target meaning representation. We observe that current LLMs fail to detect unanswerable questions; and as a result, cannot handle questions corresponding to missing slots. We address this by fine-tuning a language model on public QA datasets using synthetic negative samples. Experimental results show that our QA-based decomposition paired with the fine-tuned LLM can zero-shot parse $\approx$ 16% of utterances in the MTOP dataset.

16:00-17:30 (East Foyer / Virtual)
**#25 Faithful Model Evaluation for Model-Based Metrics**
*Qian Hu, Palash Goyal and Rahul Gupta*
Statistical significance testing is used in natural language processing (NLP) to determine whether the results of a study or experiment are likely to be due to chance or if they reflect a genuine relationship. A key step in significance testing is the estimation of confidence interval which is a function of sample variance. Sample variance calculation is straightforward when evaluating against ground truth. However, in many cases, a metric model is often used for evaluation. For example, to compare toxicity of two large language models, a toxicity classifier is used for evaluation. Existing works usually do not consider the variance change due to metric model errors, which can lead to wrong conclusions. In this work, we establish the mathematical foundation of significance testing for model-based metrics. With experiments on public benchmark datasets and a production system, we show that considering metric model errors to calculate sample variances for model-based metrics changes the conclusions in certain experiments.

16:00-17:30 (East Foyer / Virtual)
**#26 KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing**
*Seonmin Koo, Chanjun Park, Jinsung Kim, Jaehyung Seo, Sugyeong Eo, Hyeonseok Moon and Heuiseok Lim*
Automatic Speech Recognition (ASR) systems are instrumental across various applications, with their performance being critically tied to user satisfaction. Conventional evaluation metrics for ASR systems produce a singular aggregate score, which is insufficient for understanding specific system vulnerabilities. Therefore, we aim to address the limitations of the previous ASR evaluation methods by introducing the Korean Error Explainable Benchmark Dataset for ASR and Post-processing (KEBAP). KEBAP enables comprehensive analysis of ASR systems at both speech- and text levels, thereby facilitating a more balanced assessment encompassing speech recognition accuracy and user readability. KEBAP provides 37 newly defined speech-level resources incorporating diverse noise environments and speaker characteristics categories, also presenting 13 distinct text-level error types. This paper demonstrates detailed statistical analyses of colloquial noise categories and textual error types. Furthermore, we conduct extensive validation and analysis on commercially deployed ASR systems, providing valuable insights into their performance. As a more fine-grained and real-world-centric evaluation method, KEBAP contributes to identifying and mitigating potential weaknesses in ASR systems.

16:00-17:30 (East Foyer / Virtual)
**#27 Enhancing Chat Language Models by Scaling High-quality Instructional Conversations**
*Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun and Bowen Zhou*
Fine-tuning on instruction data has been widely validated as an effective practice for implementing chat language models like ChatGPT. Scaling the diversity and quality of such data, although straightforward, stands a great chance of leading to improved performance. This paper aims to push the upper bound of open-source models further. We first provide a systematically designed, diverse, informative, large-scale dataset of instructional conversations, UltraChat, which does not involve human queries. Our objective is to capture the breadth of interactions between a human user and an AI assistant and employs a comprehensive framework to generate multi-turn conversation iteratively. UltraChat contains 1.5 million high-quality multi-turn dialogues and covers a wide range of topics and instructions. Our statistical analysis of UltraChat

reveals its superiority in various key metrics, including scale, average length, diversity, coherence, etc., solidifying its position as a leading open-source dataset. Building upon UltraChat, we fine-tune a LLaMA model to create a powerful conversational model, UltraLM. Our evaluations indicate that UltraLM consistently outperforms other open-source models, including WizardLM and Vicuna, the previously recognized state-of-the-art open-source models.

16:00-17:30 (East Foyer / Virtual)
**#28 ReasoningLM: Enabling Structural Subgraph Reasoning in Pre-trained Language Models for Question Answering over Knowledge Graph**
*Jinhao Jiang, Kun Zhou, Xin Zhao, Yaliang Li and Ji-Rong Wen*
Question Answering over Knowledge Graph (KGQA) aims to seek answer entities for the natural language question from a large-scale Knowledge Graph (KG). To better perform reasoning on KG, recent work typically adopts a pre-trained language model (PLM) to model the question, and a graph neural network (GNN) based module to perform multi-hop reasoning on the KG. Despite the effectiveness, due to the divergence in model architecture, the PLM and GNN are not closely integrated, limiting the knowledge sharing and fine-grained feature interactions. To solve it, we aim to simplify the above two-module approach, and develop a more capable PLM that can directly support subgraph reasoning for KGQA, namely ReasoningLM. In our approach, we propose a subgraph-aware self-attention mechanism to imitate the GNN for performing structured reasoning, and also adopt an adaptation tuning strategy to adapt the model parameters with 20,000 subgraphs with synthesized questions. After adaptation, the PLM can be parameter-efficient fine-tuned on downstream tasks. Experiments show that ReasoningLM surpasses state-of-the-art models by a large margin, even with fewer updated parameters and less training data. Our codes and data are publicly available at https://github.com/RUCAIBox/ReasoningLM.

16:00-17:30 (East Foyer / Virtual)
**#29 CiteBench: A Benchmark for Scientific Citation Text Generation**
*Martin Funkquist, Ilia Kuznetsov, Yufang Hou and Iryna Gurevych*
Science progresses by building upon the prior body of knowledge documented in scientific publications. The acceleration of research makes it hard to stay up-to-date with the recent developments and to summarize the ever-growing body of prior work. To address this, the task of citation text generation aims to produce accurate textual summaries given a set of papers-to-cite and the citing paper context. Due to otherwise rare explicit anchoring of cited documents in the citing paper, citation text generation provides an excellent opportunity to study how humans aggregate and synthesize textual knowledge from sources. Yet, existing studies are based upon widely diverging task definitions, which makes it hard to study this task systematically. To address this challenge, we propose CiteBench: a benchmark for citation text generation that unifies multiple diverse datasets and enables standardized evaluation of citation text generation models across task designs and domains. Using the new benchmark, we investigate the performance of multiple strong baselines, test their transferability between the datasets, and deliver new insights into the task definition and evaluation to guide future research in citation text generation. We make the code for CiteBench publicly available at https://github.com/UKPLab/citebench.

16:00-17:30 (East Foyer / Virtual)
**#30 Cabbage Sweeter than Cake? Analysing the Potential of Large Language Models for Learning Conceptual Spaces**
*Usashi Chatterjee, Amit Gajbhiye and Steven Schockaert*
The theory of Conceptual Spaces is an influential cognitive-linguistic framework for representing the meaning of concepts. Conceptual spaces are constructed from a set of quality dimensions, which essentially correspond to primitive perceptual features (e.g. hue or size). These quality dimensions are usually learned from human judgements, which means that applications of conceptual spaces tend to be limited to narrow domains (e.g. modelling colour or taste). Encouraged by recent findings about the ability of Large Language Models (LLMs) to learn perceptually grounded representations, we explore the potential of such models for learning conceptual spaces. Our experiments show that LLMs can indeed be used for learning meaningful representations to some extent. However, we also find that fine-tuned models of the BERT family are able to match or even outperform the largest GPT-3 model, despite being 2 to 3 orders of magnitude smaller.

16:00-17:30 (East Foyer / Virtual)
**#31 Solving Hard Analogy Questions with Relation Embedding Chains**
*Nitesh Kumar and Steven Schockaert*
Modelling how concepts are related is a central topic in Lexical Semantics. A common strategy is to rely on knowledge graphs (KGs) such as ConceptNet, and to model the relation between two concepts as a set of paths. However, KGs are limited to a fixed set of relation types, and they are incomplete and often noisy. Another strategy is to distill relation embeddings from a fine-tuned language model. However, this is less suitable for words that are only indirectly related and it does not readily allow us to incorporate structured domain knowledge. In this paper, we aim to combine the best of both worlds. We model relations as paths but associate their edges with relation embeddings. The paths are obtained by first identifying suitable intermediate words and then selecting those words for which informative relation embeddings can be obtained. We empirically show that our proposed representations are useful for solving hard analogy questions.

16:00-17:30 (East Foyer / Virtual)
**#32 Superlim: A Swedish Language Understanding Evaluation Benchmark**
*Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen and Nina Tahmasebi*
We present Superlim, a multi-task NLP benchmark and analysis platform for evaluating Swedish language models, a counterpart to the English-language (Super)GLUE suite. We describe the dataset, the tasks, the leaderboard and report the baseline results yielded by a reference implementation. The tested models do not approach ceiling performance on any of the tasks, which suggests that Superlim is truly difficult, a desirable quality for a benchmark. We address methodological challenges, such as mitigating the Anglocentric bias when creating datasets for a less-resourced language; choosing the most appropriate measures; documenting the datasets and making the leaderboard convenient and transparent. We also highlight other potential usages of the dataset, such as, for instance, the evaluation of cross-lingual transfer learning.

16:00-17:30 (East Foyer / Virtual)
**#33 Building Persona Consistent Dialogue Agents with Offline Reinforcement Learning**
*Ryan Shea and Zhou Yu*
Maintaining a consistent persona is a key quality for any open domain dialogue system. Current state-of-the-art systems do this by training agents with supervised learning or online reinforcement learning (RL). However, systems trained with supervised learning often lack consistency as they are never punished for uttering contradictions. Additional training with RL can alleviate some of these issues, however the training process is expensive. Instead, we propose an offline RL framework to improve the persona consistency of dialogue systems. Our framework allows us to combine the advantages of previous methods as we can inexpensively train our model on existing data as in supervised learning, while punishing and rewarding specific utterances as in RL. We also introduce a simple importance sampling method to reduce the variance of importance weights in offline RL training which we call Variance-Reducing MLE-Initialized (VaRMI) importance sampling. Our

automatic and human evaluations show that our framework improves both the persona consistency and dialogue quality of a state-of-the-art social chatbot.

16:00-17:30 (East Foyer / Virtual)
**#34 Can We Edit Multimodal Large Language Models?**
*Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen and Ningyu Zhang*
In this paper, we focus on editing multimodal Large Language Models (LLMs). Compared to editing single-modal LLMs, multimodal model editing is more challenging, which demands a higher level of scrutiny and careful consideration in the editing process. To facilitate research in this area, we construct a new benchmark, dubbed MMEdit, for editing multimodal LLMs and establishing a suite of innovative metrics for evaluation. We conduct comprehensive experiments involving various model editing baselines and analyze the impact of editing different components for multimodal LLMs. Empirically, we notice that previous baselines can implement editing multimodal LLMs to some extent, but the effect is still barely satisfactory, indicating the potential difficulty of this task. We hope that our work can provide the NLP community with insights.

16:00-17:30 (East Foyer / Virtual)
**#35 Knowledge-Augmented Language Model Verification**
*Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park and Sung Ju Hwang*
Recent Language Models (LMs) have shown impressive capabilities in generating texts with the knowledge internalized in parameters. Yet, LMs often generate the factually incorrect responses to the given queries, since their knowledge may be inaccurate, incomplete, and outdated. To address this problem, previous works propose to augment LMs with the knowledge retrieved from an external knowledge source. However, such approaches often show suboptimal text generation performance due to two reasons: 1) the model may fail to retrieve the knowledge relevant to the given query, or 2) the model may not faithfully reflect the retrieved knowledge in the generated text. To overcome these, we propose to verify the output and the knowledge of the knowledge-augmented LMs with a separate verifier, which is a small LM that is trained to detect those two types of errors through instruction-finetuning. Then, when the verifier recognizes an error, we can rectify it by either retrieving new knowledge or generating new text. Further, we use an ensemble of the outputs from different instructions with a single verifier to enhance the reliability of the verification processes. We validate the effectiveness of the proposed verification steps on multiple question answering benchmarks, whose results show that the proposed verifier effectively identifies retrieval and generation errors, allowing LMs to provide more factually correct outputs. Our code is available at https://github.com/JinheonBaek/KALMV.

16:00-17:30 (East Foyer / Virtual)
**#36 Retrieval-Generation Alignment for End-to-End Task-Oriented Dialogue System**
*Weizhou Shen, Yingqi Gao, Canbin Huang, Fanqi Wan, Xiaojun Quan and Wei Bi*
Developing an efficient retriever to retrieve knowledge from a large-scale knowledge base (KB) is critical for task-oriented dialogue systems to effectively handle localized and specialized tasks. However, widely used generative models such as T5 and ChatGPT often struggle to differentiate subtle differences among the retrieved KB records when generating responses, resulting in suboptimal quality of generated responses. In this paper, we propose the application of maximal marginal likelihood to train a perceptive retriever by utilizing signals from response generation for supervision. In addition, our approach goes beyond considering solely retrieved entities and incorporates various meta knowledge to guide the generator, thus improving the utilization of knowledge. We evaluate our approach on three task-oriented dialogue datasets using T5 and ChatGPT as the backbone models. The results demonstrate that when combined with meta knowledge, the response generator can effectively leverage high-quality knowledge records from the retriever and enhance the quality of generated responses. The code of this work is available at https://github.com/shenwzh3/MK-TOD.

16:00-17:30 (East Foyer / Virtual)
**#37 Answering Questions by Meta-Reasoning over Multiple Chains of Thought**
*Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch and Jonathan Berant*
Modern systems for multi-hop question answering (QA) typically break questions into a sequence of reasoning steps, termed chain-of-thought (CoT), before arriving at a final answer. Often, multiple chains are sampled and aggregated through a voting mechanism over the final answers, but the intermediate steps themselves are discarded. While such approaches improve performance, they do not consider the relations between intermediate steps across chains and do not provide a unified explanation for the predicted answer. We introduce Multi-Chain Reasoning (MCR), an approach which prompts large language models to meta-reason over multiple chains of thought, rather than aggregate their answers. MCR examines different reasoning chains, mixes information between them and selects the most relevant facts in generating an explanation and predicting the answer. MCR outperforms strong baselines on 7 multi-hop QA datasets. Moreover, our analysis reveals that MCR explanations exhibit high quality, enabling humans to verify its answers.

16:00-17:30 (East Foyer / Virtual)
**#38 FinEntity: Entity-level Sentiment Classification for Financial Texts**
*Yixuan Tang, Yi Yang, Allen H Huang, Andy Tam and Justin Z. Tang*
In the financial domain, conducting entity-level sentiment analysis is crucial for accurately assessing the sentiment directed toward a specific financial entity. To our knowledge, no publicly available dataset currently exists for this purpose. In this work, we introduce an entity-level sentiment classification dataset, called FinEntity, that annotates financial entity spans and their sentiment (positive, neutral, and negative) in financial news. We document the dataset construction process in the paper. Additionally, we benchmark several pre-trained models (BERT, FinBERT, etc.) and ChatGPT on entity-level sentiment classification. In a case study, we demonstrate the practical utility of using FinEntity in monitoring cryptocurrency markets. The data and code of FinEntity is available at https://github.com/yixuantt/FinEntity.

16:00-17:30 (East Foyer / Virtual)
**#39 CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset**
*Susanna Rücker and Alan Akbik*
The CoNLL-03 corpus is arguably the most well-known and utilized benchmark dataset for named entity recognition (NER). However, prior works found significant numbers of annotation errors, incompleteness, and inconsistencies in the data. This poses challenges to objectively comparing NER approaches and analyzing their errors, as current state-of-the-art models achieve F1-scores that are comparable to or even exceed the estimated noise level in CoNLL-03. To address this issue, we present a comprehensive relabeling effort assisted by automatic consistency checking that corrects 7.0% of all labels in the English CoNLL-03. Our effort adds a layer of entity linking annotation both for better explainability of NER labels and as additional safeguard of annotation quality. Our experimental evaluation finds not only that state-of-the-art approaches reach significantly higher F1-scores (97.1%) on our data, but crucially that the share of correct predictions falsely counted as errors due to annotation noise drops from 47% to 6%. This indicates that our resource is well suited to analyze the remaining errors made by state-of-the-art models, and that the theoretical upper bound even on high resource, coarse-grained NER is not yet reached. To facilitate such analysis, we make CleanCoNLL publicly available to the research community.

16:00-17:30 (East Foyer / Virtual)

**#40 Question Answering as Programming for Solving Time-Sensitive Questions**
*Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou and Yujiu Yang*
Question answering plays a pivotal role in human daily life because it involves our acquisition of knowledge about the world. However, due to the dynamic and ever-changing nature of real-world facts, the answer can be completely different when the time constraint in the question changes. Recently, Large Language Models (LLMs) have shown remarkable intelligence in question answering, while our experiments reveal that the aforementioned problems still pose a significant challenge to existing LLMs. This can be attributed to the LLMs' inability to perform rigorous reasoning based on surface-level text semantics. To overcome this limitation, rather than requiring LLMs to directly answer the question, we propose a novel approach where we reframe the **Q**uestion **A**nswering task **a**s **P**rogramming (**QAaP**). Concretely, by leveraging modern LLMs' superior capability in understanding both natural language and programming language, we endeavor to harness LLMs to represent diversely expressed text as well-structured code and select the best matching answer from multiple candidates through programming. We evaluate our QAaP framework on several time-sensitive question answering datasets and achieve decent improvement, up to $14.5\%$ over strong baselines.

16:00-17:30 (East Foyer / Virtual)
**#41 EDeR: Towards Understanding Dependency Relations Between Events**
*Ruiqi Li, Patrik Haslum and Leyang Cui*
Relation extraction is a crucial task in natural language processing (NLP) and information retrieval (IR). Previous work on event relation extraction mainly focuses on hierarchical, temporal and causal relations. Such relationships consider two events to be independent in terms of syntax and semantics, but they fail to recognize the interdependence between events. To bridge this gap, we introduce a human-annotated Event Dependency Relation dataset (EDeR). The annotation is done on a sample of documents from the OntoNotes dataset, which has the additional benefit that it integrates with existing, orthogonal, annotations of this dataset. We investigate baseline approaches for EDeR's event dependency relation prediction. We show that recognizing such event dependency relations can further benefit critical NLP tasks, including semantic role labelling and co-reference resolution.

16:00-17:30 (East Foyer / Virtual)
**#42 Optimized Tokenization for Transcribed Error Correction**
*Tomer Wullach and Shlomo Chazan*
The challenges facing speech recognition systems, such as variations in pronunciations, adverse audio conditions, and the scarcity of labeled data, emphasize the necessity for a post-processing step that corrects recurring errors. Previous research has shown the advantages of employing dedicated error correction models, yet training such models requires large amounts of labeled data which is not easily obtained. To overcome this limitation, synthetic transcribed-like data is often utilized, however, bridging the distribution gap between transcribed errors and synthetic noise is not trivial. In this paper, we demonstrate that the performance of correction models can be significantly increased by training solely using synthetic data. Specifically, we empirically show that: (1) synthetic data generated using the error distribution derived from a set of transcribed data outperforms the common approach of applying random perturbations; (2) applying language-specific adjustments to the vocabulary of a BPE tokenizer strike a balance between adapting to unseen distributions and retaining knowledge of transcribed errors. We showcase the benefits of these key observations, and evaluate our approach using multiple languages, speech recognition systems and prominent speech recognition datasets.

16:00-17:30 (East Foyer / Virtual)
**#43 Interview Evaluation: A Novel Approach for Automatic Evaluation of Conversational Question Answering Models**
*Xibo Li, Bowei Zou, Yifan Fan, Yanling Li, Ai Ti Aw and Yu Hong*
Conversational Question Answering (CQA) aims to provide natural language answers to users in information-seeking dialogues. Existing CQA benchmarks often evaluate models using pre-collected human-human conversations. However, replacing the model-predicted dialogue history with ground truth compromises the naturalness and sustainability of CQA evaluation. While previous studies proposed using predicted history and rewriting techniques to address unresolved coreferences and incoherencies, this approach renders the question self-contained from the conversation. In this paper, we propose a novel automatic evaluation approach, interview evaluation. Specifically, ChatGPT acts as the interviewer (Q agent) with a set of carefully designed prompts, and the CQA model under test serves as the interviewee (A agent). During the interview evaluation, questions are dynamically generated by the Q agent to guide the A agent in predicting the correct answer through an interactive process. We evaluated four different models on QuAC and two models on CoQA in our experiments. The experiment results demonstrate that our interview evaluation has advantages over previous CQA evaluation approaches, particularly in terms of naturalness and coherence. The source code is made publicly available.

16:00-17:30 (East Foyer / Virtual)
**#44 Exploring the Boundaries of GPT-4 in Radiology**
*Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Tajdin Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya V. Nori, Matthew P. Lungren, Ozan Oktay and Javier Alvarez-Valle*
The recent success of general-domain large language models (LLMs) has significantly changed the natural language processing paradigm towards a unified foundation model across domains and applications. In this paper, we focus on assessing the performance of GPT-4, the most capable LLM so far, on the text-based applications for radiology reports, comparing against state-of-the-art (SOTA) radiology-specific models. Exploring various prompting strategies, we evaluated GPT-4 on a diverse range of common radiology tasks and we found GPT-4 either outperforms or is on par with current SOTA radiology models. With zero-shot prompting, GPT-4 already obtains substantial gains ($\approx$ 10% absolute improvement) over radiology models in temporal sentence similarity classification (accuracy) and natural language inference ($F_1$). For tasks that require learning dataset-specific style or schema (e.g. findings summarisation), GPT-4 improves with example-based prompting and matches supervised SOTA. Our extensive error analysis with a board-certified radiologist shows GPT-4 has a sufficient level of radiology knowledge with only occasional errors in complex context that require nuanced domain knowledge. For findings summarisation, GPT-4 outputs are found to be overall comparable with existing manually-written impressions.

16:00-17:30 (East Foyer / Virtual)
**#45 Retrofitting Light-weight Language Models for Emotions using Supervised Contrastive Learning**
*Sapan Shah, Sreedhar Reddy and Pushpak Bhattacharyya*
We present a novel retrofitting method to induce emotion aspects into pre-trained language models (PLMs) such as BERT and RoBERTa. Our method updates pre-trained network weights using contrastive learning so that the text fragments exhibiting similar emotions are encoded nearby in the representation space, and the fragments with different emotion content are pushed apart. While doing so, it also ensures that the linguistic knowledge already present in PLMs is not inadvertently perturbed. The language models retrofitted by our method, i.e., BERTEmo and RoBERTaEmo, produce emotion-aware text representations, as evaluated through different clustering and retrieval metrics. For the downstream tasks on sentiment analysis and sarcasm detection, they perform better than their pre-trained counterparts (about 1% improvement in F1-score) and other existing approaches. Additionally, a more significant boost in performance is observed for the retrofitted models over pre-trained ones in few-shot learning setting.

16:00-17:30 (East Foyer / Virtual)
**#46 Speech-enriched Memory for Inference-time Adaptation of ASR Models to Word Dictionaries**
*Ashish Mittal, Sunita Sarawagi, Preethi Jyothi, George Saon and Gakuto Kurata*
Despite the impressive performance of ASR models on mainstream benchmarks, their performance on rare words is unsatisfactory. In enterprise settings, often a focused list of entities (such as locations, names, etc) are available which can be used to adapt the model to the terminology of specific domains. In this paper, we present a novel inference algorithm that improves the prediction of state-of-the-art ASR models using nearest-neighbor-based matching on an inference-time word list. We consider both the Transducer architecture that is useful in the streaming setting, and state-of-the-art encoder-decoder models such as Whisper. In our approach, a list of rare entities is indexed in a memory by synthesizing speech for each entry, and then storing the internal acoustic and language model states obtained from the best possible alignment on the ASR model. The memory is organized as a trie which we harness to perform a stateful lookup during inference. A key property of our extension is that we prevent spurious matches by restricting to only word-level matches. In our experiments on publicly available datasets and private benchmarks, we show that our method is effective in significantly improving rare word recognition.

16:00-17:30 (East Foyer / Virtual)
**#47 DUMB: A Dutch Model Benchmark**
*Wietse de Vries, Martijn Wieling and Malvina Nissim*
We introduce the Dutch Model Benchmark: DUMB. The benchmark includes a diverse set of datasets for low-, medium- and high-resource tasks. The total set of nine tasks includes four tasks that were previously not available in Dutch. Instead of relying on a mean score across tasks, we propose Relative Error Reduction (RER), which compares the DUMB performance of language models to a strong baseline which can be referred to in the future even when assessing different sets of language models. Through a comparison of 14 pre-trained language models (mono- and multi-lingual, of varying sizes), we assess the internal consistency of the benchmark tasks, as well as the factors that likely enable high performance. Our results indicate that current Dutch monolingual models under-perform and suggest training larger Dutch models with other architectures and pre-training objectives. At present, the highest performance is achieved by DeBERTaV3 (large), XLM-R (large) and mDeBERTaV3 (base). In addition to highlighting best strategies for training larger Dutch models, DUMB will foster further research on Dutch. A public leaderboard is available at https://dumbench.nl.

16:00-17:30 (East Foyer / Virtual)
**#48 MILDSum: A Novel Benchmark Dataset for Multilingual Summarization of Indian Legal Case Judgments**
*Debtanu Datta, Shubham Soni, Rajdeep Mukherjee and Saptarshi Ghosh*
Automatic summarization of legal case judgments is a practically important problem that has attracted substantial research efforts in many countries. In the context of the Indian judiciary, there is an additional complexity – Indian legal case judgments are mostly written in complex English, but a significant portion of India's population lacks command of the English language. Hence, it is crucial to summarize the legal documents in Indian languages to ensure equitable access to justice. While prior research primarily focuses on summarizing legal case judgments in their source languages, this study presents a pioneering effort toward cross-lingual summarization of English legal documents into Hindi, the most frequently spoken Indian language. We construct the first high-quality legal corpus comprising of 3,122 case judgments from prominent Indian courts in English, along with their summaries in both English and Hindi, drafted by legal practitioners. We benchmark the performance of several diverse summarization approaches on our corpus and demonstrate the need for further research in cross-lingual summarization in the legal domain.

16:00-17:30 (East Foyer / Virtual)
**#49 When Do Decompositions Help for Machine Reading?**
*Kangda Wei, Dawn Lawrie, Benjamin Van Durme, Yunmo Chen and Orion Weller*
Answering complex questions often requires multi-step reasoning in order to obtain the final answer. Most research into decompositions of complex questions involves open-domain systems, which have shown success in using these decompositions for improved retrieval. In the machine reading setting, however, work to understand when decompositions are helpful is understudied. We conduct experiments on decompositions in machine reading to unify recent work in this space, using a range of models and datasets. We find that decompositions can be helpful in zero or limited-data settings, giving several points of improvement in exact match. However, we also show that when models are given access to around a few hundred or more examples, decompositions are not helpful (and can actually be detrimental). Thus, our analysis implies that models can learn decompositions implicitly even with limited data.

16:00-17:30 (East Foyer / Virtual)
**#50 Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation**
*Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang and Ashwin Kalyan*
In this paper, we present a novel approach for distilling math word problem solving capabilities from large language models (LLMs) into smaller, more efficient student models. Our approach is designed to consider the student model's weaknesses and foster a tailored learning experience by generating targeted exercises aligned with educational science principles, such as knowledge tracing and personalized learning. Concretely, we let GPT-3 be a math tutor and run two steps iteratively: 1) assessing the student model's current learning status on a GPT-generated exercise book, and 2) improving the student model by training it with tailored exercise samples generated by GPT-3. Experimental results reveal that our approach outperforms LLMs (e.g., GPT-3 and PaLM) in accuracy across three distinct benchmarks while employing significantly fewer parameters. Furthermore, we provide a comprehensive analysis of the various components within our methodology to substantiate their efficacy.

16:00-17:30 (East Foyer / Virtual)
**#51 Counter Turing Test (CT2): AI-Generated Text Detection is Not as Easy as You May Think - Introducing AI Detectability Index (ADI)**
*Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar R Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit P. Sheth and Amitava Das*
With the rise of prolific ChatGPT, the risk and consequences of AI-generated text has increased alarmingly. This triggered a series of events, including an open letter, signed by thousands of researchers and tech leaders in March 2023, demanding a six-month moratorium on the training of AI systems more sophisticated than GPT-4. To address the inevitable question of ownership attribution for AI-generated artifacts, the US Copyright Office released a statement stating that "if the content is traditional elements of authorship produced by a machine, the work lacks human authorship and the office will not register it for copyright". Furthermore, both the US and the EU governments have recently drafted their initial proposals regarding the regulatory framework for AI. Given this cynosural spotlight on generative AI, 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. This paper introduces the Counter Turing Test (CT2), a benchmark consisting of techniques aiming to offer a comprehensive evaluation of the robustness of existing AGTD techniques. Our empirical findings unequivocally highlight the fragility of the proposed AGTD methods under scrutiny. Amidst the extensive deliberations on policy-making for regulating AI development, it is of utmost importance to assess the detectability of content generated by LLMs. Thus, to

establish a quantifiable spectrum facilitating the evaluation and ranking of LLMs according to their detectability levels, we propose the AI Detectability Index (ADI). We conduct a thorough examination of 15 contemporary LLMs, empirically demonstrating that larger LLMs tend to have a lower ADI, indicating they are less detectable compared to smaller LLMs. We firmly believe that ADI holds significant value as a tool for the wider NLP community, with the potential to serve as a rubric in AI-related policy-making.

16:00-17:30 (East Foyer / Virtual)
**#52 FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W Question-Answering**
*Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Ashok Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit P. Sheth and Amitava Das*
Combating disinformation is one of the burning societal crises - about 67% of the American population believes that disinformation produces a lot of uncertainty, and 10% of them knowingly propagate disinformation. Evidence shows that disinformation can manipulate democratic processes and public opinion, causing disruption in the share market, panic and anxiety in society, and even death during crises. Therefore, disinformation should be identified promptly and, if possible, mitigated. With approximately 3.2 billion images and 720,000 hours of video shared online daily on social media platforms, scalable detection of multimodal disinformation requires efficient fact verification. Despite progress in automatic text-based fact verification (e.g., FEVER, LIAR), the research community lacks substantial effort in multimodal fact verification. To address this gap, we introduce FACTIFY 3M, a dataset of 3 million samples that pushes the boundaries of the domain of fact verification via a multimodal fake news dataset, in addition to offering explainability through the concept of 5W question-answering. Salient features of the dataset include: (i) textual claims, (ii) ChatGPT-generated paraphrased claims, (iii) associated images, (iv) stable diffusion-generated additional images (i.e., visual paraphrases), (v) pixel-level image heatmap to foster image-text explainability of the claim, (vi) 5W QA pairs, and (vii) adversarial fake news stories.

16:00-17:30 (East Foyer / Virtual)
**#53 LM vs LM: Detecting Factual Errors via Cross Examination**
*Roi Cohen, May Hamri, Mor Geva and Amir Globerson*
A prominent weakness of modern language models (LMs) is their tendency to generate factually incorrect text, which hinders their usability. A natural question is whether such factual errors can be detected automatically. Inspired by truth-seeking mechanisms in law, we propose a factuality evaluation framework for LMs that is based on cross-examination. Our key idea is that an incorrect claim is likely to result in inconsistency with other claims that the model generates. To discover such inconsistencies, we facilitate a multi-turn interaction between the LM that generated the claim and another LM (acting as an examiner) which introduces questions to discover inconsistencies. We empirically evaluate our method on factual claims made by multiple recent LMs on four benchmarks, finding that it outperforms existing methods and baselines, often by a large gap. Our results demonstrate the potential of using interacting LMs for capturing factual errors.

16:00-17:30 (East Foyer / Virtual)
**#54 On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research**
*Luiza Amador Pozzobon, Beyza Ermis, Patrick Lewis and Sara Hooker*
Perception of toxicity evolves over time and often differs between geographies and cultural backgrounds. Similarly, black-box commercially available APIs for detecting toxicity, such as the Perspective API, are not static, but frequently retrained to address any unattended weaknesses and biases. We evaluate the implications of these changes on the reproducibility of findings that compare the relative merits of models and methods that aim to curb toxicity. Our findings suggest that research that relied on inherited automatic toxicity scores to compare models and techniques may have resulted in inaccurate findings. Rescoring all models from HELM, a widely respected living benchmark, for toxicity with the recent version of the API led to a different ranking of widely used foundation models. We suggest caution in applying apples-to-apples comparisons between studies and call for a more structured approach to evaluating toxicity over time.

16:00-17:30 (East Foyer / Virtual)
**#55 Natural Disaster Tweets Classification Using Multimodal Data**
*Mohammad Abdul Basit, Bashir Alam, Zubaida Fatima and Salman Shaikh*
Social media platforms are extensively used for expressing opinions or conveying information. The information available on such platforms can be used for various humanitarian and disaster-related tasks as distributing messages in different formats through social media is quick and easy. Often this useful information during disaster events goes to waste as efficient systems don't exist which can turn these unstructured data into meaningful format which can ultimately assist aid agencies. In disaster identification and assessment, information available is naturally multimodal, however, most existing work has been solely focused on single modalities e.g. images or texts separately. When information from different modalities are integrated , it produces significantly better results. In this paper, we have explored different models which can lead to the development of a system that deals with multimodal datasets and can perform sequential hierarchical classification. Specifically, we aim to find the damage and its severity along with classifying the data into humanitarian categories. The different stages in the hierarchical classification have had their respective models selected by researching with many different modality specific models and approaches of multimodal classification including multi task learning. The hierarchical model can give results at different abstraction levels according to the use cases. Through extensive quantitative and qualitative analysis, we show how our system is effective in classifying the multimodal tweets along with an excellent computational efficiency and assessment performance. With the help of our approach, we aim to support disaster management through identification of situations involving humanitarian tragedies and aid in assessing the severity and type of damage.

16:00-17:30 (East Foyer / Virtual)
**#56 IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions**
*Ziheng Zeng, Kellen Tan Cheng, Srihari Venkat Nanniyur, Jianing Zhou and Suma Bhat*
Idiomatic expression (IE) processing and comprehension have challenged pre-trained language models (PTLMs) because their meanings are non-compositional. Unlike prior works that enable IE comprehension through fine-tuning PTLMs with sentences containing IEs, in this work, we construct IEKG, a commonsense knowledge graph for figurative interpretations of IEs. This extends the established $ATOMIC^{20}_{20}$ converting PTLMs into knowledge models (KMs) that encode and infer commonsense knowledge related to IE use. Experiments show that various PTLMs can be converted into KMs with IEKG. We verify the quality of IEKG and the ability of the trained KMs with automatic and human evaluation. Through applications in natural language understanding, we show that a PTLM injected with knowledge from IEKG exhibits improved IE comprehension ability and can generalize to IEs unseen during training.

16:00-17:30 (East Foyer / Virtual)
**#57 StrAE: Autoencoding for Pre-Trained Embeddings using Explicit Structure**
*Mattia Opper, Victor Prokhorov and Siddharth N*
This work presents StrAE: a Structured Autoencoder framework that through strict adherence to explicit structure, and use of a novel contrastive objective over tree-structured representations, enables effective learning of multi-level representations. Through comparison over different forms of structure, we verify that our results are directly attributable to the informativeness of the structure provided as input, and

show that this is not the case for existing tree models. We then further extend StrAE to allow the model to define its own compositions using a simple localised-merge algorithm. This variant, called Self-StrAE, outperforms baselines that don't involve explicit hierarchical compositions, and is comparable to models given informative structure (e.g. constituency parses). Our experiments are conducted in a data-constrained (circa 10M tokens) setting to help tease apart the contribution of the inductive bias to effective learning. However, we find that this framework can be robust to scale, and when extended to a much larger dataset (circa 100M tokens), our 430 parameter model performs comparably to a 6-layer RoBERTa many orders of magnitude larger in size. Our findings support the utility of incorporating explicit composition as an inductive bias for effective representation learning.

16:00-17:30 (East Foyer / Virtual)
**#58 CoF-CoT: Enhancing Large Language Models with Coarse-to-Fine Chain-of-Thought Prompting for Multi-domain NLU Tasks**
*Hoang H Nguyen, Ye Liu, Chenwei Zhang, Tao Zhang and Philip S. Yu*
While Chain-of-Thought prompting is popular in reasoning tasks, its application to Large Language Models (LLMs) in Natural Language Understanding (NLU) is under-explored. Motivated by multi-step reasoning of LLMs, we propose Coarse-to-Fine Chain-of-Thought (CoF-CoT) approach that breaks down NLU tasks into multiple reasoning steps where LLMs can learn to acquire and leverage essential concepts to solve tasks from different granularities. Moreover, we propose leveraging semantic-based Abstract Meaning Representation (AMR) structured knowledge as an intermediate step to capture the nuances and diverse structures of utterances, and to understand connections between their varying levels of granularity. Our proposed approach is demonstrated effective in assisting the LLMs adapt to the multi-grained NLU tasks under both zero-shot and few-shot multi-domain settings.

16:00-17:30 (East Foyer / Virtual)
**#59 TOD-Flow: Modeling the Structure of Task-Oriented Dialogues**
*Sungryull Sohn, Yiwei Lyu, Anthony Zhe Liu, Lajanugen Logeswaran, Dong-Ki Kim, Dongsub Shim and Honglak Lee*
Task-Oriented Dialogue (TOD) systems have become crucial components in interactive artificial intelligence applications. While recent advances have capitalized on pre-trained language models (PLMs), they exhibit limitations regarding transparency and controllability. To address these challenges, we propose a novel approach focusing on inferring the TOD-flow graph from dialogue data annotated with dialog acts, uncovering the underlying task structure in the form of a graph. The inferred TOD-flow graph can be easily integrated with any dialogue model to improve its prediction performance, transparency, and controllability. Our TOD-flow graph learns what a model can, should, and should not predict, effectively reducing the search space and providing a rationale for the model's prediction. We show that the proposed TOD-flow graph better resemble human-annotated graphs compared to prior approaches. Furthermore, when combined with several dialogue policies and end-to-end dialogue models, we demonstrate that our approach significantly improves dialog act classification and end-to-end response generation performance in the MultiWOZ and SGD benchmarks.

16:00-17:30 (East Foyer / Virtual)
**#60 Accented Speech Recognition With Accent-specific Codebooks**
*Darshan Deepak Prabhu, Preethi Jyothi, Sriram Ganapathy and Vinit Unni*
Speech accents pose a significant challenge to state-of-the-art automatic speech recognition (ASR) systems. Degradation in performance across underrepresented accents is a severe deterrent to the inclusive adoption of ASR. In this work, we propose a novel accent adaptation approach for end-to-end ASR systems using cross-attention with a trainable set of codebooks. These learnable codebooks capture accent-specific information and are integrated within the ASR encoder layers. The model is trained on accented English speech, while the test data also contained accents which were not seen during training. On the Mozilla Common Voice multi-accented dataset, we show that our proposed approach yields significant performance gains not only on the seen English accents (up to 37% relative improvement in word error rate) but also on the unseen accents (up to 5% relative improvement in WER). Further, we illustrate benefits for a zero-shot transfer setup on the L2Artic dataset. We also compare the performance with other approaches based on accent adversarial training.

16:00-17:30 (East Foyer / Virtual)
**#61 Ideology Takes Multiple Looks: A High-Quality Dataset for Multifaceted Ideology Detection**
*Songtao Liu, Ziling Luo, Minghua Xu, Lixiao Wei, Ziyao Wei, Han Yu, Wei Xiang and Bang Wang*
Ideology detection (ID) is important for gaining insights about peoples' opinions and stances on our world and society, which can find many applications in politics, economics and social sciences. It is not uncommon that a piece of text can contain descriptions of various issues. It is also widely accepted that a person can take different ideological stances in different facets. However, existing datasets for the ID task only label a text as ideologically left- or right-leaning as a whole, regardless whether the text containing one or more different issues. Moreover, most prior work annotates texts from data resources with known ideological bias through distant supervision approaches, which may result in many false labels. With some theoretical help from social sciences, this work first designs an ideological schema containing five domains and twelve facets for a new multifaceted ideology detection (MID) task to provide a more complete and delicate description of ideology. We construct a MITweet dataset for the MID task, which contains 12,594 English Twitter posts, each annotated with a Relevance and an Ideology label for all twelve facets. We also design and test a few of strong baselines for the MID task under in-topic and cross-topic settings, which can serve as benchmarks for further research.

16:00-17:30 (East Foyer / Virtual)
**#62 A Challenging Multimodal Video Summary: Simultaneously Extracting and Generating Keyframe-Caption Pairs from Video**
*Keito Kudo, Haruki Nagasawa, Jun Suzuki and Nobuyuki Shimizu*
This paper proposes a practical multimodal video summarization task setting and a dataset to train and evaluate the task. The target task involves summarizing a given video into a predefined number of keyframe-caption pairs and displaying them in a listable format to grasp the video content quickly. This task aims to extract crucial scenes from the video in the form of images (keyframes) and generate corresponding captions explaining each keyframe's situation. This task is useful as a practical application and presents a highly challenging problem worthy of study. Specifically, achieving simultaneous optimization of the keyframe selection performance and caption quality necessitates careful consideration of the mutual dependence on both preceding and subsequent keyframes and captions. To facilitate subsequent research in this field, we also construct a dataset by expanding upon existing datasets and propose an evaluation framework. Furthermore, we develop two baseline systems and report their respective performance.

16:00-17:30 (East Foyer / Virtual)
**#63 Semantic Space Grounded Weighted Decoding for Multi-Attribute Controllable Dialogue Generation**
*Zhiling Zhang, Mengyue Wu and Kenny Q. Zhu*
Controlling chatbot utterance generation with multiple attributes such as personalities, emotions and dialogue acts is a practically useful but under-studied problem. We propose a novel framework called DASC that possesses strong controllability with a weighted decoding paradigm, while improving generation quality with the grounding in an attribute semantics space. Generation with multiple attributes is then intuitively implemented with an interpolation of multiple attribute embeddings, which results in substantial reduction in the model sizes. Experiments show that DASC can achieve high control accuracy in generation task with the simultaneous control of 3 aspects while also producing interesting and reasonably sensible responses, even in an out-of-distribution robustness test.

16:00-17:30 (East Foyer / Virtual)
**#64 CESAR: Automatic Induction of Compositional Instructions for Multi-turn Dialogs**
*Taha Aksu, Devamanyu Hazarika, Shikib Mehri, Seokhwan Kim, Dilek Hakkani-Tur, Yang Liu and Mahdi Namazifar*
Instruction-based multitasking has played a critical role in the success of large language models (LLMs) in multi-turn dialog applications. While publicly available LLMs have shown promising performance, when exposed to complex instructions with multiple constraints, they lag against state-of-the-art models like ChatGPT. In this work, we hypothesize that the availability of large-scale complex demonstrations is crucial in bridging this gap. Focusing on dialog applications, we propose a novel framework, CESAR, that unifies a large number of dialog tasks in the same format and allows programmatic induction of complex instructions without any manual effort. We apply CESAR on InstructDial, a benchmark for instruction-based dialog tasks. We further enhance InstructDial with new datasets and tasks and utilize CESAR to induce complex tasks with compositional instructions. This results in a new benchmark called InstructDial++, which includes 63 datasets with 86 basic tasks and 68 composite tasks. Through rigorous experiments, we demonstrate the scalability of CESAR in providing rich instructions. Models trained on InstructDial++ can follow compositional prompts, such as prompts that ask for multiple stylistic constraints.

16:00-17:30 (East Foyer / Virtual)
**#65 Training Simultaneous Speech Translation with Robust and Random Wait-k-Tokens Strategy**
*Linlin Zhang, Kai Fan, Jiajun Bu and Zhongqiang Huang*
Simultaneous Speech Translation (SimulST) is a task focused on ensuring high-quality translation of speech in low-latency situations. Despite this, the modality gap (*e.g.*, unknown word boundaries) between audio and text presents a challenge. This gap hinders the effective application of policies from simultaneous text translation (SimulMT) and compromises the performance of offline speech translation. To address this issue, we first leverage the Montreal Forced Aligner (MFA) and utilize audio transcription pairs in pre-training the acoustic encoder, and introduce a token-level cross-modal alignment that allows the wait-$k$ policy from SimulMT to better adapt to SimulST. This token-level boundary alignment simplifies the decision-making process for predicting read/write actions, as if the decoder were directly processing text tokens. Subsequently, to optimize the SimulST task, we propose a robust and random wait-$k$-tokens strategy. This strategy allows a single model to meet various latency requirements and minimizes error accumulation of boundary alignment during inference. Our experiments on the MuST-C dataset show that our method achieves better trade-off between translation quality and latency.

16:00-17:30 (East Foyer / Virtual)
**#66 Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory**
*Ziang Xiao, Susu Zhang, Vivian Lai and Q. Vera Liao*
We address a fundamental challenge in Natural Language Generation (NLG) model evaluation—the design and evaluation of evaluation metrics. Recognizing the limitations of existing automatic metrics and noises from how current human evaluation was conducted, we propose MetricEval, a framework informed by measurement theory, the foundation of educational test design, for conceptualizing and evaluating the reliability and validity of NLG evaluation metrics. The framework formalizes the source of measurement error and offers statistical tools for evaluating evaluation metrics based on empirical data. With our framework, one can quantify the uncertainty of the metrics to better interpret the result. To exemplify the use of our framework in practice, we analyzed a set of evaluation metrics for summarization and identified issues related to conflated validity structure in human-eval and reliability in LLM-based metrics. Through MetricEval, we aim to promote the design, evaluation, and interpretation of valid and reliable metrics to advance robust and effective NLG models.

16:00-17:30 (East Foyer / Virtual)
**#67 CLEME: Debiasing Multi-reference Evaluation for Grammatical Error Correction**
*Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng and Ying Shen*
Evaluating the performance of Grammatical Error Correction (GEC) systems is a challenging task due to its subjectivity. Designing an evaluation metric that is as objective as possible is crucial to the development of GEC task. However, mainstream evaluation metrics, i.e., reference-based metrics, introduce bias into the multi-reference evaluation by extracting edits without considering the presence of multiple references. To overcome this issue, we propose Chunk-LE Multi-reference Evaluation (CLEME), designed to evaluate GEC systems in the multi-reference evaluation setting. CLEME builds chunk sequences with consistent boundaries for the source, the hypothesis and references, thus eliminating the bias caused by inconsistent edit boundaries. Furthermore, we observe the consistent boundary could also act as the boundary of grammatical errors, based on which the $F_{0.5}$ score is then computed following the correction independence assumption. We conduct experiments on six English reference sets based on the CoNLL-2014 shared task. Extensive experiments and detailed analyses demonstrate the correctness of our discovery and the effectiveness of CLEME. Further analysis reveals that CLEME is robust to evaluate GEC systems across reference sets with varying numbers of references and annotation styles. All the source codes of CLEME are released at https://github.com/THUKElab/CLEME.

16:00-17:30 (East Foyer / Virtual)
**#68 Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models**
*Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang and Ji-Rong Wen*
The recent success of large language models (LLMs) has shown great potential to develop more powerful conversational recommender systems (CRSs), which rely on natural language conversations to satisfy user needs. In this paper, we embark on an investigation into the utilization of ChatGPT for CRSs, revealing the inadequacy of the existing evaluation protocol. It might overemphasize the matching with ground-truth items annotated by humans while neglecting the interactive nature of CRSs. To overcome the limitation, we further propose an **i**nteractive **Eva**luation approach based on **L**L**M**s, named **iEvaLM**, which harnesses LLM-based user simulators. Our evaluation approach can simulate various system-user interaction scenarios. Through the experiments on two public CRS datasets, we demonstrate notable improvements compared to the prevailing evaluation protocol. Furthermore, we emphasize the evaluation of explainability, and ChatGPT showcases persuasive explanation generation for its recommendations. Our study contributes to a deeper comprehension of the untapped potential of LLMs for CRSs and provides a more flexible and realistic evaluation approach for future research about LLM-based CRSs.

16:00-17:30 (East Foyer / Virtual)
**#69 Query Rewriting in Retrieval-Augmented Large Language Models**
*Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao and Nan Duan*
Large Language Models (LLMs) play powerful, black-box readers in the retrieve-then-read pipeline, making remarkable progress in knowledge-intensive tasks. This work introduces a new framework, Rewrite-Retrieve-Read instead of the previous retrieve-then-read for the retrieval-augmented LLMs from the perspective of the query rewriting. Unlike prior studies focusing on adapting either the retriever or the reader, our approach pays attention to the adaptation of the search query itself, for there is inevitably a gap between the input text and the needed knowledge in retrieval. We first prompt an LLM to generate the query, then use a web search engine to retrieve contexts. Furthermore, to better align the query to the frozen modules, we propose a trainable scheme for our pipeline. A small language model is adopted as a trainable rewriter to cater to the black-box LLM reader. The rewriter is trained using the feedback of the LLM reader by reinforcement learning. Evaluation is conducted on downstream tasks, open-domain QA and multiple-choice QA. Experiments results show consistent performance improvement, indicating that our framework is proven effective and scalable, and brings a new framework for retrieval-augmented LLM.

16:00-17:30 (East Foyer / Virtual)
**#70 A Framework for Vision-Language Warm-up Tasks in Multimodal Dialogue Models**
*Jaewook Lee, Seongsik Park, Seong-Heum Park, Hongjin Kim and Harksoo Kim*
Most research on multimodal open-domain dialogue agents has focused on pretraining and multi-task learning using additional rich datasets beyond a given target dataset. However, methods for exploiting these additional datasets can be quite limited in real-world settings, creating a need for more efficient methods for constructing agents based solely on the target dataset. To address these issues, we present a new learning strategy called vision-language warm-up tasks for multimodal dialogue models (VLAW-MDM). This strategy does not require the use of large pretraining or multi-task datasets but rather relies solely on learning from target data. Moreover, our proposed approach automatically generate captions for images and incorporate them into the model's input to improve the contextualization of visual information. Using this novel approach, we empirically demonstrate that our learning strategy is effective for limited data and relatively small models. The result show that our method achieved comparable and in some cases superior performance compared to existing state-of-the-art models on various evaluation metrics.

16:00-17:30 (East Foyer / Virtual)
**#71 Causal Document-Grounded Dialogue Pre-training**
*Yingxiu Zhao, Bowen Yu, Bowen Li, Haiyang Yu, Jinyang Li, Chao Wang, Fei Huang, Yongbin Li and Nevin L. Zhang*
The goal of document-grounded dialogue (DocGD) is to generate a response by anchoring the evidence in a supporting document in accordance with the dialogue context. This entails four causally interconnected variables. While task-specific pre-training has significantly enhanced performances on numerous downstream tasks, existing DocGD methods still rely on general pre-trained language models without a specifically tailored pre-training approach that explicitly captures the causal relationships. To address this, we present the first causally-complete dataset construction strategy for developing million-scale DocGD pre-training corpora. Additionally, we propose a causally-perturbed pre-training strategy to better capture causality by introducing perturbations on the variables and optimizing the overall causal effect. Experiments conducted on three benchmark datasets demonstrate that our causal pre-training yields substantial and consistent improvements in fully-supervised, low-resource, few-shot, and zero-shot settings.

16:00-17:30 (East Foyer / Virtual)
**#72 API-Assisted Code Generation for Question Answering on Varied Table Structures**
*Yihan Cao, Shuyi Chen, Ryan Liu, Zhiruo Wang and Daniel Fried*
A persistent challenge to table question answering (TableQA) by generating executable programs has been adapting to varied table structures, typically requiring domain-specific logical forms. In response, this paper introduces a unified TableQA framework that: (1) provides a unified representation for structured tables as multi-index Pandas data frames, (2) uses Python as a powerful querying language, and (3) uses few-shot prompting to translate NL questions into Python programs, which are executable on Pandas data frames. Furthermore, to answer complex relational questions with extended program functionality and external knowledge, our framework allows customized APIs that Python programs can call. We experiment with four TableQA datasets that involve tables of different structures — relational, multi-table, and hierarchical matrix shapes — and achieve prominent improvements over past state-of-the-art systems. In ablation studies, we (1) show benefits from our multi-index representation and APIs over baselines that use only an LLM, and (2) demonstrate that our approach is modular and can incorporate additional APIs.

16:00-17:30 (East Foyer / Virtual)
**#73 OssCSE: Overcoming Surface Structure Bias in Contrastive Learning for Unsupervised Sentence Embedding**
*Zhan Shi, Guoyin Wang, Ke Bai, Jiwei Li, Xiang Li, Qingjun Cui, Belinda Zeng, Trishul Chilimbi and Xiaodan Zhu*
Contrastive learning has been demonstrated effective in unsupervised sentence representation learning. Given one sentence, positive pairs are obtained by passing the sentence to the encoder twice using the different dropout masks, and negative pairs are obtained by taking another sentence in the same mini-batch. However, the method suffers from the surface structure bias, i.e., sentences with similar surface structures will be regarded as close in semantics while sentences with dissimilar surface structures will be viewed as distinct in semantics. This leads to the result that paraphrasing a sentence that is dissimilar in surface structure will receive a lower semantic similarity score than inserting a negative word into the sentence. In this paper, we first verify the bias by collecting a sentence transformation testset. Then we systematically probe the existing models by proposing novel splits based on benchmark datasets in accordance with semantic and surface structure similarity. We tackle the bias in two aspects: balancing the learning target by augmenting with data that counters the bias, and meanwhile preserving word semantics by leveraging recall loss to prevent catastrophic forgetting. We evaluate our model on standard semantic textual similarity (STS) tasks using different pre-trained backbones and achieve state-of-the-art averaged performance across the STS benchmarks. Particularly, our models that are fine-tuned with $RoBERTa_{base}$ and $RoBERTa_{large}$ achieve significantly better performance on most benchmark datasets.

16:00-17:30 (East Foyer / Virtual)
**#74 M$^3$Seg: A Maximum-Minimum Mutual Information Paradigm for Unsupervised Topic Segmentation in ASR Transcripts**
*Ke Wang, Xiutian Zhao, Yanghui Li and Wei Peng*
Topic segmentation aims to detect topic boundaries and split automatic speech recognition transcriptions (e.g., meeting transcripts) into segments that are bounded by thematic meanings. In this work, we propose M$^3$Seg, a novel Maximum-Minimum Mutual information paradigm for linear topic segmentation without using any parallel data. Specifically, by employing sentence representations provided by pre-trained language models, M$^3$Seg first learns a region-based segment encoder based on the maximization of mutual information between the global segment representation and the local contextual sentence representation. Secondly, an edge-based boundary detection module aims to segment the whole by topics based on minimizing the mutual information between different segments. Experiment results on two public datasets demonstrate the effectiveness of M$^3$Seg, which outperform the state-of-the-art methods by a significant (18%–37% improvement) margin.

16:00-17:30 (East Foyer / Virtual)
**#75 ART: rule bAsed futuRe-inference deducTion**
*Mengze Li, Tianqi Zhao, Bai Jionghao, Baoyi He, Jiaxu Miao, Wei Ji, Zheqi Lv, Zhou Zhao, Shengyu Zhang, Wenqiao Zhang and Fei Wu*
Deductive reasoning is a crucial cognitive ability of humanity, allowing us to derive valid conclusions from premises and observations. However, existing works mainly focus on language-based premises and generally neglect deductive reasoning from visual observations. In this work, we introduce rule bAsed futuRe-inference deducTion (ART), which aims at deducing the correct future event based on the visual phenomenon (a video) and the rule-based premises, along with an explanation of the reasoning process. To advance this field, we construct a large-scale densely annotated dataset (Video-ART), where the premises, future event candidates, the reasoning process explanation, and auxiliary commonsense knowledge (e.g., actions and appearance) are annotated by annotators. Upon Video-ART, we develop a strong baseline named ARTNet. In essence, guided by commonsense knowledge, ARTNet learns to identify the target video character and perceives its visual clues related to the future event. Then, ARTNet rigorously applies the given premises to conduct reasoning from the identified information to future events, through a non-parametric rule reasoning network and a reasoning-path review module. Empirical studies validate the rationality

of ARTNet in deductive reasoning upon visual observations and the effectiveness over existing works.

16:00-17:30 (East Foyer / Virtual)
**#76 RoBoCoP: A Comprehensive ROmance BOrrowing COgnate Package and Benchmark for Multilingual Cognate Identification**
*Liviu P Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Daniela Dinu, Ioan-Bogdan Iordache, Simona Georgescu and Laurentiu Zoicas*
The identification of cognates is a fundamental process in historical linguistics, on which any further research is based. Even though there are several cognate databases for Romance languages, they are rather scattered, incomplete, noisy, contain unreliable information, or have uncertain availability. In this paper we introduce a comprehensive database of Romance cognates and borrowings based on the etymological information provided by the dictionaries. We extract pairs of cognates between any two Romance languages by parsing electronic dictionaries of Romanian, Italian, Spanish, Portuguese and French. Based on this resource, we propose a strong benchmark for the automatic detection of cognates, by applying machine learning and deep learning based methods on any two pairs of Romance languages. We find that automatic identification of cognates is possible with accuracy averaging around 94% for the more difficult task formulations.

16:00-17:30 (East Foyer / Virtual)
**#77 CS2W: A Chinese Spoken-to-Written Style Conversion Dataset with Multiple Conversion Types**
*Zishan Guo, Linhao Yu, Minghui Xu, Renren Jin and Deyi Xiong*
Spoken texts (either manual or automatic transcriptions from automatic speech recognition (ASR)) often contain disfluencies and grammatical errors, which pose tremendous challenges to downstream tasks. Converting spoken into written language is hence desirable. Unfortunately, the availability of datasets for this is limited. To address this issue, we present CS2W, a Chinese Spoken-to-Written style conversion dataset comprising 7,237 spoken sentences extracted from transcribed conversational texts. Four types of conversion problems are covered in CS2W: disfluencies, grammatical errors, ASR transcription errors, and colloquial words. Our annotation convention, data, and code are publicly available at https://github.com/guozishan/CS2W.

16:00-17:30 (East Foyer / Virtual)
**#78 StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding**
*Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang and Zheng Zhang*
Analogy-making between narratives is crucial for human reasoning. In this paper, we evaluate the ability to identify and generate analogies by constructing a first-of-its-kind large-scale story-level analogy corpus, STORYANALOGY, which contains 24K story pairs from diverse domains with human annotations on two similarities from the extended Structure-Mapping Theory. We design a set of tests on STORYANALOGY, presenting the first evaluation of story-level analogy identification and generation. Interestingly, we find that the analogy identification tasks are incredibly difficult not only for sentence embedding models but also for the recent large language models (LLMs) such as ChatGPT and LLaMa. ChatGPT, for example, only achieved around 30% accuracy in multiple-choice questions (compared to over 85% accuracy for humans). Furthermore, we observe that the data in STORYANALOGY can improve the quality of analogy generation in LLMs, where a fine-tuned FlanT5-xxl model achieves comparable performance to zero-shot ChatGPT.

16:00-17:30 (East Foyer / Virtual)
**#79 The ACL OCL Corpus: Advancing Open Science in Computational Linguistics**
*Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Anand Unnithan and Min-Yen Kan*
We present ACL OCL, a scholarly corpus derived from the ACL Anthology to assist Open scientific research in the Computational Linguistics domain. Integrating and enhancing the previous versions of the ACL Anthology, the ACL OCL contributes metadata, PDF files, citation graphs and additional structured full texts with sections, figures, and links to a large knowledge resource (Semantic Scholar). The ACL OCL spans seven decades, containing 73K papers, alongside 210K figures. We spotlight how ACL OCL applies to observe trends in computational linguistics. By detecting paper topics with a supervised neural model, we note that interest in "Syntax: Tagging, Chunking and Parsing" is waning and "Natural Language Generation" is resurging. Our dataset is available from HuggingFace (https://huggingface.co/datasets/WINGNUS/ACL-OCL).

16:00-17:30 (East Foyer / Virtual)
**#80 BLESS: Benchmarking Large Language Models on Sentence Simplification**
*Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego and Matthew Shardlow*
We present BLESS, a comprehensive performance benchmark of the most recent state-of-the-art Large Language Models (LLMs) on the task of text simplification (TS). We examine how well off-the-shelf LLMs can solve this challenging task, assessing a total of 44 models, differing in size, architecture, pre-training methods, and accessibility, on three test sets from different domains (Wikipedia, news, and medical) under a few-shot setting. Our analysis considers a suite of automatic metrics, as well as a large-scale quantitative investigation into the types of common edit operations performed by the different models. Furthermore, we perform a manual qualitative analysis on a subset of model outputs to better gauge the quality of the generated simplifications. Our evaluation indicates that the best LLMs, despite not being trained on TS perform comparably with state-of-the-art TS baselines. Additionally, we find that certain LLMs demonstrate a greater range and diversity of edit operations. Our performance benchmark will be available as a resource for the development of future TS methods and evaluation metrics.

16:00-17:30 (East Foyer / Virtual)
**#81 mRedditSum: A Multimodal Abstractive Summarization Dataset of Reddit Threads with Images**
*Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park and Gunhee Kim*
The growing number of multimodal online discussions necessitates automatic summarization to save time and reduce content overload. However, existing summarization datasets are not suitable for this purpose, as they either do not cover discussions, multiple modalities, or both. To this end, we present mRedditSum, the first multimodal discussion summarization dataset. It consists of 3,033 discussion threads where a post solicits advice regarding an issue described with an image and text, and respective comments express diverse opinions. We annotate each thread with a human-written summary that captures both the essential information from the text, as well as the details available only in the image. Experiments show that popular summarization models—GPT-3.5, BART, and T5—consistently improve in performance when visual information is incorporated. We also introduce a novel method, cluster-based multi-stage summarization, that outperforms existing baselines and serves as a competitive baseline for future work.

16:00-17:30 (East Foyer / Virtual)
**#82 DiffS2UT: A Semantic Preserving Diffusion Model for Textless Direct Speech-to-Speech Translation**
*Yongxin Zhu, Zhujin Gao, Xinyuan Zhou, Ye Zhongyi and Linli Xu*
While Diffusion Generative Models have achieved great success on image generation tasks, how to efficiently and effectively incorporate them into speech generation especially translation tasks remains a non-trivial problem. Specifically, due to the low information density of speech data, the transformed discrete speech unit sequence is much longer than the corresponding text transcription, posing significant challenges to existing auto-regressive models. Furthermore, it is not optimal to brutally apply discrete diffusion on the speech unit sequence while

disregarding the continuous space structure, which will degrade the generation performance significantly. In this paper, we propose a novel diffusion model by applying the diffusion forward process in the continuous speech representation space, while employing the diffusion backward process in the discrete speech unit space. In this way, we preserve the semantic structure of the continuous speech representation space in the diffusion process and integrate the continuous and discrete diffusion models. We conduct extensive experiments on the textless direct speech-to-speech translation task, where the proposed method achieves comparable results to the computationally intensive auto-regressive baselines (500 steps on average) with significantly fewer decoding steps (50 steps).

16:00-17:30 (East Foyer / Virtual)
**#83 HutCRS: Hierarchical User-Interest Tracking for Conversational Recommender System**
*Mingjie Qian, Yongsen Zheng, Jinghui Qin and Liang Lin*
Conversational Recommender System (CRS) aims to explicitly acquire user preferences towards items and attributes through natural language conversations. However, existing CRS methods ask users to provide explicit answers (yes/no) for each attribute they require, regardless of users' knowledge or interest, which may significantly reduce the user experience and semantic consistency. Furthermore, these methods assume that users like all attributes of the target item and dislike those unrelated to it, which can introduce bias in attribute-level feedback and impede the system's ability to accurately identify the target item. To address these issues, we propose a more realistic, user-friendly, and explainable CRS framework called Hierarchical User-Interest Tracking for Conversational Recommender System (HutCRS). HutCRS portrays the conversation as a hierarchical interest tree that consists of two stages. In stage I, the system identifies the aspects that the user prefers while the system asks about attributes related to these positive aspects or recommends items in stage II. In addition, we develop a Hierarchical-Interest Policy Learning (HIPL) module to integrate the decision-making process of which aspects to ask and when to ask about attributes or recommend items. Moreover, we classify the attribute-level feedback results to further enhance the system's ability to capture special information, such as attribute instances that are accepted by users but not presented in their historical interactive data. Extensive experiments on four benchmark datasets demonstrate the superiority of our method. The implementation of HutCRS is publicly available at https://github.com/xinle1129/HutCRS.

16:00-17:30 (East Foyer / Virtual)
**#84 Conversational Semantic Parsing using Dynamic Context Graphs**
*Parag Jain and Mirella Lapata*
In this paper we consider the task of conversational semantic parsing over general purpose knowledge graphs (KGs) with millions of entities, and thousands of relation-types. We focus on models which are capable of interactively mapping user utterances into executable logical forms (e.g., Sparql) in the context of the conversational history. Our key idea is to represent information about an utterance and its context via a subgraph which is created dynamically, i.e., the number of nodes varies per utterance. Rather than treating the subgraph as a sequence, we exploit its underlying structure and encode it with a graph neural network which further allows us to represent a large number of (unseen) nodes. Experimental results show that dynamic context modeling is superior to static approaches, delivering performance improvements across the board (i.e., for simple and complex questions). Our results further confirm that modeling the structure of context is better at processing discourse information, (i.e., at handling ellipsis and resolving coreference) and longer interactions.

16:00-17:30 (East Foyer / Virtual)
**#85 Back Transcription as a Method for Evaluating Robustness of Natural Language Understanding Models to Speech Recognition Errors**
*Marek Kubis, Paweł Marek Skórzewski, Marcin Sowański and Tomasz Ziętkiewicz*
In a spoken dialogue system, an NLU model is preceded by a speech recognition system that can deteriorate the performance of natural language understanding. This paper proposes a method for investigating the impact of speech recognition errors on the performance of natural language understanding models. The proposed method combines the back transcription procedure with a fine-grained technique for categorizing the errors that affect the performance of NLU models. The method relies on the usage of synthesized speech for NLU evaluation. We show that the use of synthesized speech in place of audio recording does not change the outcomes of the presented technique in a significant way.

16:00-17:30 (East Foyer / Virtual)
**#86 Large Language Models are Complex Table Parsers**
*Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng and Xiaobo Zhang*
With the Generative Pre-trained Transformer 3.5 (GPT-3.5) exhibiting remarkable reasoning and comprehension abilities in Natural Language Processing (NLP), most Question Answering (QA) research has primarily centered around general QA tasks based on GPT, neglecting the specific challenges posed by Complex Table QA. In this paper, we propose to incorporate GPT-3.5 to address such challenges, in which complex tables are reconstructed into tuples and specific prompt designs are employed for dialogues. Specifically, we encode each cell's hierarchical structure, position information, and content as a tuple. By enhancing the prompt template with an explanatory description of the meaning of each tuple and the logical reasoning process of the task, we effectively improve the hierarchical structure awareness capability of GPT-3.5 to better parse the complex tables. Extensive experiments and results on Complex Table QA datasets, i.e., the open-domain dataset HiTAB and the aviation domain dataset AIT-QA show that our approach significantly outperforms previous work on both datasets, leading to state-of-the-art (SOTA) performance.

16:00-17:30 (East Foyer / Virtual)
**#87 Dialogizer: Context-aware Conversational-QA Dataset Generation from Textual Sources**
*Yerin Hwang, Yongil Kim, Hyunkyung Bae, Hwanhee Lee, Jeesoo Bang and Kyomin Jung*
To address the data scarcity issue in Conversational question answering (ConvQA), a dialog inpainting method, which utilizes documents to generate ConvQA datasets, has been proposed. However, the original dialog inpainting model is trained solely on the dialog reconstruction task, resulting in the generation of questions with low contextual relevance due to insufficient learning of question-answer alignment. To overcome this limitation, we propose a novel framework called Dialogizer, which has the capability to automatically generate ConvQA datasets with high contextual relevance from textual sources. The framework incorporates two training tasks: question-answer matching (QAM) and topic-aware dialog generation (TDG). Moreover, re-ranking is conducted during the inference phase based on the contextual relevance of the generated questions. Using our framework, we produce four ConvQA datasets by utilizing documents from multiple domains as the primary source. Through automatic evaluation using diverse metrics, as well as human evaluation, we validate that our proposed framework exhibits the ability to generate datasets of higher quality compared to the baseline dialog inpainting model.

16:00-17:30 (East Foyer / Virtual)
**#88 Graph vs. Sequence: An Empirical Study on Knowledge Forms for Knowledge-Grounded Dialogue**
*Yizhe Yang, Heyan Huang, Yuhang Liu and Yang Gao*
Knowledge-grounded dialogue is a task of gener- ating an informative response based on both the dialogue history and external knowledge source. In general, there are two forms of knowledge: manu- ally annotated knowledge graphs and knowledge text from website. From various evaluation viewpoints, each type of knowledge has advantages and downsides. To further distinguish the principles and determinants

from the intricate factors, we conduct a thorough experiment and study on the task to answer three essential questions. The ques- tions involve the choice of appropriate knowledge form, the degree of mutual effects between knowl- edge and the model selection, and the few-shot performance of knowledge. Supported by statistical shreds of evidence, we offer conclusive solutions and sensible suggestions for directions and standards of future research.

16:00-17:30 (East Foyer / Virtual)
**#89 Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators**
*Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua and Kam-Fai Wong*
Large language models (LLMs) outperform information retrieval techniques for downstream knowledge-intensive tasks when being prompted to generate world knowledge. However, community concerns abound regarding the factuality and potential implications of using this uncensored knowledge. In light of this, we introduce CONNER, a COmpreheNsive kNowledge Evaluation fRamework, designed to systematically and automatically evaluate generated knowledge from six important perspectives – Factuality, Relevance, Coherence, Informativeness, Helpfulness and Validity. We conduct an extensive empirical analysis of the generated knowledge from three different types of LLMs on two widely studied knowledge-intensive tasks, i.e., open-domain question answering and knowledge-grounded dialogue. Surprisingly, our study reveals that the factuality of generated knowledge, even if lower, does not significantly hinder downstream tasks. Instead, the relevance and coherence of the outputs are more important than small factual mistakes. Further, we show how to use CONNER to improve knowledge-intensive tasks by designing two strategies: Prompt Engineering and Knowledge Selection. Our evaluation code and LLM-generated knowledge with human annotations will be released to facilitate future research.

16:00-17:30 (East Foyer / Virtual)
**#90 A Fair and In-Depth Evaluation of Existing End-to-End Entity Linking Systems**
*Hannah Bast, Matthias Hertel and Natalie Prange*
Existing evaluations of entity linking systems often say little about how the system is going to perform for a particular application. There are two fundamental reasons for this. One is that many evaluations only use aggregate measures (like precision, recall, and F1 score), without a detailed error analysis or a closer look at the results. The other is that all of the widely used benchmarks have strong biases and artifacts, in particular: a strong focus on named entities, an unclear or missing specification of what else counts as an entity mention, poor handling of ambiguities, and an over- or underrepresentation of certain kinds of entities. We provide a more meaningful and fair in-depth evaluation of a variety of existing end-to-end entity linkers. We characterize their strengths and weaknesses and also report on reproducibility aspects. The detailed results of our evaluation can be inspected under https://elevant.cs.uni-freiburg.de/emnlp2023. Our evaluation is based on several widely used benchmarks, which exhibit the problems mentioned above to various degrees, as well as on two new benchmarks, which address the problems mentioned above. The new benchmarks can be found under https://github.com/ad-freiburg/fair-entity-linking-benchmarks.

16:00-17:30 (East Foyer / Virtual)
**#91 Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models**
*Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park and Jaewoo Kang*
Questions in open-domain question answering are often ambiguous, allowing multiple interpretations. One approach to handling them is to identify all possible interpretations of the ambiguous question (AQ) and to generate a long-form answer addressing them all, as suggested by Stelmakh et al., (2022). While it provides a comprehensive response without bothering the user for clarification, considering multiple dimensions of ambiguity and gathering corresponding knowledge remains a challenge. To cope with the challenge, we propose a novel framework, Tree of Clarifications (ToC): It recursively constructs a tree of disambiguations for the AQ—via few-shot prompting leveraging external knowledge—and uses it to generate a long-form answer. ToC outperforms existing baselines on ASQA in a few-shot setup across the metrics, while surpassing fully-supervised baselines trained on the whole training set in terms of Disambig-F1 and Disambig-ROUGE. Code is available at https://github.com/gankim/tree-of-clarifications.

16:00-17:30 (East Foyer / Virtual)
**#92 Finding Authentic Counterhate Arguments: A Case Study with Public Figures**
*Abdullah Albanyan, Ahmed Hassan and Eduardo Blanco*
We explore authentic counterhate arguments for online hateful content toward individuals. Previous efforts are limited to counterhate to fight against hateful content toward groups. Thus, we present a corpus of 54,816 hateful tweet-paragraph pairs, where the paragraphs are candidate counterhate arguments. The counterhate arguments are retrieved from 2,500 online articles from multiple sources. We propose a methodology that assures the authenticity of the counter argument and its specificity to the individual of interest. We show that finding arguments in online articles is an efficient alternative to counterhate generation approaches that may hallucinate unsupported arguments. We also present linguistic insights on the language used in counterhate arguments. Experimental results show promising results. It is more challenging, however, to identify counterhate arguments for hateful content toward individuals not included in the training set.

16:00-17:30 (East Foyer / Virtual)
**#93 MarkQA: A large scale KBQA dataset with numerical reasoning**
*Xiang Huang, Sitao Cheng, Yuheng Bao, Shanshan Huang and Yuzhong Qu*
While question answering over knowledge bases (KBQA) has shown progress in addressing factoid questions, KBQA with numerical reasoning remains relatively unexplored. In this paper, we focus on the complex numerical reasoning in KBQA, and propose a new task, NR-KBQA, which necessitates the ability to perform both multi-hop reasoning and numerical reasoning. We also design a logic form in Python format called PyQL to represent the reasoning process of numerical reasoning questions. To facilitate the development of NR-KBQA, we present a large NR-KBQA dataset called MarkQA, which is automatically constructed by a small set of seeds. Each question in MarkQA is annotated with its corresponding SPARQL query, alongside the step-by-step reasoning path in the QDMR format and PyQL program. Experimental results of some state-of-the-art QA methods performed on the MarkQA dataset show that complex numerical reasoning in KBQA faces great challenges.

16:00-17:30 (East Foyer / Virtual)
**#94 Find-2-Find: Multitask Learning for Anaphora Resolution and Object Localization**
*Cennet Oguz, Pascal Denis, Emmanuel Vincent, Simon Ostermann and Josef van Genabith*
In multimodal understanding tasks, visual and linguistic ambiguities can arise. Visual ambiguity can occur when visual objects require a model to ground a referring expression in a video without strong supervision, while linguistic ambiguity can occur from changes in entities in action flows. As an example from the cooking domain, "oil" mixed with "salt" and "pepper" could later be referred to as a "mixture". Without a clear visual-linguistic alignment, we cannot know which among several objects shown is referred to by the language expression "mixture", and without resolved antecedents, we cannot pinpoint what the mixture is. We define this chicken-and-egg problem as Visual-linguistic Ambiguity. In this paper, we present Find2Find, a joint anaphora resolution and object localization dataset targeting the problem of *visual-linguistic ambiguity*, consisting of 500 anaphora-annotated recipes with corresponding videos. We present experimental results of a novel end-to-end joint multitask learning framework for Find2Find that fuses visual and textual information and shows improvements both for anaphora resolution and object localization with one joint model in multitask learning, as compared to a strong single-task baseline.

16:00-17:30 (East Foyer / Virtual)
**#95 Best of Both Worlds: Towards Improving Temporal Knowledge Base Question Answering via Targeted Fact Extraction**
*Nithish Kannen, Udit Sharma, Sumit Neelam, Dinesh Khandelwal, Shajith Ikbal, Hima Karanam and L Venkata Subramaniam*
Temporal question answering (QA) is a special category of complex question answering task that requires reasoning over facts asserting time intervals of events. Previous works have predominately relied on Knowledge Base Question Answering (KBQA) for temporal QA. One of the major challenges faced by these systems is their inability to retrieve all relevant facts due to factors such as incomplete KB and entity/relation linking errors. A failure to fetch even a single fact will block KBQA from computing the answer. Such cases of KB incompleteness are even more profound in the temporal context. To address this issue, we explore an interesting direction where a targeted temporal fact extraction technique is used to assist KBQA whenever it fails to retrieve temporal facts from the KB. We model the extraction problem as an open-domain question answering task using off-the-shelf language models. This way, we target to extract from textual resources those facts that failed to get retrieved from the KB. Experimental results on two temporal QA benchmarks show promising $\sim$30% & $\sim$10% relative improvements in answer accuracies without any additional training cost.

16:00-17:30 (East Foyer / Virtual)
**#96 Benchmarking and Improving Text-to-SQL Generation under Ambiguity**
*Adithya Bhaskar, Tushar Tomar, Ashutosh Sathe and Sunita Sarawagi*
Research in Text-to-SQL conversion has been largely benchmarked against datasets where each text query corresponds to one correct SQL. However, natural language queries over real-life databases frequently involve significant ambiguity about the intended SQL due to overlapping schema names and multiple confusing relationship paths. To bridge this gap, we develop a novel benchmark called AmbiQT with over 3000 examples where each text is interpretable as two plausible SQLs due to lexical and/or structural ambiguity. When faced with ambiguity, an ideal top-$k$ decoder should generate all valid interpretations for possible disambiguation by the user. We evaluate several Text-to-SQL systems and decoding algorithms, including those employing state-of-the-art LLMs, and find them to be far from this ideal. The primary reason is that the prevalent beam search algorithm and its variants, treat SQL queries as a string and produce unhelpful token-level diversity in the top-$k$. We propose LogicalBeam, a new decoding algorithm that navigates the SQL logic space using a blend of plan-based template generation and constrained infilling. Counterfactually generated plans diversify templates while in-filling with a beam-search that branches solely on schema names provides value diversity. LogicalBeam is up to $2.5$ times more effective than state-of-the-art models at generating all candidate SQLs in the top-$k$ ranked outputs. It also enhances the top-$5$ Exact and Execution Match Accuracies on SPIDER and Kaggle DBQA.

16:00-17:30 (East Foyer / Virtual)
**#97 From Parse-Execute to Parse-Execute-Refine: Improving Semantic Parser for Complex Question Answering over Knowledge Base**
*Wangzhen Guo, Linyin Luo, Hanjiang Lai and Jian Yin*
Parsing questions into executable logical forms has showed impressive results for knowledge-base question answering (KBQA). However, complex KBQA is a more challenging task that requires to perform complex multi-step reasoning. Recently, a new semantic parser called KoPL has been proposed to explicitly model the reasoning processes, which achieved the state-of-the-art on complex KBQA. In this paper, we further explore how to unlock the reasoning ability of semantic parsers by a simple proposed parse-execute-refine paradigm. We refine and improve the KoPL parser by demonstrating the executed intermediate reasoning steps to the KBQA model. We show that such simple strategy can significantly improve the ability of complex reasoning. Specifically, we propose three components: a parsing stage, an execution stage and a refinement stage, to enhance the ability of complex reasoning. The parser uses the KoPL to generate the transparent logical forms. Then, the execution stage aligns and executes the logical forms over knowledge base to obtain intermediate reasoning processes. Finally, the intermediate step-by-step reasoning processes are demonstrated to the KBQA model in the refinement stage. With the explicit reasoning processes, it is much easier to answer the complex questions. Experiments on benchmark dataset shows that the proposed PER-KBQA performs significantly better than the stage-of-the-art baselines on the complex KBQA.

16:00-17:30 (East Foyer / Virtual)
**#98 Diversify Question Generation with Retrieval-Augmented Style Transfer**
*Qi Gou, Zehua Xia, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li and Nguyen Cam-Tu*
Given a textual passage and an answer, humans are able to ask questions with various expressions, but this ability is still challenging for most question generation (QG) systems. Existing solutions mainly focus on the internal knowledge within the given passage or the semantic word space for diverse content planning. These methods, however, have not considered the potential of external knowledge for expression diversity. To bridge this gap, we propose RAST, a framework for Retrieval-Augmented Style Transfer, where the objective is to utilize the style of diverse templates for question generation. For training RAST, we develop a novel Reinforcement Learning (RL) based approach that maximizes a weighted combination of diversity reward and consistency reward. Here, the consistency reward is computed by a Question-Answering (QA) model, whereas the diversity reward measures how much the final output mimics the retrieved template. Experimental results show that our method outperforms previous diversity-driven baselines on diversity while being comparable in terms of consistency scores. Our code is available at https://github.com/gouqi666/RAST.

16:00-17:30 (East Foyer / Virtual)
**#99 PRCA: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter**
*Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li and Jing Xiao*
The Retrieval Question Answering (ReQA) task employs the retrieval-augmented framework, composed of a retriever and generator. The generators formulate the answer based on the documents retrieved by the retriever. Incorporating Large Language Models (LLMs) as generators is beneficial due to their advanced QA capabilities, but they are typically too large to be fine-tuned with budget constraints while some of them are only accessible via APIs. To tackle this issue and further improve ReQA performance, we propose a trainable Pluggable Reward-Driven Contextual Adapter (PRCA), keeping the generator as a black box. Positioned between the retriever and generator in a Pluggable manner, PRCA refines the retrieved information by operating in a token-autoregressive strategy via maximizing rewards of the reinforcement learning phase. Our experiments validate PRCA's effectiveness in enhancing ReQA performance on three datasets by up to 20% improvement to fit black-box LLMs into existing frameworks, demonstrating its considerable potential in the LLMs era.

16:00-17:30 (East Foyer / Virtual)
**#100 Re$^3$Dial: Retrieve, Reorganize and Rescale Conversations for Long-Turn Open-Domain Dialogue Pre-training**
*Jiaxin Wen, Hao Zhou, Jian Guan, Jie Zhou and Minlie Huang*
Pre-training on large-scale open-domain dialogue data can substantially improve the performance of dialogue models. However, the pre-trained dialogue model's ability to utilize long-range context is limited due to the scarcity of long-turn dialogue sessions. Most dialogues in existing pre-training corpora contain fewer than three turns of dialogue. To alleviate this issue, we propose the Retrieve, Reorganize and

Rescale framework (Re$^3$Dial), which can automatically construct billion-scale long-turn dialogues by reorganizing existing short-turn ones. Given a short-turn session, Re$^3$Dial first employs a session retriever to retrieve coherent consecutive sessions. To this end, we train the retriever to capture semantic and discourse relations within multi-turn dialogues through contrastive training. Next, Re$^3$Dial samples a session from retrieved results following a diversity sampling strategy, which is designed to penalize repetitive or generic sessions. A longer session is then derived by concatenating the original session and the sampled session. By repeating the above process, Re$^3$Dial can yield a coherent long-turn dialogue. Extensive experiments on multiple multi-turn dialogue benchmarks demonstrate that Re$^3$Dial significantly improves the dialogue model's ability to utilize long-range context and thus generate more sensible and informative responses. Finally, we build a toolkit for efficiently rescaling conversations with Re$^3$Dial, which enables us to construct a corpus containing 1B Chinese dialogue sessions with 11.3 turns on average (5X longer than the original corpus). We will release our retriever model, toolkit, and data for public use.

16:00-17:30 (East Foyer / Virtual)
**#101 SEER : A Knapsack approach to Exemplar Selection for In-Context HybridQA**
*Jonathan Tonglet, Manon Reusens, Philipp Borchert and Bart Baesens*
Question answering over hybrid contexts is a complex task, which requires the combination of information extracted from unstructured texts and structured tables in various ways. Recently, In-Context Learning demonstrated significant performance advances for reasoning tasks. In this paradigm, a large language model performs predictions based on a small set of supporting exemplars. The performance of In-Context Learning depends heavily on the selection procedure of the supporting exemplars, particularly in the case of HybridQA, where considering the diversity of reasoning chains and the large size of the hybrid contexts becomes crucial. In this work, we present Selection of ExEmplars for hybrid Reasoning (SEER), a novel method for selecting a set of exemplars that is both representative and diverse. The key novelty of SEER is that it formulates exemplar selection as a Knapsack Integer Linear Program. The Knapsack framework provides the flexibility to incorporate diversity constraints that prioritize exemplars with desirable attributes, and capacity constraints that ensure that the prompt size respects the provided capacity budgets. The effectiveness of SEER is demonstrated on FinQA and TAT-QA, two real-world benchmarks for HybridQA, where it outperforms previous exemplar selection methods.

16:00-17:30 (East Foyer / Virtual)
**#102 Non-Autoregressive Math Word Problem Solver with Unified Tree Structure**
*Yi Bin, Mengqun Han, Wenhao Shi, Lei Wang, Yang Yang, See-Kiong Ng and Heng Tao Shen*
Existing MWP solvers employ sequence or binary tree to present the solution expression and decode it from given problem description. However, such structures fail to handle the variants that can be derived via mathematical manipulation, e.g., $(a_1+a_2)*a_3$ and $a_1*a_3+a_2*a_3$ can both be possible valid solutions for a same problem but formulated as different expression sequences or trees. The multiple solution variants depicting different possible solving procedures for the same input problem would raise two issues: 1) making it hard for the model to learn the mapping function between the input and output spaces effectively, and 2) wrongly indicating *wrong* when evaluating a valid expression variant. To address these issues, we introduce a unified tree structure to present a solution expression, where the elements are permutable and identical for all the expression variants. We propose a novel non-autoregressive solver, named *MWP-NAS*, to parse the problem and deduce the solution expression based on the unified tree. For evaluating the possible expression variants, we design a path-based metric to evaluate the partial accuracy of expressions of a unified tree. The results from extensive experiments conducted on Math23K and MAWPS demonstrate the effectiveness of our proposed MWP-NAS. The codes and checkpoints are available at: https://github.com/mengqunhan/MWP-NAS.

16:00-17:30 (East Foyer / Virtual)
**#103 TRAVEL: Tag-Aware Conversational FAQ Retrieval via Reinforcement Learning**
*Yue Chen, Dingnan Jin, Chen Huang, Jia Liu and Wenqiang Lei*
Efficiently retrieving FAQ questions that match users' intent is essential for online customer service. Existing methods aim to fully utilize the dynamic conversation context to enhance the semantic association between the user query and FAQ questions. However, the conversation context contains noise, e.g., users may click questions they don't like, leading to inaccurate semantics modeling. To tackle this, we introduce tags of FAQ questions, which can help us eliminate irrelevant information. We later integrate them into a reinforcement learning framework and minimize the negative impact of irrelevant information in the dynamic conversation context. We experimentally demonstrate our efficiency and effectiveness on conversational FAQ retrieval compared to other baselines.

16:00-17:30 (East Foyer / Virtual)
**#104 PreWoMe: Exploiting Presuppositions as Working Memory for Long Form Question Answering**
*Wookje Han, Jinsol Park and Kyungjae Lee*
Information-seeking questions in long-form question answering (LFQA) often prove misleading due to ambiguity or false presupposition in the question. While many existing approaches handle misleading questions, they are tailored to limited questions, which are insufficient in a real-world setting with unpredictable input characteristics. In this work, we propose PreWoMe, a unified approach capable of handling any type of information-seeking question. The key idea of PreWoMe involves extracting presuppositions in the question and exploiting them as working memory to generate feedback and action about the question. Our experiment shows that PreWoMe is effective not only in tackling misleading questions but also in handling normal ones, thereby demonstrating the effectiveness of leveraging presuppositions, feedback, and action for real-world QA settings.

16:00-17:30 (East Foyer / Virtual)
**#105 A Diffusion Weighted Graph Framework for New Intent Discovery**
*Wenkai Shi, Wenbin An, Feng Tian, Qinghua Zheng, QianYing Wang and Ping Chen*
New Intent Discovery (NID) aims to recognize both new and known intents from unlabeled data with the aid of limited labeled data containing only known intents. Without considering structure relationships between samples, previous methods generate noisy supervisory signals which cannot strike a balance between quantity and quality, hindering the formation of new intent clusters and effective transfer of the pre-training knowledge. To mitigate this limitation, we propose a novel *Diffusion Weighted Graph Framework* (DWGF) to capture both semantic similarities and structure relationships inherent in data, enabling more sufficient and reliable supervisory signals. Specifically, for each sample, we diffuse neighborhood relationships along semantic paths guided by the nearest neighbors for multiple hops to characterize its local structure discriminately. Then, we sample its positive keys and weigh them based on semantic similarities and local structures for contrastive learning. During inference, we further propose *Graph Smoothing Filter* (GSF) to explicitly utilize the structure relationships to filter high-frequency noise embodied in semantically ambiguous samples on the cluster boundary. Extensive experiments show that our method outperforms state-of-the-art models on all evaluation metrics across multiple benchmark datasets. Code and data will be made public.

16:00-17:30 (East Foyer / Virtual)
**#106 Paraphrase Types for Generation and Detection**
*Jan Philip Wahle, Bela Gipp and Terry Ruas*
Current approaches in paraphrase generation and detection heavily rely on a single general similarity score, ignoring the intricate linguistic

properties of language. This paper introduces two new tasks to address this shortcoming by considering paraphrase types - specific linguistic perturbations at particular text positions. We name these tasks Paraphrase Type Generation and Paraphrase Type Detection. Our results suggest that while current techniques perform well in a binary classification scenario, i.e., paraphrased or not, the inclusion of fine-grained paraphrase types poses a significant challenge. While most approaches are good at generating and detecting general semantic similar content, they fail to understand the intrinsic linguistic variables they manipulate. Models trained in generating and identifying paraphrase types also show improvements in tasks without them. In addition, scaling these models further improves their ability to understand paraphrase types. We believe paraphrase types can unlock a new paradigm for developing paraphrase models and solving tasks in the future.

16:00-17:30 (East Foyer / Virtual)
**#107 PRESTO: A Multilingual Dataset for Parsing Realistic Task-Oriented Dialogs**
*Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Chuan He, Rattima Nitisaroj, Anna Trukhina, Shachi Paul, Pararth Shah, Rushin Shah and Zhou Yu*
Research interest in task-oriented dialogs has increased as systems such as Google Assistant, Alexa and Siri have become ubiquitous in everyday life. However, the impact of academic research in this area has been limited by the lack of datasets that realistically capture the wide array of user pain points. To enable research on some of the more challenging aspects of parsing realistic conversations, we introduce PRESTO, a public dataset of over 550K contextual multilingual conversations between humans and virtual assistants. PRESTO contains a diverse array of challenges that occur in real-world NLU tasks such as disfluencies, code-switching, and revisions. It is the only large scale human generated conversational parsing dataset that provides structured context such as a user's contacts and lists for each example. Our mT5 model based baselines demonstrate that the conversational phenomenon present in PRESTO are challenging to model, which is further pronounced in a low-resource setup.

16:00-17:30 (East Foyer / Virtual)
**#108 A Simple Baseline for Knowledge-Based Visual Question Answering**
*Alexandros Xenos, Themos Stafylakis, Ioannis Patras and Georgios Tzimiropoulos*
This paper is on the problem of Knowledge-Based Visual Question Answering (KB-VQA). Recent works have emphasized the significance of incorporating both explicit (through external databases) and implicit (through LLMs) knowledge to answer questions requiring external knowledge effectively. A common limitation of such approaches is that they consist of relatively complicated pipelines and often heavily rely on accessing GPT-3 API. Our main contribution in this paper is to propose a much simpler and readily reproducible pipeline which, in a nutshell, is based on efficient in-context learning by prompting LLaMA (1 and 2) using question-informative captions as contextual information. Contrary to recent approaches, our method is training-free, does not require access to external databases or APIs, and yet achieves state-of-the-art accuracy on the OK-VQA and A-OK-VQA datasets. Finally, we perform several ablation studies to understand important aspects of our method. Our code is publicly available at https://github.com/alexandrosXe/ASimple-Baseline-For-Knowledge-Based-VQA

16:00-17:30 (East Foyer / Virtual)
**#109 Weakly Supervised Semantic Parsing with Execution-based Spurious Program Filtering**
*Kang-il Lee, Segwang Kim and Kyomin Jung*
The problem of spurious programs is a longstanding challenge when training a semantic parser from weak supervision. To eliminate such programs that have wrong semantics but correct denotation, existing methods focus on exploiting similarities between examples based on domain-specific knowledge. In this paper, we propose a domain-agnostic filtering mechanism based on program execution results. Specifically, for each program obtained through the search process, we first construct a representation that captures the program's semantics as execution results under various inputs. Then, we run a majority vote on these representations to identify and filter out programs with significantly different semantics from the other programs. In particular, our method is orthogonal to the program search process so that it can easily augment any of the existing weakly supervised semantic parsing frameworks. Empirical evaluations on the Natural Language Visual Reasoning and WikiTableQuestions demonstrate that applying our method to the existing semantic parsers induces significantly improved performances.

16:00-17:30 (East Foyer / Virtual)
**#110 Mitigating Temporal Misalignment by Discarding Outdated Facts**
*Michael J.Q. Zhang and Eunsol Choi*
While large language models are able to retain vast amounts of world knowledge seen during pretraining, such knowledge is prone to going out of date and is nontrivial to update. Furthermore, these models are often used under temporal misalignment, tasked with answering questions about the present, despite having only been trained on data collected in the past. To mitigate the effects of temporal misalignment, we propose fact duration prediction: the task of predicting how long a given fact will remain true. In our experiments, we demonstrate that identifying which facts are prone to rapid change can help models avoid reciting outdated information and determine which predictions require seeking out up-to-date knowledge sources. We also show how modeling fact duration improves calibration for knowledge-intensive tasks, such as open-retrieval question answering, under temporal misalignment, by discarding volatile facts.

16:00-17:30 (East Foyer / Virtual)
**#111 Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models**
*Daman Arora, Himanshu Gaurav Singh and Mausam*
The performance of large language models (LLMs) on existing reasoning benchmarks has significantly improved over the past years. In response, we present JEEBench, a considerably more challenging benchmark dataset for evaluating the problem solving abilities of LLMs. We curate 515 challenging pre-engineering mathematics, physics and chemistry problems from the highly competitive IIT JEE-Advanced exam. Long-horizon reasoning on top of deep in-domain knowledge is essential for solving problems in this benchmark. Our evaluation on various open-source and proprietary models reveals that the highest performance, even after using techniques like self-consistency, self-refinement and chain-of-thought prompting, is less than 40%. The typical failure modes of GPT-4, the best model, are errors in algebraic manipulation, difficulty in grounding abstract concepts into mathematical equations accurately and failure in retrieving relevant domain-specific concepts. We also observe that by mere prompting, GPT-4 is unable to assess risk introduced by negative marking for incorrect answers. For this, we develop a post-hoc confidence-thresholding method over self-consistency, which enables effective response selection. We hope that our challenging benchmark will guide future re-search in problem-solving using LLMs.

16:00-17:30 (East Foyer / Virtual)
**#112 Continual Dialogue State Tracking via Example-Guided Question Answering**
*Hyundong Justin Cho, Andrea Madotto, Zhaojiang Lin, Khyathi Chandu, Satwik Kottur, Jing Xu, Jonathan May and Chinnadhurai Sankar*
Dialogue systems are frequently updated to accommodate new services, but naively updating them by continually training with data for new services in diminishing performance on previously learnt services. Motivated by the insight that dialogue state tracking (DST), a crucial component of dialogue systems that estimates the user's goal as a conversation proceeds, is a simple natural language understanding task, we propose reformulating it as a bundle of granular example-guided question answering tasks to minimize the task shift between services and thus benefit continual learning. Our approach alleviates service-specific memorization and teaches a model to contextualize the given

question and example to extract the necessary information from the conversation. We find that a model with just 60M parameters can achieve a significant boost by learning to learn from in-context examples retrieved by a retriever trained to identify turns with similar dialogue state changes. Combining our method with dialogue-level memory replay, our approach attains state of the art performance on DST continual learning metrics without relying on any complex regularization or parameter expansion methods.

16:00-17:30 (East Foyer / Virtual)
**#113 Seeing through the mess: evolutionary dynamics of lexical polysemy**
*Andreas Baumann, Andreas Stephan and Benjamin Roth*
Evidently, words can have multiple senses. For example, the word mess refers to a place to have food or to a confusing situation. How exactly multiple senses emerge is less clear. In this work, we propose and analyze a mathematical model of the evolution of lexical meaning to investigate mechanisms leading to polysemy. This model features factors that have been discussed to impact the semantic processing and transmission of words: word frequency, non-conformism, and semantic discriminability. We formally derive conditions under which a sense of a word tends to diversify itself into multiple senses that coexist stably. The model predicts that diversification is promoted by low frequency, a strong bias for non-conformist usage, and high semantic discriminability. We statistically validate these predictions with historical language data covering semantic developments of a set of English words. Multiple alternative measures are used to operationalize each variable involved, and we confirm the predicted tendencies for twelve combinations of measures.

16:00-17:30 (East Foyer / Virtual)
**#114 C-STS: Conditional Semantic Textual Similarity**
*Ameet Deshpande, Carlos E Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen and Karthik R Narasimhan*
Semantic textual similarity (STS) has been a cornerstone task in NLP that measures the degree of similarity between a pair of sentences, with applications in information retrieval, question answering, and embedding methods. However, it is an inherently ambiguous task, with the sentence similarity depending on the specific aspect of interest. We resolve this ambiguity by proposing a novel task called conditional STS (C-STS) which measures similarity conditioned on an aspect elucidated in natural language (hereon, condition). As an example, the similarity between the sentences "The NBA player shoots a three-pointer." and "A man throws a tennis ball into the air to serve." is higher for the condition "The motion of the ball." (both upward) and lower for "The size of the ball." (one large and one small). C-STS's advantages are two-fold: (1) it reduces the subjectivity and ambiguity of STS, and (2) enables fine-grained similarity evaluation using diverse conditions. C-STS contains almost 20,000 instances from diverse domains and we evaluate several state-of-the-art models to demonstrate that even the most performant fine-tuning and in-context learning models (GPT-4, Flan, SimCSE) find it challenging, with Spearman correlation scores of <50. We encourage the community to evaluate their models on C-STS to provide a more holistic view of semantic similarity and natural language understanding.

16:00-17:30 (East Foyer / Virtual)
**#115 Selectively Answering Ambiguous Questions**
*Jeremy R. Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra and Jacob Eisenstein*
Trustworthy language models should abstain from answering questions when they do not know the answer. However, the answer to a question can be unknown for a variety of reasons. Prior research has focused on the case in which the question is clear and the answer is unambiguous but possibly unknown. However, the answer to a question can also be unclear due to uncertainty of the questioner's intent or context. We investigate question answering from this perspective, focusing on answering a subset of questions with a high degree of accuracy, from a set of questions in which many are inherently ambiguous. In this setting, we find that the most reliable approach to calibration involves quantifying repetition within a set of sampled model outputs, rather than the model's likelihood or self-verification as used in prior work. We find this to be the case across different types of uncertainty, varying model scales and both with or without instruction tuning. Our results suggest that sampling-based confidence scores help calibrate answers to relatively unambiguous questions, with more dramatic improvements on ambiguous questions.

16:00-17:30 (East Foyer / Virtual)
**#116 Active Retrieval Augmented Generation**
*Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan and Graham Neubig*
Despite the remarkable ability of large language models (LMs) to comprehend and generate language, they have a tendency to hallucinate and create factually inaccurate output. Augmenting LMs by retrieving information from external knowledge resources is one promising solution. Most existing retrieval augmented LMs employ a retrieve-and-generate setup that only retrieves information once based on the input. This is limiting, however, in more general scenarios involving generation of long texts, where continually gathering information throughout generation is essential. In this work, we provide a generalized view of active retrieval augmented generation, methods that actively decide when and what to retrieve across the course of the generation. We propose Forward-Looking Active REtrieval augmented generation (FLARE), a generic method which iteratively uses a prediction of the upcoming sentence to anticipate future content, which is then utilized as a query to retrieve relevant documents to regenerate the sentence if it contains low-confidence tokens. We test FLARE along with baselines comprehensively over 4 long-form knowledge-intensive generation tasks/datasets. FLARE achieves superior or competitive performance on all tasks, demonstrating the effectiveness of our method.

16:00-17:30 (East Foyer / Virtual)
**#117 BERTie Bott's Every Flavor Labels: A Tasty Introduction to Semantic Role Labeling for Galician**
*Micaella Bruton and Meriem Beloucif*
In this paper, we leverage existing corpora, WordNet, and dependency parsing to build the first Galician dataset for training semantic role labeling systems in an effort to expand available NLP resources. Additionally, we introduce verb indexing, a new pre-processing method, which helps increase the performance when semantically parsing highly-complex sentences. We use transfer-learning to test both the resource and the verb indexing method. Our results show that the effects of verb indexing were amplified in scenarios where the model was both pre-trained and fine-tuned on datasets utilizing the method, but improvements are also noticeable when only used during fine-tuning. The best-performing Galician SRL model achieved an f1 score of 0.74, introducing a baseline for future Galician SRL systems. We also tested our method on Spanish where we achieved an f1 score of 0.83, outperforming the baseline set by the 2009 CoNLL Shared Task by 0.025 showing the merits of our verb indexing method for pre-processing.

16:00-17:30 (East Foyer / Virtual)
**#118 Support or Refute: Analyzing the Stance of Evidence to Detect Out-of-Context Mis- and Disinformation**
*Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang and Shujun Li*
Mis- and disinformation online have become a major societal problem as major sources of online harms of different kinds. One common form of mis- and disinformation is out-of-context (OOC) information, where different pieces of information are falsely associated, e.g., a real image combined with a false textual caption or a misleading textual description. Although some past studies have attempted to defend against OOC mis- and disinformation through external evidence, they tend to disregard the role of different pieces of evidence with different stances.

Motivated by the intuition that the stance of evidence represents a bias towards different detection results, we propose a stance extraction network (SEN) that can extract the stances of different pieces of multi-modal evidence in a unified framework. Moreover, we introduce a support-refutation score calculated based on the co-occurrence relations of named entities into the textual SEN. Extensive experiments on a public large-scale dataset demonstrated that our proposed method outperformed the state-of-the-art baselines, with the best model achieving a performance gain of 3.2% in accuracy.

16:00-17:30 (East Foyer / Virtual)
**#119 End-to-End Single-Channel Speaker-Turn Aware Conversational Speech Translation**
*Juan Pablo Zuluaga-Gomez, Zhaocheng Huang, Xing Niu, Rohit Paturi, Sundararajan Srinivasan, Prashant Mathur, Brian Thompson and Marcello Federico*
Conventional speech-to-text translation (ST) systems are trained on single-speaker utterances, and they may not generalize to real-life scenarios where the audio contains conversations by multiple speakers. In this paper, we tackle single-channel multi-speaker conversational ST with an end-to-end and multi-task training model, named Speaker-Turn Aware Conversational Speech Translation, that combines automatic speech recognition, speech translation and speaker turn detection using special tokens in a serialized labeling format. We run experiments on the Fisher-CALLHOME corpus, which we adapted by merging the two single-speaker channels into one multi-speaker channel, thus representing the more realistic and challenging scenario with multi-speaker turns and cross-talk. Experimental results across single- and multi-speaker conditions and against conventional ST systems, show that our model outperforms the reference systems on the multi-speaker condition, while attaining comparable performance on the single-speaker condition. We release scripts for data processing and model training.

16:00-17:30 (East Foyer / Virtual)
**#120 Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata**
*Silei Xu, Shicheng Liu, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina Semnani and Monica Lam*
While large language models (LLMs) can answer many questions correctly, they can also hallucinate and give wrong answers. Wikidata, with its over 12 billion facts, can be used to ground LLMs to improve their factuality. This paper presents WikiWebQuestions, a high-quality question answering benchmark for Wikidata. Ported over from WebQuestions for Freebase, it consists of real-world data with SPARQL annotation. This paper presents a few-shot sequence-to-sequence semantic parser for Wikidata. We modify SPARQL to use the unique domain and property names instead of their IDs. We train the parser to use either the results from an entity linker or mentions in the query. We fine-tune LLaMA by adding the few-shot training data to that used to fine-tune Alpaca. Our experimental results demonstrate the effectiveness of this methodology, establishing a strong baseline of 76% and 65% answer accuracy in the dev and test sets of WikiWebQuestions, respectively. By pairing our semantic parser with GPT-3, we combine verifiable results with qualified GPT-3 guesses to provide useful answers to 96% of the questions in dev. We also show that our method outperforms the state-of-the-art for the QALD-7 Wikidata dataset by 3.6% in F1 score.

16:00-17:30 (East Foyer / Virtual)
**#121 Bridging Continuous and Discrete Spaces: Interpretable Sentence Representation Learning via Compositional Operations**
*James Y. Huang, Wenlin Yao, Kaiqiang Song, Hongming Zhang, Muhao Chen and Dong Yu*
Traditional sentence embedding models encode sentences into vector representations to capture useful properties such as the semantic similarity between sentences. However, in addition to similarity, sentence semantics can also be interpreted via compositional operations such as sentence fusion or difference. It is unclear whether the compositional semantics of sentences can be directly reflected as compositional operations in the embedding space. To more effectively bridge the continuous embedding and discrete text spaces, we explore the plausibility of incorporating various compositional properties into the sentence embedding space that allows us to interpret embedding transformations as compositional sentence operations. We propose InterSent, an end-to-end framework for learning interpretable sentence embeddings that supports compositional sentence operations in the embedding space. Our method optimizes operator networks and a bottleneck encoder-decoder model to produce meaningful and interpretable sentence embeddings. Experimental results demonstrate that our method significantly improves the interpretability of sentence embeddings on four textual generation tasks over existing approaches while maintaining strong performance on traditional semantic similarity tasks.

16:00-17:30 (East Foyer / Virtual)
**#122 Exploring Chain of Thought Style Prompting for Text-to-SQL**
*Chang-Yu Tai, Ziru Chen, Tianshu Zhang, Xiang Deng and Huan Sun*
In-context learning with large language models (LLMs) has recently caught increasing attention due to its superior few-shot performance on various tasks. However, its performance on text-to-SQL parsing still has much room for improvement. In this paper, we hypothesize that a crucial aspect of LLMs to improve for text-to-SQL parsing is their multi-step reasoning ability. Thus, we systematically study how to enhance LLMs' reasoning ability through chain of thought (CoT) style prompting, including the original chain-of-thought prompting and least-to-most prompting. Our experiments demonstrate that iterative prompting as in least-to-most prompting may be unnecessary for text-to-SQL parsing, and using detailed reasoning steps tends to have more error propagation issues. Based on these findings, we propose a new CoT-style prompting method for text-to-SQL parsing. It brings 5.2 and 6.5 point absolute gains on the Spider development set and the Spider Realistic set, respectively, compared to the standard prompting method without reasoning steps; 2.4 and 1.5 point absolute gains, compared to the least-to-most prompting method.

16:00-17:30 (East Foyer / Virtual)
**#123 Towards a Unified Conversational Recommendation System: Multi-task Learning via Contextualized Knowledge Distillation**
*Yeongseo Jung, Eunseo Jung and Lei Chen*
In Conversational Recommendation System (CRS), an agent is asked to recommend a set of items to users within natural language conversations. To address the need for both conversational capability and personalized recommendations, prior works have utilized separate recommendation and dialogue modules. However, such approach inevitably results in a discrepancy between recommendation results and generated responses. To bridge the gap, we propose a multi-task learning for a unified CRS, where a single model jointly learns both tasks via Contextualized Knowledge Distillation (ConKD). We introduce two versions of ConKD: hard gate and soft gate. The former selectively gates between two task-specific teachers, while the latter integrates knowledge from both teachers. Our gates are computed on-the-fly in a context-specific manner, facilitating flexible integration of relevant knowledge. Extensive experiments demonstrate that our single model significantly improves recommendation performance while enhancing fluency, and achieves comparable results in terms of diversity.

16:00-17:30 (East Foyer / Virtual)
**#124 DeSIQ: Towards an Unbiased, Challenging Benchmark for Social Intelligence Understanding**
*Xiao-Yu Guo, Yuan-Fang Li and Reza Haf*
Social intelligence is essential for understanding and reasoning about human expressions, intents and interactions. One representative benchmark for its study is Social Intelligence Queries (Social-IQ), a dataset of multiple-choice questions on videos of complex social interactions. We define a comprehensive methodology to study the soundness of Social-IQ, as the soundness of such benchmark datasets is crucial to the investigation of the underlying research problem. We define a comprehensive methodology to study the soundness of Social-IQ, as the soundness of such benchmark datasets is crucial to the investigation of the underlying research problem. Our analysis reveals that Social-IQ

contains substantial biases, which can be exploited by a moderately strong language model to learn spurious correlations to achieve perfect performance without being given the context or even the question. We introduce DeSIQ, a new challenging dataset, constructed by applying simple perturbations to Social-IQ. Our empirical analysis shows De-SIQ significantly reduces the biases in the original Social-IQ dataset. Furthermore, we examine and shed light on the effect of model size, model style, learning settings, commonsense knowledge, and multi-modality on the new benchmark performance. Our new dataset, observations and findings open up important research questions for the study of social intelligence.

16:00-17:30 (East Foyer / Virtual)
**#125 We're Afraid Language Models Aren't Modeling Ambiguity**
*Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith and Yejin Choi*
Ambiguity is an intrinsic feature of natural language. Managing ambiguity is a key part of human language understanding, allowing us to anticipate misunderstanding as communicators and revise our interpretations as listeners. As language models are increasingly employed as dialogue interfaces and writing aids, handling ambiguous language is critical to their success. We capture ambiguity in a sentence through its effect on entailment relations with another sentence, and collect AmbiEnt, a linguist-annotated benchmark of 1,645 examples with diverse kinds of ambiguity. We design a suite of tests based on AmbiEnt, presenting the first evaluation of pretrained LMs to recognize ambiguity and disentangle possible meanings. We find that the task remains extremely challenging, including for GPT-4, whose generated disambiguations are considered correct only 32% of the time in crowdworker evaluation, compared to 90% for disambiguations in our dataset. Finally, to illustrate the value of ambiguity-sensitive tools, we show that a multilabel NLI model can flag political claims in the wild that are misleading due to ambiguity. We encourage the field to rediscover the importance of ambiguity for NLP.

16:00-17:30 (East Foyer / Virtual)
**#126 Length is a Curse and a Blessing for Document-level Semantics**
*Chenghao Xiao, Yizhi Li, G Thomas Hudson, Chenghua Lin and Noura Al Moubayed*
In recent years, contrastive learning (CL) has been extensively utilized to recover sentence and document-level encoding capability from pre-trained language models. In this work, we question the length generalizability of CL-based models, i.e., their vulnerability towards length-induced semantic shift. We verify not only that length vulnerability is a significant yet overlooked research gap, but we can devise unsupervised CL methods solely depending on the semantic signal provided by document length. We first derive the theoretical foundations underlying length attacks, showing that elongating a document would intensify the high intra-document similarity that is already brought by CL. Moreover, we found that isotropy promised by CL is highly dependent on the length range of text exposed in training. Inspired by these findings, we introduce a simple yet universal document representation learning framework, **LA(SER)$^3$**: length-agnostic self-reference for semantically robust sentence representation learning, achieving state-of-the-art unsupervised performance on the standard information retrieval benchmark. [Our code is publicly available.](https://github.com/gowitheflow-1998/LA-SER-cubed)

16:00-17:30 (East Foyer / Virtual)
**#127 Well Begun is Half Done: Generator-agnostic Knowledge Pre-Selection for Knowledge-Grounded Dialogue**
*Lang Qin, Yao Zhang, Hongru Liang, Jun Wang and Zhenglu Yang*
Accurate knowledge selection is critical in knowledge-grounded dialogue systems. Towards a closer look at it, we offer a novel perspective to organize existing literature, i.e., knowledge selection coupled with, after, and before generation. We focus on the third under-explored category of study, which can not only select knowledge accurately in advance, but has the advantage to reduce the learning, adjustment, and interpretation burden of subsequent response generation models, especially LLMs. We propose `GATE`, a generator-agnostic knowledge selection method, to prepare knowledge for subsequent response generation models by selecting context-related knowledge among different knowledge structures and variable knowledge requirements. Experimental results demonstrate the superiority of `GATE`, and indicate that knowledge selection before generation is a lightweight yet effective way to facilitate LLMs (e.g., ChatGPT) to generate more informative responses.

16:00-17:30 (East Foyer / Virtual)
**#128 AMR Parsing with Causal Hierarchical Attention and Pointers**
*Chao Lou and Kewei Tu*
Translation-based AMR parsers have recently gained popularity due to their simplicity and effectiveness. They predict linearized graphs as free texts, avoiding explicit structure modeling. However, this simplicity neglects structural locality in AMR graphs and introduces unnecessary tokens to represent coreferences. In this paper, we introduce new target forms of AMR parsing and a novel model, CHAP, which is equipped with causal hierarchical attention and the pointer mechanism, enabling the integration of structures into the Transformer decoder. We empirically explore various alternative modeling options. Experiments show that our model outperforms baseline models on four out of five benchmarks in the setting of no additional data.

16:00-17:30 (East Foyer / Virtual)
**#129 Chain-of-Questions Training with Latent Answers for Robust Multistep Question Answering**
*Wang Zhu, Jesse Thomason and Robin Jia*
We propose Chain-of-Questions, a framework that trains a model to robustly answer multistep questions by generating and answering sub-questions. We obtain supervision for sub-questions from human-annotated question decomposition meaning representation (QDMR), but QDMR does not include annotated answers to sub-questions. To overcome this technical challenge, we treat sub-answers as latent variables and infer them with a novel dynamic mixture of Hard-EM and MAPO. Chain-of-Questions is effective and robust, greatly outperforming strong neuro-symbolic methods by 9.0 F1 on a DROP contrast set and GPT-3.5 by 24.3 F1 on a HotpotQA adversarial set.

16:00-17:30 (East Foyer / Virtual)
**#130 Scalable-DSC: A Structural Template Prompt Approach to Scalable Dialogue State Correction**
*Haoxiang Su, Hongyan Xie, Hao Huang, Shuangyong Song, Ruiyu Fang, Xiaomeng Huang and Sijie Feng*
Dialogue state error correction has recently been proposed to correct wrong slot values in predicted dialogue states, thereby mitigating the error propagation problem for dialogue state tracking (DST). These approaches, though effective, are heavily intertwined with specific DST models, limiting their applicability to other DST models. To solve this problem, we propose Scalable Dialogue State Correction (Scalable-DSC), which can correct wrong slot values in the dialogue state predicted by any DST model. Specifically, we propose a Structural Template Prompt (STP) that converts predicted dialogue state from any DST models into a standardized natural language sequence as a part of the historical context, associates them with dialogue history information, and generates a corrected dialogue state sequence based on predefined template options. We further enhance Scalable-DSC by introducing two training strategies. The first employs a predictive state simulator to simulate the predicted dialogue states as the training data to enhance the generalization ability of the model. The second involves using the dialogue state predicted by DST as the training data, aiming at mitigating the inconsistent error type distribution between the training and inference. Experiments confirm that our model achieves state-of-the-art results on MultiWOZ 2.0-2.4.

16:00-17:30 (East Foyer / Virtual)
**#131 SuperDialseg: A Large-scale Dataset for Supervised Dialogue Segmentation**
*Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi and Akiko Aizawa*
Dialogue segmentation is a crucial task for dialogue systems allowing a better understanding of conversational texts. Despite recent progress in unsupervised dialogue segmentation methods, their performances are limited by the lack of explicit supervised signals for training. Furthermore, the precise definition of segmentation points in conversations still remains as a challenging problem, increasing the difficulty of collecting manual annotations. In this paper, we provide a feasible definition of dialogue segmentation points with the help of document-grounded dialogues and release a large-scale supervised dataset called SuperDialseg, containing 9,478 dialogues based on two prevalent document-grounded dialogue corpora, and also inherit their useful dialogue-related annotations. Moreover, we provide a benchmark including 18 models across five categories for the dialogue segmentation task with several proper evaluation metrics. Empirical studies show that supervised learning is extremely effective in in-domain datasets and models trained on SuperDialseg can achieve good generalization ability on out-of-domain data. Additionally, we also conducted human verification on the test set and the Kappa score confirmed the quality of our automatically constructed dataset. We believe our work is an important step forward in the field of dialogue segmentation.

16:00-17:30 (East Foyer / Virtual)
**#132 Building Multi-domain Dialog State Trackers from Single-domain Dialogs**
*Qi Zhu, Zheng Zhang, Xiaoyan Zhu and Minlie Huang*
Existing multi-domain dialog state tracking (DST) models are developed based on multi-domain dialogs, which require significant manual effort to define domain relations and collect data. This process can be challenging and expensive, particularly when numerous domains are involved. In this paper, we propose a divide-and-conquer (DAC) DST paradigm and a multi-domain dialog synthesis framework, which makes building multi-domain DST models from single-domain dialogs possible. The DAC paradigm segments a multi-domain dialog into multiple single-domain dialogs for DST, which makes models generalize better on dialogs involving unseen domain combinations. The multi-domain dialog synthesis framework merges several potentially related single-domain dialogs into one multi-domain dialog and modifies the dialog to simulate domain relations. The synthesized dialogs can help DST models capture the value transfer between domains. Experiments with three representative DST models on two datasets demonstrate the effectiveness of our proposed DAC paradigm and data synthesis framework.

16:00-17:30 (East Foyer / Virtual)
**#133 Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model**
*Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer and David R Mortensen*
Large language models (LLMs) have recently reached an impressive level of linguistic capability, prompting comparisons with human language skills. However, there have been relatively few systematic inquiries into the linguistic capabilities of the latest generation of LLMs, and those studies that do exist (i) ignore the remarkable ability of humans to generalize, (ii) focus only on English, and (iii) investigate syntax or semantics and overlook other capabilities that lie at the heart of human language, like morphology. Here, we close these gaps by conducting the first rigorous analysis of the morphological capabilities of ChatGPT in four typologically varied languages (specifically, English, German, Tamil, and Turkish). We apply a version of Berko's (1958) wug test to ChatGPT, using novel, uncontaminated datasets for the four examined languages. We find that ChatGPT massively underperforms purpose-built systems, particularly in English. Overall, our results—through the lens of morphology—cast a new light on the linguistic capabilities of ChatGPT, suggesting that claims of human-like language skills are premature and misleading.

16:00-17:30 (East Foyer / Virtual)
**#134 Speech Recognition and Meaning Interpretation: Towards Disambiguation of Structurally Ambiguous Spoken Utterances in Indonesian**
*Ruhiyah Faradishi Widiaputri, Ayu Purwarianti, Dessi Puji Lestari, Kurniawati Azizah, Dipta Tanaya and Sakriani Sakti*
Despite being the world's fourth-most populous country, the development of spoken language technologies in Indonesia still needs improvement. Most automatic speech recognition (ASR) systems that have been developed are still limited to transcribing the exact word-by-word, which, in many cases, consists of ambiguous sentences. In fact, speakers use prosodic characteristics of speech to convey different interpretations, which, unfortunately, these systems often ignore. In this study, we attempt to resolve structurally ambiguous utterances into unambiguous texts in Indonesian using prosodic information. To the best of our knowledge, this might be the first study to address this problem in the ASR context. Our contributions include (1) collecting the Indonesian speech corpus on structurally ambiguous sentences; (2) conducting a survey on how people disambiguate structurally ambiguous sentences presented in both text and speech forms; and (3) constructing an Indonesian ASR and meaning interpretation system by utilizing both cascade and direct approaches to map speech to text, along with two additional prosodic information signals (pause and pitch). The experimental results reveal that it is possible to disambiguate these utterances. In this study, the proposed cascade system, utilizing Mel-spectrograms concatenated with F0 and energy as input, achieved a disambiguation accuracy of 79.6%, while the proposed direct system with the same input yielded an even more impressive disambiguation accuracy of 82.2%.

16:00-17:30 (East Foyer / Virtual)
**#135 Dialogue Chain-of-Thought Distillation for Commonsense-aware Conversational Agents**
*Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang and Jinyoung Yeo*
Human-like chatbots necessitate the use of commonsense reasoning in order to effectively comprehend and respond to implicit information present within conversations. Achieving such coherence and informativeness in responses, however, is a non-trivial task. Even for large language models (LLMs), the task of identifying and aggregating key evidence within a single hop presents a substantial challenge. This complexity arises because such evidence is scattered across multiple turns in a conversation, thus necessitating integration over multiple hops. Hence, our focus is to facilitate such multi-hop reasoning over a dialogue context, namely dialogue chain-of-thought (CoT) reasoning. To this end, we propose a knowledge distillation framework that leverages LLMs as unreliable teachers and selectively distills consistent and helpful rationales via alignment filters. We further present DOCTOR, a DialOgue Chain-of-ThOught Reasoner that provides reliable CoT rationales for response generation. We conduct extensive experiments to show that enhancing dialogue agents with high-quality rationales from DOCTOR significantly improves the quality of their responses.

16:00-17:30 (East Foyer / Virtual)
**#136 PromptST: Abstract Prompt Learning for End-to-End Speech Translation**
*Tengfei Yu, Liang Ding, Xuebo Liu, Kehai Chen, Meishan Zhang, Dacheng Tao and Min Zhang*
An end-to-end speech-to-text (S2T) translation model is usually initialized from a pre-trained speech recognition encoder and a pre-trained text-to-text (T2T) translation decoder. Although this straightforward setting has been shown empirically successful, there do not exist clear answers to the research questions: 1) how are speech and text modalities fused in S2T model and 2) how to better fuse the two modalities? In this paper, we take the first step toward understanding the fusion of speech and text features in S2T model. We first design and release a 10GB linguistic probing benchmark, namely Speech-Senteval, to investigate the acoustic and linguistic behaviors of S2T models. Preliminary analysis reveals that the uppermost encoder layers of the S2T model can not learn linguistic knowledge efficiently, which is crucial for accurate

translation. Based on the finding, we further propose a simple plug-in prompt-learning strategy on the uppermost encoder layers to broaden the abstract representation power of the encoder of S2T models. We call such a prompt-enhanced S2T model PromptST. Experimental results on four widely-used S2T datasets show that PromptST can deliver significant improvements over a strong baseline by capturing richer linguistic knowledge. Benchmarks, code, and scripts are freely available at https://github.com/ytf-philp/PromptST.

16:00-17:30 (East Foyer / Virtual)
**#137 Pointwise Mutual Information Based Metric and Decoding Strategy for Faithful Generation in Document Grounded Dialogs**
*Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi and Luis A. Lastras*
A major concern in using deep learning based generative models for document-grounded dialogs is the potential generation of responses that are not faithful to the underlying document. Existing automated metrics used for evaluating the faithfulness of response with respect to the grounding document measure the degree of similarity between the generated response and the document's content. However, these automated metrics are far from being well aligned with human judgments. Therefore, to improve the measurement of faithfulness, we propose a new metric that utilizes (Conditional) Point-wise Mutual Information (PMI) between the generated response and the source document, conditioned on the dialogue. PMI quantifies the extent to which the document influences the generated response – with a higher PMI indicating a more faithful response. We build upon this idea to create a new decoding technique that incorporates PMI into the response generation process to predict more faithful responses. Our experiments on the BEGIN benchmark demonstrate an improved correlation of our metric with human evaluation. We also show that our decoding technique is effective in generating more faithful responses when compared to standard decoding techniques on a set of publicly available document-grounded dialog datasets.

16:00-17:30 (East Foyer / Virtual)
**#138 Enhancing Task-oriented Dialogue Systems with Generative Post-processing Networks**
*Atsumoto Ohashi and Ryuichiro Higashinaka*
Recently, post-processing networks (PPNs), which modify the outputs of arbitrary modules including non-differentiable ones in task-oriented dialogue systems, have been proposed. PPNs have successfully improved the dialogue performance by post-processing natural language understanding (NLU), dialogue state tracking (DST), and dialogue policy (Policy) modules with a classification-based approach. However, they cannot be applied to natural language generation (NLG) modules because the post-processing of the utterance output by the NLG module requires a generative approach. In this study, we propose a new post-processing component for NLG, generative post-processing networks (GenPPNs). For optimizing GenPPNs via reinforcement learning, the reward function incorporates dialogue act contribution, a new measure to evaluate the contribution of GenPPN-generated utterances with regard to task completion in dialogue. Through simulation and human evaluation experiments based on the MultiWOZ dataset, we confirmed that GenPPNs improve the task completion performance of task-oriented dialogue systems.

16:00-17:30 (East Foyer / Virtual)
**#139 P5: Plug-and-Play Persona Prompting for Personalized Response Selection**
*Joosung Lee, Minsik Oh and Donghun Lee*
The use of persona-grounded retrieval-based chatbots is crucial for personalized conversations, but there are several challenges that need to be addressed. 1) In general, collecting persona-grounded corpus is very expensive. 2) The chatbot system does not always respond in consideration of persona at real applications. To address these challenges, we propose a plug-and-play persona prompting method. Our system can function as a standard open-domain chatbot if persona information is not available. We demonstrate that this approach performs well in the zero-shot setting, which reduces the dependence on persona-ground training data. This makes it easier to expand the system to other languages without the need to build a persona-grounded corpus. Additionally, our model can be fine-tuned for even better performance. In our experiments, the zero-shot model improved the standard model by 7.71 and 1.04 points in the original persona and revised persona, respectively. The fine-tuned model improved the previous state-of-the-art system by 1.95 and 3.39 points in the original persona and revised persona, respectively. To the best of our knowledge, this is the first attempt to solve the problem of personalized response selection using prompt sequences. Our code is available on github.

16:00-17:30 (East Foyer / Virtual)
**#140 AnyTOD: A Programmable Task-Oriented Dialog System**
*Jeffrey Zhao, Yuan Cao, Raghav Gupta, Harrison Lee, Abhinav Rastogi, Mingqiu Wang, Hagen Soltau, Izhak Shafran and Yonghui Wu*
We propose AnyTOD, an end-to-end, zero-shot task-oriented dialog (TOD) system capable of zero-shot adaptation onto unseen tasks or domains. We view TOD as a program executed by a language model (LM), where program logic and ontology is provided by a designer as a schema. To enable generalization to unseen schemas and programs without prior training, AnyTOD adopts a neuro-symbolic approach. A neural LM keeps track of events that occur during a conversation, and a symbolic program implementing dialog policy is executed to recommend actions AnyTOD should take. This approach drastically reduces data annotation and model training requirements, addressing the enduring challenge of rapidly adapting a TOD system to unseen tasks and domains. We demonstrate state-of-the-art results on STAR, ABCD and SGD benchmarks. We also demonstrate strong zero-shot transfer ability in low-resource settings, such as zero-shot transfer onto MultiWOZ. In addition, we release STARv2, an updated version of the STAR dataset with richer annotations, for benchmarking zero-shot task transfer for end-to-end TOD models.

16:00-17:30 (East Foyer / Virtual)
**#141 Learning From Free-Text Human Feedback – Collect New Datasets Or Extend Existing Ones?**
*Dominic Petrak, Nafise Sadat Moosavi, Ye Tian, Nikolai Rozanov and Iryna Gurevych*
Continuous learning from free-text human feedback, such as error corrections, new knowledge, or alternative responses, is essential for today's chatbots and virtual assistants to stay up-to-date, engaging, and socially acceptable. However, for research on methods for learning from such data, annotated data is scarce. To address this, we examine the error and user response types of six popular dialogue datasets from various types, including MultiWoZ, PersonaChat, Wizards-of-Wikipedia, and others, to assess their extendibility with the needed annotations. For this corpus study, we manually annotate a subset of each dataset with error and user response types using an improved version of the Integrated Error Taxonomy and a newly proposed user response type taxonomy. We provide the resulting dataset (EURTAD) to the community. Our findings provide new insights into dataset composition, including error types, user response types, and the relations between them.

16:00-17:30 (East Foyer / Virtual)
**#142 ViT-TTS: Visual Text-to-Speech with Scalable Diffusion Transformer**
*Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He and Zhou Zhao*
Text-to-speech(TTS) has undergone remarkable improvements in performance, particularly with the advent of Denoising Diffusion Probabilistic Models (DDPMs). However, the perceived quality of audio depends not solely on its content, pitch, rhythm, and energy, but also on the physical environment.In this work, we propose ViT-TTS, the first visual TTS model with scalable diffusion transformers. ViT-TTS complement the phoneme sequence with the visual information to generate high-perceived audio, opening up new avenues for practical applications of AR and VR to allow a more immersive and realistic audio experience. To mitigate the data scarcity in learning visual acoustic information, we 1) introduce a self-supervised learning framework to enhance both the visual-text encoder and denoiser decoder; 2) leverage

the diffusion transformer scalable in terms of parameters and capacity to learn visual scene information. Experimental results demonstrate that ViT-TTS achieves new state-of-the-art results, outperforming cascaded systems and other baselines regardless of the visibility of the scene. With low-resource data (1h, 2h, 5h), ViT-TTS achieves comparative results with rich-resource baselines.

16:00-17:30 (East Foyer / Virtual)
**#143 To Split or Not to Split: Composing Compounds in Contextual Vector Spaces**
*Christopher William Jenkins, Filip Miletic and Sabine Schulte im Walde*
We investigate the effect of sub-word tokenization on representations of German noun compounds: single orthographic words which are composed of two or more constituents but often tokenized into units that are not morphologically motivated or meaningful. Using variants of BERT models and tokenization strategies on domain-specific restricted diachronic data, we introduce a suite of evaluations relying on the masked language modelling task and compositionality prediction. We obtain the most consistent improvements by pre-splitting compounds into constituents.

16:00-17:30 (East Foyer / Virtual)
**#144 Hop, Union, Generate: Explainable Multi-hop Reasoning without Rationale Supervision**
*Wenting Zhao, Justin T Chiu, Claire Cardie and Alexander M Rush*
Explainable multi-hop question answering (QA) not only predicts answers but also identifies rationales, i. e. subsets of input sentences used to derive the answers. Existing methods rely on supervision for both answers and rationales. This problem has been extensively studied under the supervised setting, where both answer and rationale annotations are given. Because rationale annotations are expensive to collect and not always available, recent efforts have been devoted to developing methods that do not rely on supervision for rationales. However, such methods have limited capacities in modeling interactions between sentences, let alone reasoning across multiple documents. This work proposes a principled, probabilistic approach for training explainable multi-hop QA systems without rationale supervision. Our approach performs multi-hop reasoning by explicitly modeling rationales as sets, enabling the model to capture interactions between documents and sentences within a document. Experimental results show that our approach is more accurate at selecting rationales than the previous methods, while maintaining similar accuracy in predicting answers.

16:00-17:30 (East Foyer / Virtual)
**#145 Can language models learn analogical reasoning? Investigating training objectives and comparisons to human performance**
*Molly Petersen and Lonneke van der Plas*
While analogies are a common way to evaluate word embeddings in NLP, it is also of interest to investigate whether or not analogical reasoning is a task in itself that can be learned. In this paper, we test several ways to learn basic analogical reasoning, specifically focusing on analogies that are more typical of what is used to evaluate analogical reasoning in humans than those in commonly used NLP benchmarks. Our experiments find that models are able to learn analogical reasoning, even with a small amount of data. We additionally compare our models to a dataset with a human baseline, and find that after training models approach human performance.

16:00-17:30 (East Foyer / Virtual)
**#146 Interactive Text-to-SQL Generation via Editable Step-by-Step Explanations**
*Yuan Tian, Zheng Zhang, Zheng Ning, Toby Jia-Jun Li, Jonathan K. Kummerfeld and Tianyi Zhang*
Relational databases play an important role in business, science, and more. However, many users cannot fully unleash the analytical power of relational databases, because they are not familiar with database languages such as SQL. Many techniques have been proposed to automatically generate SQL from natural language, but they suffer from two issues: (1) they still make many mistakes, particularly for complex queries, and (2) they do not provide a flexible way for non-expert users to validate and refine incorrect queries. To address these issues, we introduce a new interaction mechanism that allows users to directly edit a step-by-step explanation of a query to fix errors. Our experiments on multiple datasets, as well as a user study with 24 participants, demonstrate that our approach can achieve better performance than multiple SOTA approaches. Our code and datasets are available at https://github.com/magic-YuanTian/STEPS.

16:00-17:30 (East Foyer / Virtual)
**#147 ToolWriter: Question Specific Tool Synthesis for Tabular Data**
*Carlos Gemmell and Jeff Dalton*
Tabular question answering (TQA) presents a challenging setting for neural systems by requiring joint reasoning of natural language with large amounts of semi-structured data. Unlike humans who use programmatic tools like filters to transform data before processing, language models in TQA process tables directly, resulting in information loss as table size increases. In this paper we propose ToolWriter to generate query specific programs and detect when to apply them to transform tables and align them with the TQA model's capabilities. Focusing Toolwriter to generate row-filtering tools improves the state-of-the-art for WikiTableQuestions and WikiSQL with the most performance gained on long tables. By investigating headroom, our work highlights the broader potential for programmatic tools combined with neural components to manipulate large amounts of structured data.

16:00-17:30 (East Foyer / Virtual)
**#148 TaskDiff: A Similarity Metric for Task-Oriented Conversations**
*Ankita Bhaumik, Praveen Venkateswaran, Yara Rizk and Vatche Isahagian*
The popularity of conversational digital assistants has resulted in the availability of large amounts of conversational data which can be utilized for improved user experience and personalized response generation. Building these assistants using popular large language models like ChatGPT also require additional emphasis on prompt engineering and evaluation methods. Textual similarity metrics are a key ingredient for such analysis and evaluations. While many similarity metrics have been proposed in the literature, they have not proven effective for task-oriented conversations as they do not take advantage of unique conversational features. To address this gap, we present TaskDiff, a novel conversational similarity metric that utilizes different dialogue components (utterances, intents, and slots) and their distributions to compute similarity. Extensive experimental evaluation of TaskDiff on a benchmark dataset demonstrates its superior performance and improved robustness over other related approaches.

16:00-17:30 (East Foyer / Virtual)
**#149 Zero-Shot Multi-Label Topic Inference with Sentence Encoders and LLMs**
*Souvika Sarkar, Dongji Feng and Shubhra Kanti Karmaker Santu*
In this paper, we conducted a comprehensive study with the latest Sentence Encoders and Large Language Models (LLMs) on the challenging task of "definition-wild zero-shot topic inference", where users define or provide the topics of interest in real-time. Through extensive experimentation on seven diverse data sets, we observed that LLMs, such as ChatGPT-3.5 and PaLM, demonstrated superior generality compared to other LLMs, e.g., BLOOM and GPT-NeoX. Furthermore, Sentence-BERT, a BERT-based classical sentence encoder, outperformed PaLM and achieved performance comparable to ChatGPT-3.5.

16:00-17:30 (East Foyer / Virtual)

**#150 DP-Parse: Finding Word Boundaries from Raw Speech with an Instance Lexicon**
*Robin Algayres, Tristan Ricoul, Julien Karadayi, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, Emmanuel Dupoux and Hugo Laurençon*
Finding word boundaries in continuous speech is challenging as there is little or no equivalent of a space delimiter between words.Popular Bayesian non-parametric models for text segmentation use a Dirichlet process to jointly segment sentences and build a lexicon of word types. We introduce DP-Parse, which uses similar principles but only relies on an instance lexicon of word tokens, avoiding the clustering errors that arise with a lexicon of word types.On the Zero Resource Speech Benchmark 2017, our model sets a new speech segmentation state-of-the-art in 5 languages.The algorithm monotonically improves with better input representations, achieving yet higher scores when fed with weakly supervised inputs. Despite lacking a type lexicon, DP-Parse can be pipelined to a language model and learn semantic and syntactic representations as assessed by a new spoken word embedding benchmark.

16:00-17:30 (East Foyer / Virtual)
**#151 On Graph-based Reentrancy-free Semantic Parsing**
*Alban Petit and Caio Corro*
We propose a novel graph-based approach for semantic parsing that resolves two problems observed in the literature:(1) seq2seq models fail on compositional generalization tasks;(2) previous work using phrase structure parsers cannot cover all the semantic parses observed in treebanks.We prove that both MAP inference and latent tag anchoring (required for weakly-supervised learning) are NP-hard problems.We propose two optimization algorithms based on constraint smoothing and conditional gradient to approximately solve these inference problems.Experimentally, our approach delivers state-of-the-art results on GeoQuery, Scan and Clevr, both for i.i.d. splits and for splits that test for compositional generalization.

16:00-17:30 (East Foyer / Virtual)
**#152 Exploring Contrast Consistency of Open-Domain Question Answering Systems on Minimally Edited Questions**
*Zhihan Zhang, Wenhao Yu, Zheng Ning, Mingxuan Ju and Meng Jiang*
Contrast consistency, the ability of a model to make consistently correct predictions in the presence of perturbations, is an essential aspect in NLP. While studied in tasks such as sentiment analysis and reading comprehension, it remains unexplored in open-domain question answering (OpenQA) due to the difficulty of collecting perturbed questions that satisfy factuality requirements. In this work, we collect minimally edited questions as challenging contrast sets to evaluate OpenQA models. Our collection approach combines both human annotation and large language model generation. We find that the widely used dense passage retriever (DPR) performs poorly on our contrast sets, despite fitting the training set well and performing competitively on standard test sets. To address this issue, we introduce a simple and effective query-side contrastive loss with the aid of data augmentation to improve DPR training. Our experiments on the contrast sets demonstrate that DPR's contrast consistency is improved without sacrificing its accuracy on the standard test sets.

16:00-17:30 (East Foyer / Virtual)
**#153 Intent-calibrated Self-training for Answer Selection in Open-domain Dialogues**
*Wentao Deng, Jiahuan Pei, Zhaochun Ren, Zhumin Chen and Pengjie Ren*
Answer selection in open-domain dialogues aims to select an accurate answer from candidates. Recent success of answer selection models hinges on training with large amounts of labeled data. However, collecting large-scale labeled data is labor-intensive and time-consuming. In this paper, we introduce the predicted intent labels to calibrate answer labels in a self-training paradigm. Specifically, we propose the ICAST to improve the quality of pseudo answer labels through the intent-calibrated answer selection paradigm, in which we employ pseudo intent labels to help improve pseudo answer labels. We carry out extensive experiments on two benchmark datasets with open-domain dialogues. The experimental results show that ICAST outperforms baselines consistently with 1%, 5% and 10% labeled data. Specifically, it improves 2.06% and 1.00% of F1 score on the two datasets, compared with the strongest baseline with only 5% labeled data.

16:00-17:30 (East Foyer / Virtual)
**#154 Optimal Transport Posterior Alignment for Cross-lingual Semantic Parsing**
*Tom Sherborne, Mirella Lapata and Tom Hosking*
Cross-lingual semantic parsing transfers parsing capability from a high-resource language (e.g., English) to low-resource languages with scarce training data. Previous work has primarily considered silver-standard data augmentation or zero-shot methods, however, exploiting few-shot gold data is comparatively unexplored. We propose a new approach to cross-lingual semantic parsing by explicitly minimizing cross-lingual divergence between probabilistic latent variables using Optimal Transport. We demonstrate how this direct guidance improves parsing from natural languages using fewer examples and less training. We evaluate our method on two datasets, MTOP and MultiATIS++SQL, establishing state-of-the-art results under a few-shot cross-lingual regime. Ablation studies further reveal that our method improves performance even without parallel input translations. In addition, we show that our model better captures cross-lingual structure in the latent space to improve semantic representation similarity.

16:00-17:30 (East Foyer / Virtual)
**#155 Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision**
*Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi and Neil Zeghidour*
We introduce SPEAR-TTS, a multi-speaker text-to-speech (TTS) system that can be trained with minimal supervision. By combining two types of discrete speech representations, we cast TTS as a composition of two sequence-to-sequence tasks: from text to high-level semantic tokens (akin to "reading") and from semantic tokens to low-level acoustic tokens ("speaking"). Decoupling these two tasks enables training of the "speaking" module using abundant audio-only data, and unlocks the highly efficient combination of pretraining and backtranslation to reduce the need for parallel data when training the "reading" component. To control the speaker identity, we adopt example prompting, which allows SPEAR-TTS to generalize to unseen speakers using only a short sample of 3 seconds, without any explicit speaker representation or speaker labels. Our experiments demonstrate that SPEAR-TTS achieves a character error rate that is competitive with state-of-the-art methods using only 15 minutes of parallel data, while matching ground-truth speech in naturalness and acoustic quality.

16:00-17:30 (East Foyer / Virtual)
**#156 PaniniQA: Enhancing Patient Education Through Interactive Question Answering**
*Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz and Hong Yu*
Patient portal allows discharged patients to access their personalized discharge instructions in electronic health records (EHRs). However, many patients have difficulty understanding or memorizing their discharge instructions. In this paper, we present PaniniQA, a patient-centric interactive question answering system designed to help patients understand their discharge instructions. PaniniQA first identifies important clinical content from patients' discharge instructions and then formulates patient-specific educational questions.In addition, PaniniQA is also equipped with answer verification functionality to provide timely feedback to correct patients' misunderstandings. Our comprehensive automatic and human evaluation results demonstrate our PaniniQA is capable of improving patients' mastery of their medical instructions through

effective interactions.

16:00-17:30 (East Foyer / Virtual)
**#157 ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation**
*Zhengxuan Wu, Christopher Manning and Christopher Potts*
Compositional generalization benchmarks seek to assess whether models can accurately compute meanings for novel sentences, but operationalize this in terms of logical form (LF) prediction. This raises the concern that semantically irrelevant details of the chosen LFs could shape model performance. We argue that this concern is realized for the COGS benchmark (Kim and Linzen, 2020). COGS poses generalization splits that appear impossible for present-day models, which could be taken as an indictment of those models. However, we show that the negative results trace to incidental features of COGS LFs. Converting these LFs to semantically equivalent ones and factoring out capabilities unrelated to semantic interpretation, we find that even baseline models get traction. A recent variable-free translation of COGS LFs suggests similar conclusions, but we observe this format is not semantically equivalent; it is incapable of accurately representing some COGS meanings. These findings inform our proposal for ReCOGS, a modified version of COGS that comes closer to assessing the target semantic capabilities while remaining very challenging. Overall, our results reaffirm the importance of compositional generalization and careful benchmark task design.

16:00-17:30 (East Foyer / Virtual)
**#158 Calibrated Interpretation: Confidence Estimation in Semantic Parsing**
*Elias Stengel-Eskin and Benjamin Van Durme*
Sequence generation models are increasingly being used to translate natural language into programs, i.e. to perform executable semantic parsing. The fact that semantic parsing aims to predict programs that can lead to executed actions in the real world motivates developing safe systems. This in turn makes measuring calibration – a central component to safety – particularly important. We investigate the calibration of popular generation models across four popular semantic parsing datasets, finding that it varies across models and datasets. We then analyze factors associated with calibration error and release new confidence-based challenge splits of two parsing datasets. To facilitate the inclusion of calibration in semantic parsing evaluations, we release a library for computing calibration metrics.

16:00-17:30 (East Foyer / Virtual)
**#159 QAmeleon: Multilingual QA with Only 5 Examples**
*Chris Alberti, Priyanka Agrawal, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das and Mirella Lapata*
The availability of large, high-quality datasets has been a major driver of recent progress in question answering (QA). Such annotated datasets, however, are difficult and costly to collect, and rarely exist in languages other than English, rendering QA technology inaccessible to underrepresented languages. An alternative to building large monolingual training datasets is to leverage pre-trained language models (PLMs) under a few-shot learning setting. Our approach, QAmeleon, uses a PLM to automatically generate multilingual data upon which QA models are fine-tuned, thus avoiding costly annotation. Prompt tuning the PLM with only five examples per language delivers accuracy superior to translation-based baselines; it bridges nearly 60% of the gap between an English-only baseline and a fully-supervised upper bound fine-tuned on almost 50,000 hand-labeled examples; and consistently leads to improvements compared to directly fine-tuning a QA model on labeled examples in low resource settings. Experiments on the TyDiQA-GoldP and MLQA benchmarks show that few-shot prompt tuning for data synthesis scales across languages and is a viable alternative to large-scale annotation.

16:00-17:30 (East Foyer / Virtual)
**#160 MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis**
*Haifeng Hu and Ronghao Lin*
When applying multimodal machine learning in downstream inference, both joint and coordinated multimodal representations rely on the complete presence of modalities as in training. However, the modal-incomplete data, where certain modalities are missing, greatly reduces performance in Multimodal Sentiment Analysis (MSA) due to varying input forms and semantic information deficiencies. This limits the applicability of the predominant MSA methods in the real world, where the completeness of multimodal data is uncertain and variable. The generation-based methods attempt to generate the missing modality, yet they require complex hierarchical architecture with huge computational costs and struggle with the representation gaps across different modalities. Diversely, we propose a novel representation learning approach named MissModal, devoting to increasing robustness to missing modality in a classification approach. Specifically, we adopt constraints with geometric contrastive loss, distribution distance loss, and sentiment semantic loss to align the representations of modal-missing and modal-complete data, without impacting the sentiment inference for the complete modalities. Furthermore, we do not demand any changes in the multimodal fusion stage, highlighting the generality of our method in other multimodal learning systems. Extensive experiments demonstrate that the proposed method achieves superior performance with minimal computational costs in various missing modalities scenarios (flexibility), including severely missing modality (efficiency) on two public MSA datasets.

16:00-17:30 (East Foyer / Virtual)
**#161 Languages through the Looking Glass of BPE Compression**
*Ximena Gutierrez-Vasques, Christian Bentz and Tanja Samardžić*
Byte-pair encoding (BPE) is widely used in NLP for performing subword tokenization. It uncovers redundantpatterns for compressing the data, and hence alleviates the sparsity problem in downstream applications.Subwords discovered during the first merge operations tend to have the most substantial impact on thecompression of texts. However, the structural underpinnings of this effect have not been analyzedcrosslinguistically. We conduct in-depth analyses across 47 typologically diverse languages and threeparallel corpora, and thereby show that the types of recurrent patterns that have the strongest impact oncompression are an indicator of morphological typology. For languages with richer inflectional morphologythere is a preference for highly productive subwords on the early merges, while for languages with lessinflectional morphology, idiosyncratic subwords are more prominent. Both types of patterns contribute toefficient compression. Counter to the common perception that BPE subwords are not linguistically relevant,we find patterns across languages that resemble those described in traditional typology. We thus propose anovel way to characterize languages according to their BPE subword properties, inspired by the notion ofmorphological productivity in linguistics. This allows us to have language vectors that encode typologicalknowledge induced from raw text. Our approach is easily applicable to a wider range of languages and texts,as it does not require annotated data or any external linguistic knowledge. We discuss its potentialcontributions to quantitative typology and multilingual NLP.

16:00-17:30 (East Foyer / Virtual)
**#162 Language Embeddings Sometimes Contain Typological Generalizations**
*Robert Östling and Murathan Kurfali*
To what extent can neural network models learn generalizations about language structure, and how do we find out what they have learned? We explore these questions by training neural models for a range of natural language processing tasks on a massively multilingual dataset of Bible translations in 1,295 languages. The learned language representations are then compared to existing typological databases as well as to a novel set of quantitative syntactic and morphological features obtained through annotation projection. We conclude that some generalizations are surprisingly close to traditional features from linguistic typology, but that most of our models, as well as those of previous work, do not

appear to have made linguistically meaningful generalizations. Careful attention to details in the evaluation turns out to be essential to avoid false positives. Furthermore, to encourage continued work in this field, we release several resources covering most or all of the languages in our data: (1) multiple sets of language representations, (2) multilingual word embeddings, (3) projected and predicted syntactic and morphological features, (4) software to provide linguistically sound evaluations of language representations.

16:00-17:30 (East Foyer / Virtual)
**#163 Universal Generation for Optimality Theory Is PSPACE-Complete**
*Sophie Hao*
This paper shows that the universal generation problem (Heinz, Kobele, and Riggle 2009) for Optimality Theory (OT, Prince and Smolensky 1993, 2004) is PSPACE-complete. While prior work has shown that universal generation is at least NP-hard (Eisner 1997, 2000b; Wareham 1998; Idsardi 2006) and at most EXPSPACE-hard (Riggle 2004), our results place universal generation in between those two classes, assuming that NP $\neq$ PSPACE. We additionally show that when the number of constraints is bounded in advance, universal generation is at least NL-hard and at most NPNP-hard. Our proofs rely on a close connection between OT and the intersection non-emptiness problem for finite automata, which is PSPACE-complete in general (Kozen 1977) and NL-complete when the number of automata is bounded (Jones 1975). Our analysis shows that constraint interaction is the main contributor to the complexity of OT: the ability to factor transformations into simple, interacting constraints allows OT to furnish compact descriptions of intricate phonological phenomena.

16:00-17:30 (East Foyer / Virtual)
**#164 Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering**
*Akshay Chaturvedi, Swarnadeep Bhar, Soumadeep Saha, Utpal Garain and Nicholas Asher*
Transformer-based language models have been shown to be highly effective for several NLP tasks. In this paper, we consider three transformer models, BERT, RoBERTa, and XLNet, in both small and large versions, and investigate how faithful their representations are with respect to the semantic content of texts. We formalize a notion of semantic faithfulness, in which the semantic content of a text should causally figure in a model's inferences in question answering. We then test this notion by observing a model's behavior on answering questions about a story after performing two novel semantic interventions—deletion intervention and negation intervention. While transformer models achieve high performance on standard question answering tasks, we show that they fail to be semantically faithful once we perform these interventions for a significant number of cases ($\sim$ 50% for deletion intervention, and $\sim$ 20% drop in accuracy for negation intervention). We then propose an intervention-based training regime that can mitigate the undesirable effects for deletion intervention by a significant margin (from $\sim$ 50% to $\sim$ 6%). We analyze the inner-workings of the models to better understand the effectiveness of intervention-based training for deletion intervention. But we show that this training does not attenuate other aspects of semantic unfaithfulness such as the models' inability to deal with negation intervention or to capture the predicate-argument structure of texts. We also test InstructGPT, via prompting, for its ability to handle the two interventions and to capture predicate-argument structure. While InstructGPT models do achieve very high performance on predicate-argument structure task, they fail to respond adequately to our deletion and negation interventions.

## Findings 3

16:00-17:30 (East Foyer / Virtual)

---

16:00-17:30 (East Foyer / Virtual)
**ARKitSceneRefer: Text-based Localization of Small Objects in Diverse Real-World 3D Indoor Scenes**
*Shunya Kato, Shuhei Kurita, Chenhui Chu and Sadao Kurohashi*
3D referring expression comprehension is a task to ground text representations onto objects in 3D scenes. It is a crucial task for indoor household robots or augmented reality devices to localize objects referred to in user instructions. However, existing indoor 3D referring expression comprehension datasets typically cover larger object classes that are easy to localize, such as chairs, tables, or doors, and often overlook small objects, such as cooking tools or office supplies. Based on the recently proposed diverse and high-resolution 3D scene dataset of ARKitScenes, we construct the ARKitSceneRefer dataset focusing on small daily-use objects that frequently appear in real-world indoor scenes. ARKitSceneRefer contains 15k objects of 1,605 indoor scenes, which are significantly larger than those of the existing 3D referring datasets, and covers diverse object classes of 583 from the LVIS dataset. In empirical experiments with both 2D and 3D state-of-the-art referring expression comprehension models, we observed the task difficulty of the localization in the diverse small object classes.

16:00-17:30 (East Foyer / Virtual)
**Non-parallel Accent Transfer based on Fine-grained Controllable Accent Modelling**
*Linqin Wang, Zhengtao Yu, Yuanzhang Yang, Shengxiang Gao, Cunli Mao and Yuxin Huang*
Existing accent transfer works rely on parallel data or speech recognition models. This paper focuses on the practical application of accent transfer and aims to implement accent transfer using non-parallel datasets. The study has encountered the challenge of speech representation disentanglement and modeling accents. In our accent modeling transfer framework, we manage to solve these problems by two proposed methods. First, we learn the suprasegmental information associated with tone to finely model the accents in terms of tone and rhythm. Second, we propose to use mutual information learning to disentangle the accent features and control the accent of the generated speech during the inference time. Experiments show that the proposed framework attains superior performance to the baseline models in terms of accentedness and audio quality.

16:00-17:30 (East Foyer / Virtual)
**ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback**
*Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi and Zhaopeng Tu*
Large language models (LLMs) like ChatGPT have exhibited remarkable abilities on a wide range of natural language processing (NLP) tasks, including various machine translation abilities accomplished during chat. However, these models are only accessible through restricted APIs, which creates barriers to new research and advancements in the field. Therefore, we propose ParroT, a framework to enhance and regulate the translation abilities during chat based on open-source LLMs (e.g., LLaMA), human-written translation and feedback data. Specifically, ParroT reformulates translation data into the instruction-following style, and introduces a "Hint" field for incorporating extra requirements to regulate the translation process. Accordingly, we propose three instruction types for finetuning ParroT models, including translation instruction, contrastive instruction, and error-guided instruction. Experiments on Flores subsets and WMT22 test sets suggest that translation instruction improves the translation performance of vanilla LLMs significantly while error-guided instruction can lead to further improvement, which demonstrates the importance of learning from low-quality translations annotated by humans. We also demonstrate the potential of automatic evaluation tools in providing quality information of translations, when constructing error-guided instructions for directions that lack human annotation data. Please refer to our Github project for more implementation details: https://github.com/wxjiao/ParroT.

16:00-17:30 (East Foyer / Virtual)

**Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models**
*Hongli Zhan, Desmond Ong and Junyi Jessy Li*
The emotions we experience involve complex processes; besides physiological aspects, research in psychology has studied cognitive appraisals where people assess their situations subjectively, according to their own values (Scherer, 2005). Thus, the same situation can often result in different emotional experiences. While the detection of emotion is a well-established task, there is very limited work so far on the automatic prediction of cognitive appraisals. This work fills the gap by presenting CovidET-Appraisals, the most comprehensive dataset to-date that assesses 24 appraisal dimensions, each with a natural language rationale, across 241 Reddit posts. CovidET-Appraisals presents an ideal testbed to evaluate the ability of large language models — excelling at a wide range of NLP tasks — to automatically assess and explain cognitive appraisals. We found that while the best models are performant, open-sourced LLMs fall short at this task, presenting a new challenge in the future development of emotionally intelligent models. We release our dataset at https://github.com/honglizhan/CovidET-Appraisals-Public.

16:00-17:30 (East Foyer / Virtual)
**Identifying Early Maladaptive Schemas from Mental Health Question Texts**
*Sujatha Das Gollapalli, Beng Heng Ang and See-Kiong Ng*
In Psychotherapy, maladaptive schemas– negative perceptions that an individual has of the self, others, or the world that endure despite objective reality–often lead to resistance to treatments and relapse of mental health issues such as depression, anxiety, panic attacks etc. Identification of early maladaptive schemas (EMS) is thus a crucial step during Schema Therapy-based counseling sessions, where patients go through a detailed and lengthy EMS questionnaire. However, such an approach is not practical in 'offline' counseling scenarios, such as community QA forums which are gaining popularity for people seeking mental health support. In this paper, we investigate both LLM (Large Language Models) and non-LLM approaches for identifying EMS labels using resources from Schema Therapy. Our evaluation indicates that recent LLMs can be effective for identifying EMS but their predictions lack explainability and are too sensitive to precise 'prompts'. Both LLM and non-LLM methods are unable to reliably address the null cases, i.e. cases with no EMS labels. However, we posit that the two approaches show complementary properties and together, they can be used to further devise techniques for EMS identification.

16:00-17:30 (East Foyer / Virtual)
**Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding**
*Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang and Jianfeng Gao*
Scientific literature understanding tasks have gained significant attention due to their potential to accelerate scientific discovery. Pre-trained language models (LMs) have shown effectiveness in these tasks, especially when tuned via contrastive learning. However, jointly utilizing pre-training data across multiple heterogeneous tasks (e.g., extreme multi-label paper classification, citation prediction, and literature search) remains largely unexplored. To bridge this gap, we propose a multi-task contrastive learning framework, SciMult, with a focus on facilitating common knowledge sharing across different scientific literature understanding tasks while preventing task-specific skills from interfering with each other. To be specific, we explore two techniques – task-aware specialization and instruction tuning. The former adopts a Mixture-of-Experts Transformer architecture with task-aware sub-layers; the latter prepends task-specific instructions to the input text so as to produce task-aware outputs. Extensive experiments on a comprehensive collection of benchmark datasets verify the effectiveness of our task-aware specialization strategy, where we outperform state-of-the-art scientific pre-trained LMs. Code, datasets, and pre-trained models can be found at https://scimult.github.io/.

16:00-17:30 (East Foyer / Virtual)
**Medical Text Simplification: Optimizing for Readability with Unlikelihood Training and Reranked Beam Search Decoding**
*Lorenzo Jaime Yu Flores, Heyuan Huang, Kejian Shi, Sophie Chheang and Arman Cohan*
Text simplification has emerged as an increasingly useful application of AI for bridging the communication gap in specialized fields such as medicine, where the lexicon is often dominated by technical jargon and complex constructs. Despite notable progress, methods in medical simplification sometimes result in the generated text having lower quality and diversity. In this work, we explore ways to further improve the readability of text simplification in the medical domain. We propose (1) a new unlikelihood loss that encourages generation of simpler terms and (2) a reranked beam search decoding method that optimizes for simplicity, which achieve better performance on readability metrics on three datasets. This study's findings offer promising avenues for improving text simplification in the medical field.

16:00-17:30 (East Foyer / Virtual)
**ZARA: Improving Few-Shot Self-Rationalization for Small Language Models**
*Wei-Lin Chen, An-Zi Yen, Cheng-Kuang Wu, Hen-Hsen Huang and Hsin-Hsi Chen*
Language models (LMs) that jointly generate end-task answers as well as free-text rationales are known as self-rationalization models. Recent works demonstrate great performance gain for self-rationalization by few-shot prompting LMs with rationale-augmented exemplars. However, the ability to benefit from explanations only emerges with large-scale LMs, which have poor accessibility. In this work, we explore the less-studied setting of leveraging explanations for small LMs to improve few-shot self-rationalization. We first revisit the relationship between rationales and answers. Inspired by the implicit mental process of how human beings assess explanations, we present a novel approach, Zero-shot Augmentation of Rationale-Answer pairs (ZARA), to automatically construct pseudo-parallel data for self-training by reducing the problem of plausibility judgement to natural language inference. Experimental results show ZARA achieves SOTA performance on the FEB benchmark, for both the task accuracy and the explanation metric. In addition, we conduct human and quantitative evaluation validating ZARA's ability to automatically identify plausible and accurate rationale-answer pairs.

16:00-17:30 (East Foyer / Virtual)
**Domain Adaptation for Conversational Query Production with the RAG Model Feedback**
*Ante Wang, Linfeng Song, Ge Xu and Jinsong Su*
Conversational query production is an emerging fundamental task for the dialogue system, where search queries are generated to explore the vast and continually updating knowledge from a search engine. To accelerate this line of research, previous studies have released several datasets with human-annotated search queries. However, the limited annotations still can not cover conversations of various domains. To solve this challenge, we propose a novel domain adaptation framework. It is inspired by a weakly supervised learning algorithm from previous work that guides a model using reinforcement learning with BM25 scores as feedback. Though effective, it is fragile facing noisy content on webpages from a commercial search engine and variance in conversations because of ignoring deep semantic information of dialogue contexts. Thus, we improve the algorithm by taking the advance of retrieval-augmented generation (RAG) and exploring several practical techniques such as knowledge distillation for stable training. We conduct experiments in multiple settings across different languages. Guided by the RAG model feedback, our model is more robust and performs significantly better especially in a more challenging setting over strong baselines.

16:00-17:30 (East Foyer / Virtual)
**Towards Being Parameter-Efficient: A Stratified Sparsely Activated Transformer with Dynamic Capacity**
*Haoran Xu, Maha Elbayad, Kenton Murray, Jean Maillard and Vedanuj Goswami*

Mixture-of-experts (MoE) models that employ sparse activation have demonstrated effectiveness in significantly increasing the number of parameters while maintaining low computational requirements per token. However, recent studies have established that MoE models are inherently parameter-inefficient as the improvement in performance diminishes with an increasing number of experts. We hypothesize this parameter inefficiency is a result of all experts having equal capacity, which may not adequately meet the varying complexity requirements of different tokens or tasks. In light of this, we propose Stratified Mixture of Experts (SMoE) models, which feature a stratified structure and can assign dynamic capacity to different tokens. We demonstrate the effectiveness of SMoE on three multilingual machine translation benchmarks, containing 4, 15, and 94 language pairs, respectively. We show that SMoE outperforms multiple state-of-the-art MoE models with the same or fewer parameters.

16:00-17:30 (East Foyer / Virtual)
**Orthogonal Subspace Learning for Language Model Continual Learning**
*Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui and Xuanjing Huang*
Benefiting from massive corpora and advanced hardware, large language models (LLMs) exhibit remarkable capabilities in language understanding and generation. However, their performance degrades in scenarios where multiple tasks are encountered sequentially, also known as catastrophic forgetting. In this paper, we propose orthogonal low-rank adaptation (O-LoRA), a simple and efficient approach for continual learning in language models, effectively mitigating catastrophic forgetting while learning new tasks. Specifically, O-LoRA learns tasks in different (low-rank) vector subspaces that are kept orthogonal to each other in order to minimize interference. Our method induces only marginal additional parameter costs and requires no user data storage for replay. Experimental results on continual learning benchmarks show that our method outperforms state-of-the-art methods. Furthermore, compared to previous approaches, our method excels in preserving the generalization ability of LLMs on unseen tasks.

16:00-17:30 (East Foyer / Virtual)
**On the Risk of Misinformation Pollution with Large Language Models**
*Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan and William Yang Wang*
We investigate the potential misuse of modern Large Language Models (LLMs) for generating credible-sounding misinformation and its subsequent impact on information-intensive applications, particularly Open-Domain Question Answering (ODQA) systems. We establish a threat model and simulate potential misuse scenarios, both unintentional and intentional, to assess the extent to which LLMs can be utilized to produce misinformation. Our study reveals that LLMs can act as effective misinformation generators, leading to a significant degradation (up to 87%) in the performance of ODQA systems. Moreover, we uncover disparities in the attributes associated with persuading humans and machines, presenting an obstacle to current human-centric approaches to combat misinformation. To mitigate the harm caused by LLM-generated misinformation, we propose three defense strategies: misinformation detection, vigilant prompting, and reader ensemble. These approaches have demonstrated promising results, albeit with certain associated costs. Lastly, we discuss the practicality of utilizing LLMs as automatic misinformation generators and provide relevant resources and code to facilitate future research in this area.

16:00-17:30 (East Foyer / Virtual)
**Decomposing Complex Queries for Tip-of-the-tongue Retrieval**
*Kevin Lin, Kyle Lo, Joseph E. Gonzalez and Dan Klein*
When re-finding items, users who forget or are uncertain about identifying details often rely on creative strategies for expressing their information needs—complex queries that describe content elements (e.g., book characters or events), information beyond the document text (e.g., descriptions of book covers), or personal context (e.g., when they read a book). Standard retrieval models that rely on lexical or semantic overlap between query and document text are challenged in such retrieval settings, known as tip-of-the-tongue (TOT) retrieval. We introduce a simple but effective framework for handling such complex queries by decomposing the query with an LLM into individual clues routing those as subqueries to specialized retrievers, and ensembling the results. Our approach takes advantage of off-the-shelf retrievers (e.g., CLIP for retrieving images of book covers) or incorporate retriever-specific logic (e.g., date constraints). We show that our framework incorporating query decomposition into retrievers can improve gold book recall up to 6% absolute gain for Recall@5 on a new collection of 14,441 real-world query-book pairs from an online community for resolving TOT inquiries.

16:00-17:30 (East Foyer / Virtual)
**Uncovering the Root of Hate Speech: A Dataset for Identifying Hate Instigating Speech**
*Hyoungjun Park, Ho Sung Shim and Kyuhan Lee*
While many prior studies have applied computational approaches, such as machine learning, to detect and moderate hate speech, only scant attention has been paid to the task of identifying the underlying cause of hate speech. In this study, we introduce the concept of hate instigating speech, which refers to a specific type of textual posts on online platforms that stimulate or provoke others to engage in hate speech. The identification of hate instigating speech carries substantial practical implications for effective hate speech moderation. Rather than targeting individual instances of hate speech, by focusing on their roots, i.e., hate instigating speech, it becomes possible to significantly reduce the volume of content that requires review for moderation. Additionally, targeting hate instigating speech enables early prevention of the spread and propagation of hate speech, further enhancing the effectiveness of moderation efforts. However, several challenges hinder researchers from addressing the identification of hate instigating speech. First, there is a lack of comprehensive datasets specifically annotated for hate instigation, making it difficult to train and evaluate computational models effectively. Second, the subtle and nuanced nature of hate instigating speech (e.g., seemingly non-offensive texts serve as catalysts for triggering hate speech) makes it difficult to apply off-the-shelf machine learning models to the problem. To address these challenges, in this study, we have developed and released a multilingual dataset specifically designed for the task of identifying hate instigating speech. Specifically, it encompasses both English and Korean, allowing for a comprehensive examination of hate instigating speech across different linguistic contexts. We have applied existing machine learning models to our dataset and the results demonstrate that the extant models alone are insufficient for effectively detecting hate instigating speech. This finding highlights the need for further attention from the academic community to address this specific challenge. We expect our study and dataset to inspire researchers to explore innovative methods that can enhance the accuracy of hate instigating speech detection, ultimately contributing to more effective moderation and prevention of hate speech propagation online.

16:00-17:30 (East Foyer / Virtual)
**Ask To The Point: Open-Domain Entity-Centric Question Generation**
*Yuxiang Liu, Jie Huang and Kevin Chang*
We introduce a new task called *entity-centric question generation* (ECQG), motivated by real-world applications such as topic-specific learning, assisted reading, and fact-checking. The task aims to generate questions from an entity perspective. To solve ECQG, we propose a coherent PLM-based framework GenCONE with two novel modules: content focusing and question verification. The content focusing module first identifies a focus as "what to ask" to form draft questions, and the question verification module refines the questions afterwards by verifying the answerability. We also construct a large-scale open-domain dataset from SQuAD to support this task. Our extensive experiments demonstrate that GenCONE significantly and consistently outperforms various baselines, and two modules are effective and complementary in generating high-quality questions.

16:00-17:30 (East Foyer / Virtual)
**SAC$^3$: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency**
*Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin and Sricharan Kumar*
Hallucination detection is a critical step toward understanding the trustworthiness of modern language models (LMs). To achieve this goal, we re-examine existing detection approaches based on the self-consistency of LMs and uncover two types of hallucinations resulting from 1) question-level and 2) model-level, which cannot be effectively identified through self-consistency check alone. Building upon this discovery, we propose a novel sampling-based method, i.e., semantic-aware cross-check consistency (SAC$^3$) that expands on the principle of self-consistency checking. Our SAC$^3$ approach incorporates additional mechanisms to detect both question-level and model-level hallucinations by leveraging advances including semantically equivalent question perturbation and cross-model response consistency checking. Through extensive and systematic empirical analysis, we demonstrate that SAC$^3$ outperforms the state of the art in detecting both non-factual and factual statements across multiple question-answering and open-domain generation benchmarks.

16:00-17:30 (East Foyer / Virtual)
**How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench**
*Qinyuan Ye, Harvey Yiyun Fu, Xiang Ren and Robin Jia*
We investigate the predictability of large language model (LLM) capabilities: given records of past experiments using different model families, numbers of parameters, tasks, and numbers of in-context examples, can we accurately predict LLM performance on new experiment configurations? Answering this question has practical implications for LLM users (e.g., deciding which models to try), developers (e.g., prioritizing evaluation on representative tasks), and the research community (e.g., identifying hard-to-predict capabilities that warrant further investigation). We study the performance prediction problem on experiment records from BIG-bench. On a random train-test split, an MLP-based predictor achieves an $R^2$ score greater than 95%, indicating the presence of learnable patterns within the experiment records. We then formulate the problem of searching for "small-bench," an informative subset of BIG-bench tasks from which the performance on the full set can be maximally recovered. We find a subset as informative as BIG-bench Hard for evaluating new model families, while being $3\times$ smaller. Additionally, we find competitive subsets by clustering task representations learned by our MLP-based predictor and selecting tasks close to cluster centroids, highlighting the importance of task diversity in constructing "small-bench."

16:00-17:30 (East Foyer / Virtual)
**Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting**
*Zhiyu Chen, Yujie Lu and William Yang Wang*
Mental illness remains one of the most critical public health issues of our time, due to the severe scarcity and accessibility limit of professionals. Psychotherapy requires high-level expertise to conduct deep, complex reasoning and analysis on the cognition modeling of the patients. In the era of Large Language Models, we believe it is the right time to develop AI assistance for computational psychotherapy. We study the task of cognitive distortion detection and propose the Diagnosis of Thought (DoT) prompting. DoT performs diagnosis on the patient's speech via three stages: subjectivity assessment to separate the facts and the thoughts; contrastive reasoning to elicit the reasoning processes supporting and contradicting the thoughts; and schema analysis to summarize the cognition schemas. The generated diagnosis rationales through the three stages are essential for assisting the professionals. Experiments demonstrate that DoT obtains significant improvements over ChatGPT for cognitive distortion detection, while generating high-quality rationales approved by human experts.

16:00-17:30 (East Foyer / Virtual)
**Towards Informative Few-Shot Prompt with Maximum Information Gain for In-Context Learning**
*Hongfu Liu and Ye Wang*
Large Language models (LLMs) possess the capability to engage In-context Learning (ICL) by leveraging a few demonstrations pertaining to a new downstream task as conditions. However, this particular learning paradigm suffers from high instability stemming from substantial variances induced by factors such as the input distribution of selected examples, their ordering, and prompt formats. In this work, we demonstrate that even when all these factors are held constant, the random selection of examples still results in high variance. Consequently, we aim to explore the informative ability of data examples by quantifying the Information Gain (IG) obtained in prediction after observing a given example candidate. Then we propose to sample those with maximum IG. Additionally, we identify the presence of template bias, which can lead to unfair evaluations of IG during the sampling process. To mitigate this bias, we introduce Calibration Before Sampling strategy. The experimental results illustrate that our proposed method can yield an average relative improvement of 14.3% across six classification tasks using three LLMs.

16:00-17:30 (East Foyer / Virtual)
**Implicit Sense-labeled Connective Recognition as Text Generation**
*Yui Oka and Tsutomu Hirao*
Implicit Discourse Relation Recognition (IDRR) involves identifying the sense label of an implicit connective between adjacent text spans. This has traditionally been approached as a classification task. However, some downstream tasks require more than just a sense label as well as the specific connective used. This paper presents Implicit Sense-labeled Connective Recognition (ISCR), which identifies the implicit connectives and their sense labels between adjacent text spans. ISCR can be treated as a classification task, but a large number of potential categories, sense labels, and uneven distribution of instances among them make this difficult. Instead, this paper handles the task as a text-generation task, using an encoder-decoder model to generate both connectives and their sense labels. Here, we explore a classification method and three kinds of text-generation methods. From our evaluation results on PDTB-3.0, we found that our method outperforms the conventional classification-based method.

16:00-17:30 (East Foyer / Virtual)
**MaXM: Towards Multilingual Visual Question Answering**
*Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen and Radu Soricut*
Visual Question Answering (VQA) has been primarily studied through the lens of the English language. Yet, tackling VQA in other languages in the same manner would require a considerable amount of resources. In this paper, we propose scalable solutions to multilingual visual question answering (mVQA), on both data and modeling fronts. We first propose a translation-based framework to mVQA data generation that requires much less human annotation efforts than the conventional approach of directly collection questions and answers. Then, we apply our framework to the multilingual captions in the Crossmodal-3600 dataset and develop an efficient annotation protocol to create MaXM, a test-only VQA benchmark in 7 diverse languages. Finally, we develop a simple, lightweight, and effective approach as well as benchmark state-of-the-art English and multilingual VQA models. We hope that our benchmark encourages further research on mVQA.

16:00-17:30 (East Foyer / Virtual)
**Variator: Accelerating Pre-trained Models with Plug-and-Play Compression Modules**
*Chaojun Xiao, Yuqi Luo, Wenbin Zhang, Pengle Zhang, Xu Han, Yankai Lin, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Maosong Sun and*

*Jie Zhou*
Large language models (LLMs) have achieved remarkable results on NLP tasks but at the expense of huge parameter sizes and the consequent computational costs. In this paper, we propose Variator, a parameter-efficient acceleration method that enhances computational efficiency through plug-and-play compression plugins. Compression plugins are designed to reduce the sequence length via compressing multiple hidden vectors into one and trained with original LLMs frozen. Different from traditional model acceleration methods, which compress LLMs to smaller sizes, Variator offers two distinct advantages: (1) In real-world applications, the plug-and-play nature of our compression plugins enables dynamic selection of different compression plugins with varying acceleration ratios based on the current workload. (2) The compression plugin comprises a few compact neural network layers with minimal parameters, significantly saving storage and memory overhead, particularly in scenarios with a growing number of tasks. We validate the effectiveness of Variator on seven datasets. Experimental results show that Variator can save 53% computational costs using only 0.9% additional parameters with a performance drop of less than 2%. Moreover, when the model scales to billions of parameters, Variator matches the strong performance of uncompressed LLMs. Our code and checkpoints will be released to facilitate future work.

16:00-17:30 (East Foyer / Virtual)
**Is Probing All You Need? Indicator Tasks as an Alternative to Probing Embedding Spaces**
*Tal Levy, Omer Goldman and Reut Tsarfaty*
The ability to identify and control different kinds of linguistic information encoded in vector representations of words has many use cases, especially for explainability and bias removal. This is usually done via a set of simple classification tasks, termed *probes*, to evaluate the information encoded in the embedding space. However, the involvement of a trainable classifier leads to entanglement between the probe's results and the classifier's nature. As a result, contemporary works on probing include tasks that do not involve training of auxiliary models. In this work we introduce the term *indicator tasks* for non-trainable tasks which are used to query embedding spaces for the existence of certain properties, and claim that this kind of tasks may point to a direction opposite to probes, and that this contradiction complicates the decision on whether a property exists in an embedding space. We demonstrate our claims with two test cases, one dealing with gender debiasing and another with the erasure of morphological information from embedding spaces. We show that the application of a suitable indicator provides a more accurate picture of the information captured and removed compared to probes. We thus conclude that indicator tasks should be implemented and taken into consideration when eliciting information from embedded representations.

16:00-17:30 (East Foyer / Virtual)
**Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems**
*Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng and Kai-Wei Chang*
Recent advancements in Large Language Models empower them to follow freeform instructions, including imitating generic or specific demographic personas in conversations. We define generic personas to represent demographic groups, such as "an Asian person", whereas specific personas may take the form of specific popular Asian names like "Yumi". While the adoption of personas enriches user experiences by making dialogue systems more engaging and approachable, it also casts a shadow of potential risk by exacerbating social biases within model responses, thereby causing societal harm through interactions with users. In this paper, we systematically study "persona biases", which we define to be the sensitivity of dialogue models' harmful behaviors contingent upon the personas they adopt. We categorize persona biases into biases in harmful expression and harmful agreement, and establish a comprehensive evaluation framework to measure persona biases in five aspects: Offensiveness, Toxic Continuation, Regard, Stereotype Agreement, and Toxic Agreement. Additionally, we propose to investigate persona biases by experimenting with UNIVERSALPERSONA, a systematically constructed persona dataset encompassing various types of both generic and specific model personas. Through benchmarking on four different models- including Blender, ChatGPT, Alpaca, and Vicuna- our study uncovers significant persona biases in dialogue systems. Our findings also underscore the pressing need to revisit the use of personas in dialogue agents to ensure safe application.

16:00-17:30 (East Foyer / Virtual)
**Orca: A Few-shot Benchmark for Chinese Conversational Machine Reading Comprehension**
*Nuo Chen, Hongguang Li, Junqing He, Yinan Bao, Xinshi Lin, Qi Yang, Jianfeng Liu, Ruyi Gan, Jiaxing Zhang, Baoyuan Wang and Jia Li*
The conversational machine reading comprehension (CMRC) task aims to answer questions in conversations, which has been a hot research topic in recent years because of its wide applications. However, existing CMRC benchmarks in which each conversation is assigned a static passage are inconsistent with real scenarios. Thus, model's comprehension ability towards real scenarios are hard to evaluate reasonably. To this end, we propose the first Chinese CMRC benchmark **Orca** and further provide zero-shot/few-shot settings to evaluate model's generalization ability towards diverse domains. We collect 831 hot-topic driven conversations with 4,742 turns in total. Each turn of a conversation is assigned with a response-related passage, aiming to evaluate model's comprehension ability more reasonably. The topics of conversations are collected from social media platform and cover 33 domains, trying to be consistent with real scenarios. Importantly, answers in Orca are all well-annotated natural responses rather than the specific spans or short phrase in previous datasets. Besides, we implement three strong baselines to tackle the challenge in Orca. The results indicate the great challenge of our CMRC benchmark.

16:00-17:30 (East Foyer / Virtual)
**Late Fusion of Transformers for Sentiment Analysis of Code-Switched Data**
*Gagan Sharma, R Chinmay and Raksha Sharma*
Code-switching is a common phenomenon in multilingual communities and is often used on social media. However, sentiment analysis of code-switched data is a challenging yet less explored area of research. This paper aims to develop a sentiment analysis system for code-switched data. In this paper, we present a novel approach combining two transformers using logits of their output and feeding them to a neural network for classification. We show the efficacy of our approach using two benchmark datasets, viz., English-Hindi (En-Hi), and English-Spanish (En-Es) availed by Microsoft GLUECoS. Our approach results in an F1 score of 73.66% for En-Es and 61.24% for En-Hi, significantly higher than the best model reported for the GLUECoS benchmark dataset.

16:00-17:30 (East Foyer / Virtual)
**A Hierarchical Encoding-Decoding Scheme for Abstractive Multi-document Summarization**
*Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You and Lidong Bing*
Pre-trained language models (PLMs) have achieved outstanding achievements in abstractive single-document summarization (SDS). However, such benefits may not fully extend to multi-document summarization (MDS), where the handling of cross-document information is more complex. Previous works either design new MDS architectures or apply PLMs bluntly with concatenated source documents as a reformulated SDS task. While the former does not utilize previous pre-training efforts and may not generalize well across different domains, the latter may not sufficiently attend to the intricate cross-document relationships unique to MDS tasks. Instead, we enforce hierarchy on both the encoder and decoder to better utilize a PLM to facilitate multi-document interactions for the MDS task. Across 10 MDS benchmarks from various domains, our method outperforms or is competitive with the previous best models, including those with additional MDS pre-training or with more parameters. It outperforms its corresponding PLM backbone by up to 3 Rouge-L and is favored by humans.

16:00-17:30 (East Foyer / Virtual)

**Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models**
*Paul Youssef, Osman Alperen Koraş, Meijie Li, Jörg Schlötterer and Christin Seifert*
Pre-trained Language Models (PLMs) are trained on vast unlabeled data, rich in world knowledge. This fact has sparked the interest of the community in quantifying the amount of factual knowledge present in PLMs, as this explains their performance on downstream tasks, and potentially justifies their use as knowledge bases. In this work, we survey methods and datasets that are used to probe PLMs for factual knowledge. Our contributions are: (1) We propose a categorization scheme for factual probing methods that is based on how their inputs, outputs and the probed PLMs are adapted; (2) We provide an overview of the datasets used for factual probing; (3) We synthesize insights about knowledge retention and prompt optimization in PLMs, analyze obstacles to adopting PLMs as knowledge bases and outline directions for future work.

16:00-17:30 (East Foyer / Virtual)
**MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition**
*Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko and Shervin Malmasi*
We present MULTICONER V2, a dataset for fine-grained Named Entity Recognition covering 33 entity classes across 12 languages, in both monolingual and multilingual settings. This dataset aims to tackle the following practical challenges in NER: (i) effective handling of fine-grained classes that include complex entities like movie titles, and (ii) performance degradation due to noise generated from typing mistakes or OCR errors. The dataset is compiled from open resources like Wikipedia and Wikidata, and is publicly available. Evaluation based on the XLM-RoBERTa baseline highlights the unique challenges posed by MULTICONER V2: (i) the fine-grained taxonomy is challenging, where the scores are low with macro-F1=0.63 (across all languages), and (ii) the corruption strategy significantly impairs performance, with entity corruption resulting in 9% lower performance relative to non-entity corruptions across all languages. This highlights the greater impact of entity noise in contrast to context noise.

16:00-17:30 (East Foyer / Virtual)
**NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation**
*Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula and Yejin Choi*
We present NovaCOMET, an open commonsense knowledge model, that combines the best aspects of knowledge and general task models. Compared to previous knowledge models, NovaCOMET allows open-format relations enabling direct application to reasoning tasks; compared to general task models like Flan-T5, it explicitly centers knowledge, enabling superior performance for commonsense reasoning. NovaCOMET leverages the knowledge of opaque proprietary models to create an open knowledge pipeline. First, knowledge is symbolically distilled into NovATOMIC, a publicly-releaseddiscrete knowledge graph which can be audited, critiqued, and filtered. Next, we train NovaCOMET on NovATOMIC by fine-tuning an open-source pretrained model. NovaCOMET uses an open-format training objective, replacing the fixed relation sets of past knowledge models, enabling arbitrary structures within the data to serve as inputs or outputs. The resulting generation model, optionally augmented with human annotation, matches or exceeds comparable open task models like Flan-T5 on a range of commonsense generation tasks. NovaCOMET serves as a counterexample to the contemporary focus on instruction tuning only, demonstrating a distinct advantage to explicitly modeling commonsense knowledge as well.

16:00-17:30 (East Foyer / Virtual)
**HiCL: Hierarchical Contrastive Learning of Unsupervised Sentence Embeddings**
*Zhuofeng Wu, Chaowei Xiao and VG Vinod Vydiswaran*
In this paper, we propose a hierarchical contrastive learning framework, HiCL, which considers local segment-level and global sequence-level relationships to improve training efficiency and effectiveness. Traditional methods typically encode a sequence in its entirety for contrast with others, often neglecting local representation learning, leading to challenges in generalizing to shorter texts. Conversely, HiCL improves its effectiveness by dividing the sequence into several segments and employing both local and global contrastive learning to model segment-level and sequence-level relationships. Further, considering the quadratic time complexity of transformers over input tokens, HiCL boosts training efficiency by first encoding short segments and then aggregating them to obtain the sequence representation. Extensive experiments show that HiCL enhances the prior top-performing SNCSE model across seven extensively evaluated STS tasks, with an average increase of +0.2% observed on $BERT_{large}$ and +0.44% on $RoBERTa_{large}$.

16:00-17:30 (East Foyer / Virtual)
**K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings**
*Chaewon Park, Soohwan Kim, Kyubyong Park and Kunwoo Park*
Numerous datasets have been proposed to combat the spread of online hate. Despite these efforts, a majority of these resources are English-centric, primarily focusing on overt forms of hate. This research gap calls for developing high-quality corpora in diverse languages that also encapsulate more subtle hate expressions. This study introduces K-HATERS, a new corpus for hate speech detection in Korean, comprising approximately 192K news comments with target-specific offensiveness ratings. This resource is the largest offensive language corpus in Korean and is the first to offer target-specific ratings on a three-point Likert scale, enabling the detection of hate expressions in Korean across varying degrees of offensiveness. We conduct experiments showing the effectiveness of the proposed corpus, including a comparison with existing datasets. Additionally, to address potential noise and bias in human annotations, we explore a novel idea of adopting the Cognitive Reflection Test, which is widely used in social science for assessing an individual's cognitive ability, as a proxy of labeling quality. Findings indicate that annotations from individuals with the lowest test scores tend to yield detection models that make biased predictions toward specific target groups and are less accurate. This study contributes to the NLP research on hate speech detection and resource construction. The code and dataset can be accessed at https://github.com/ssu-humane/K-HATERS.

16:00-17:30 (East Foyer / Virtual)
**Linguistically Motivated Sign Language Segmentation**
*Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling and Yoav Goldberg*
Sign language segmentation is a crucial task in sign language processing systems. It enables downstream tasks such as sign recognition, transcription, and machine translation. In this work, we consider two kinds of segmentation: segmentation into individual signs and segmentation into *phrases*, larger units comprising several signs. We propose a novel approach to jointly model these two tasks. Our method is motivated by linguistic cues observed in sign language corpora. We replace the predominant IO tagging scheme with BIO tagging to account for continuous signing. Given that prosody plays a significant role in phrase boundaries, we explore the use of optical flow features. We also provide an extensive analysis of hand shapes and 3D hand normalization. We find that introducing BIO tagging is necessary to model sign boundaries. Explicitly encoding prosody by optical flow improves segmentation in shallow models, but its contribution is negligible in deeper models. Careful tuning of the decoding algorithm atop the models further improves the segmentation quality. We demonstrate that our final models generalize to out-of-domain video content in a different signed language, even under a zero-shot setting. We observe that including optical flow and 3D hand normalization enhances the robustness of the model in this context.

16:00-17:30 (East Foyer / Virtual)

**IntenDD: A Unified Contrastive Learning Approach for Intent Detection and Discovery**
*Bhavuk Singhal, Ashim Gupta, V P Shivasankaran and Amrith Krishna*
Identifying intents from dialogue utterances forms an integral component of task-oriented dialogue systems. Intent-related tasks are typically formulated either as a classification task, where the utterances are classified into predefined categories or as a clustering task when new and previously unknown intent categories need to be discovered from these utterances. Further, the intent classification may be modeled in a multiclass (MC) or multilabel (ML) setup. While typically these tasks are modeled as separate tasks, we propose IntenDD a unified approach leveraging a shared utterance encoding backbone. IntenDD uses an entirely unsupervised contrastive learning strategy for representation learning, where pseudo-labels for the unlabeled utterances are generated based on their lexical features. Additionally, we introduce a two-step post-processing setup for the classification tasks using modified adsorption. Here, first, the residuals in the training data are propagated followed by smoothing the labels both modeled in a transductive setting. Through extensive evaluations on various benchmark datasets, we find that our approach consistently outperforms competitive baselines across all three tasks. On average, IntenDD reports percentage improvements of 2.32 %, 1.26 %, and 1.52 % in their respective metrics for few-shot MC, few-shot ML, and the intent discovery tasks respectively.

16:00-17:30 (East Foyer / Virtual)
**Exploring the Impact of Corpus Diversity on Financial Pretrained Language Models**
*Jaeyoung Choe, Keonwoong Noh, Nayeon Kim, Seyun Ahn and Woohwan Jung*
Over the past few years, various domain-specific pretrained language models (PLMs) have been proposed and have outperformed general-domain PLMs in specialized areas such as biomedical, scientific, and clinical domains. In addition, financial PLMs have been studied because of the high economic impact of financial data analysis. However, we found that financial PLMs were not pretrained on sufficiently diverse financial data. This lack of diverse training data leads to a subpar generalization performance, resulting in general-purpose PLMs, including BERT, often outperforming financial PLMs on many downstream tasks. To address this issue, we collected a broad range of financial corpus and trained the Financial Language Model (FiLM) on these diverse datasets. Our experimental results confirm that FiLM outperforms not only existing financial PLMs but also general domain PLMs. Furthermore, we provide empirical evidence that this improvement can be achieved even for unseen corpus groups.

16:00-17:30 (East Foyer / Virtual)
**Topic-Informed Dialogue Summarization using Topic Distribution and Prompt-based Modeling**
*Jaeah You and Youngjoong Ko*
Dealing with multiple topics should be considered an important issue in dialogue summarization, because dialogues, unlike documents, are prone to topic drift. Thus, we propose a new dialogue summarization model that reflects dialogue topic distribution to consider all topics present in the dialogue. First, the distribution of dialogue topics is estimated by an effective topic discovery model. Then topic-informed prompt transfers estimated topic distribution information to the output of encoder and decoder vectors. Finally, the topic extractor estimates the summary topic distribution from the output context vector of decoder to distinguish its difference from the dialogue topic distribution. To consider the proportion of each topic distribution appeared in the dialogue, the extractor is trained to reduce the difference between the distributions of the dialogue and the summary. The experimental results on SAMSum and DialogSum show that our model outperforms state-of-the-art methods on ROUGE scores. The human evaluation results also show that our framework well generates comprehensive summaries.

16:00-17:30 (East Foyer / Virtual)
**Explore the Way: Exploring Reasoning Path by Bridging Entities for Effective Cross-Document Relation Extraction**
*Junyoung Son, Jinsung Kim, Jungwoo Lim, Yoonna Jang and Heuiseok Lim*
Cross-document relation extraction (CodRED) task aims to infer the relation between two entities mentioned in different documents within a reasoning path. Previous studies have concentrated on merely capturing implicit relations between the entities. However, humans usually utilize explicit information chains such as hyperlinks or additional searches to find the relations between two entities. Inspired by this, we propose Path wIth expLOraTion (PILOT) that provides the enhanced reasoning path by exploring the explicit clue information within the documents. PILOT finds the bridging entities which directly guide the paths between the entities and then employs them as stepstones to navigate desirable paths. We show that models with PILOT outperform the baselines in the CodRED task. Furthermore, we offer a variety of analyses to verify the validity of the reasoning paths constructed through PILOT, including evaluations using large language models such as ChatGPT.

16:00-17:30 (East Foyer / Virtual)
**Entity Disambiguation on a Tight Labeling Budget**
*Audi Primadhanty and Ariadna Quattoni*
Many real-world NLP applications face the challenge of training an entity disambiguation model for a specific domain with a small labeling budget. In this setting there is often access to a large unlabeled pool of documents. It is then natural to ask the question: which samples should be selected for annotation? In this paper we propose a solution that combines feature diversity with low rank correction. Our sampling strategy is formulated in the context of bilinear tensor models. Our experiments show that the proposed approach can significantly reduce the amount of labeled data necessary to achieve a given performance.

16:00-17:30 (East Foyer / Virtual)
**Unnatural language processing: How do language models handle machine-generated prompts?**
*Corentin Kervadec, Francesca Franzon and Marco Baroni*
Language model prompt optimization research has shown that semantically and grammatically well-formed manually crafted prompts are routinely outperformed by automatically generated token sequences with no apparent meaning or syntactic structure, including sequences of vectors from a model's embedding space. We use machine-generated prompts to probe how models respond to input that is not composed of natural language expressions. We study the behavior of models of different sizes in multiple semantic tasks in response to both continuous and discrete machine-generated prompts, and compare it to the behavior in response to human-generated natural-language prompts. Even when producing a similar output, machine-generated and human prompts trigger different response patterns through the network processing pathways, including different perplexities, different attention and output entropy distributions, and different unit activation profiles. We provide preliminary insight into the nature of the units activated by different prompt types, suggesting that only natural language prompts recruit a genuinely linguistic circuit.

16:00-17:30 (East Foyer / Virtual)
**RealBehavior: A Framework for Faithfully Characterizing Foundation Models' Human-like Behavior Mechanisms**
*Enyu Zhou, Rui Zheng, Zhiheng Xi, Songyang Gao, Xiaoran Fan, Zichu Fei, Jingting Ye, Tao Gui, Qi Zhang and Xuanjing Huang*
Reports of human-like behaviors in foundation models are growing, with psychological theories providing enduring tools to investigate these behaviors. However, current research tends to directly apply these human-oriented tools without verifying the faithfulness of their outcomes. In this paper, we introduce a framework, RealBehavior, which is designed to characterize the humanoid behaviors of models faithfully. Beyond simply measuring behaviors, our framework assesses the faithfulness of results based on reproducibility, internal and external consistency, and generalizability. Our findings suggest that a simple application of psychological tools cannot faithfully characterize all human-like behav-

iors. Moreover, we discuss the impacts of aligning models with human and social values, arguing for the necessity of diversifying alignment objectives to prevent the creation of models with restricted characteristics.

16:00-17:30 (East Foyer / Virtual)
**Multilingual Coarse Political Stance Classification of Media. The Editorial Line of a ChatGPT and Bard Newspaper**
*Cristina España-Bonet*
Neutrality is difficult to achieve and, in politics, subjective. Traditional media typically adopt an editorial line that can be used by their potential readers as an indicator of the media bias. Several platforms currently rate news outlets according to their political bias. The editorial line and the ratings help readers in gathering a balanced view of news. But in the advent of instruction-following language models, tasks such as writing a newspaper article can be delegated to computers. Without imposing a biased persona, where would an AI-based news outlet lie within the bias ratings? In this work, we use the ratings of authentic news outlets to create a multilingual corpus of news with coarse stance annotations (Left and Right) along with automatically extracted topic annotations. We show that classifiers trained on this data are able to identify the editorial line of most unseen newspapers in English, German, Spanish and Catalan. We then apply the classifiers to 101 newspaper-like articles written by ChatGPT and Bard in the 4 languages at different time periods. We observe that, similarly to traditional newspapers, ChatGPT editorial line evolves with time and, being a data-driven system, the stance of the generated articles differs among languages.

16:00-17:30 (East Foyer / Virtual)
**RWKV: Reinventing RNNs for the Transformer Era**
*Bo Peng, Eric Alcaide, Quentin Gregory Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Nguyen Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu and Rui-Jie Zhu*
Transformers have revolutionized almost all natural language processing (NLP) tasks but suffer from memory and computational complexity that scales quadratically with sequence length. In contrast, recurrent neural networks (RNNs) exhibit linear scaling in memory and computational requirements but struggle to match the same performance as Transformers due to limitations in parallelization and scalability. We propose a novel model architecture, Receptance Weighted Key Value (RWKV), that combines the efficient parallelizable training of transformers with the efficient inference of RNNs. Our approach leverages a linear attention mechanism and allows us to formulate the model as either a Transformer or an RNN, thus parallelizing computations during training and maintains constant computational and memory complexity during inference. We scale our models as large as 14 billion parameters, by far the largest dense RNN ever trained, and find RWKV performs on par with similarly sized Transformers, suggesting future work can leverage this architecture to create more efficient models. This work presents a significant step towards reconciling trade-offs between computational efficiency and model performance in sequence processing tasks.

16:00-17:30 (East Foyer / Virtual)
**Inverse Reinforcement Learning for Text Summarization**
*Yu Fu, Deyi Xiong and Yue Dong*
We introduce inverse reinforcement learning (IRL) as an effective paradigm for training abstractive summarization models, imitating human summarization behaviors. Our IRL model estimates the reward function using a suite of important sub-rewards for summarization and concurrently optimizes the policy network. Experimental results across datasets in different domains (CNN/DailyMail and WikiHow) and various model sizes (BART-base and BART-large) demonstrate the superiority of our proposed IRL model for summarization over MLE and RL baselines. The resulting summaries exhibit greater similarity to human-crafted gold references, outperforming MLE and RL baselines on metrics such as ROUGE, coverage, novelty, compression ratio, factuality, and human evaluations.

16:00-17:30 (East Foyer / Virtual)
**Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation**
*Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R. Tetreault and Alejandro Jaimes*
To explore how humans can best leverage LLMs for writing and how interacting with these models affects feelings of ownership and trust in the writing process, we compared common human-AI interaction types (e.g., guiding system, selecting from system outputs, post-editing outputs) in the context of LLM-assisted news headline generation. While LLMs alone can generate satisfactory news headlines, on average, human control is needed to fix undesirable model outputs. Of the interaction methods, guiding and selecting model output added the most benefit with the lowest cost (in time and effort). Further, AI assistance did not harm participants' perception of control compared to freeform editing.

16:00-17:30 (East Foyer / Virtual)
**RoAST: Robustifying Language Models via Adversarial Perturbation with Selective Training**
*Jaehyung Kim, Yuning Mao, Rui Hou, Hanchao Yu, Davis Liang, Pascale Fung, Qifan Wang, Fuli Feng, Lifu Huang and Madian Khabsa*
Fine-tuning pre-trained language models (LMs) has become the de facto standard in many NLP tasks. Nevertheless, fine-tuned LMs are still prone to robustness issues, such as adversarial robustness and model calibration. Several perspectives of robustness for LMs have been studied independently, but lacking a unified consideration in multiple perspectives. In this paper, we propose Robustifying LMs via Adversarial perturbation with Selective Training (RoAST), a simple yet effective fine-tuning technique to enhance the multi-perspective robustness of LMs in a unified way. RoAST effectively incorporates two important sources for the model robustness, robustness on the perturbed inputs and generalizable knowledge in pre-trained LMs. To be specific, RoAST introduces adversarial perturbation during fine-tuning while the model parameters are selectively updated upon their relative importance to minimize unnecessary deviation. Under a unified evaluation of fine-tuned LMs by incorporating four representative perspectives of model robustness, we demonstrate the effectiveness of RoAST compared to state-of-the-art fine-tuning methods on six different types of LMs, which indicates its usefulness in practice.

16:00-17:30 (East Foyer / Virtual)
**The Cost of Compression: Investigating the Impact of Compression on Parametric Knowledge in Language Models**
*Satya Sai Srinath Namburi, Makesh Narsimhan Sreedhar, Srinath Srinivasan and Frederic Sala*
Compressing large language models (LLMs), often consisting of billions of parameters, provides faster inference, smaller memory footprints, and enables local deployment. The standard compression techniques are pruning and quantization, with the former eliminating redundant connections in model layers and the latter representing model parameters with as little as 4 bits. The key tradeoff is between the degree of compression and the impact on the quality of the compressed model. Existing research on LLM compression primarily focuses on performance in terms of general metrics like perplexity or downstream task accuracy. More fine-grained metrics, such as those measuring parametric knowledge, remain significantly underexplored. To help bridge this gap, we present a comprehensive analysis across multiple model families using the LAMA and LM-Harness benchmarks in order to systematically quantify the effect of commonly employed compression techniques on model performance. A particular focus is on tradeoffs involving parametric knowledge, with the goal of providing practitioners with practical insights to make informed decisions on compression.

16:00-17:30 (East Foyer / Virtual)
**Finding Common Ground: Annotating and Predicting Common Ground in Spoken Conversations**
*Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Abolghasem Mirroshandel and Owen Rambow*
When we communicate with other humans, we do not simply generate a sequence of words. Rather, we use our cognitive state (beliefs, desires, intentions) and our model of the audience's cognitive state to create utterances that affect the audience's cognitive state in the intended manner. An important part of cognitive state is the common ground, which is the content the speaker believes, and the speaker believes the audience believes, and so on. While much attention has been paid to common ground in cognitive science, there has not been much work in natural language processing. In this paper, we introduce a new annotation and corpus to capture common ground. We then describe some initial experiments extracting propositions from dialog and tracking their status in the common ground from the perspective of each speaker.

16:00-17:30 (East Foyer / Virtual)
**Descriptive Prompt Paraphrasing for Target-Oriented Multimodal Sentiment Classification**
*Dan Liu, Lin Li, Xiaohui Tao, Jian Cui and Qing Xie*
Target-Oriented Multimodal Sentiment Classification (TMSC) aims to perform sentiment polarity on a target jointly considering its corresponding multiple modalities including text, image, and others. Current researches mainly work on either of two types of targets in a decentralized manner. One type is entity, such as a person name, a location name, etc. and the other is aspect, such as 'food', 'service', etc. We believe that this target type based division in task modelling is not necessary because the sentiment polarity of the specific target is not governed by its type but its context. For this reason, we propose a unified model for target-oriented multimodal sentiment classification, so called UnifiedTMSC. It is prompt-based language modelling and performs well on four datasets spanning the above two target types. Specifically, we design descriptive prompt paraphrasing to reformulate TMSC task via (1) task paraphrasing, which obtains paraphrased prompts based on the task description through a paraphrasing rule, and (2) image prefix tuning, which optimizes a small continuous image vector throughout the multimodal representation space of text and images. Conducted on two entity-level multimodal datasets: Twitter-2015 and Twitter-2017, and two aspect-level multimodal datasets: Multi-ZOL and MASAD, the experimental results show the effectiveness of our UnifiedTMSC.

16:00-17:30 (East Foyer / Virtual)
**Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models**
*Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen and Naihao Deng*
Theory of Mind (ToM) is the ability to reason about one's own and others' mental states. ToM plays a critical role in the development of intelligence, language understanding, and cognitive processes. While previous work has primarily focused on first and second-order ToM, we explore higher-order ToM, which involves recursive reasoning on others' beliefs. %We also incorporate a new deception mechanism in ToM reasoning. We introduce Hi-ToM, a Higher Order Theory of Mind benchmark. Our experimental evaluation using various Large Language Models (LLMs) indicates a decline in performance on higher-order ToM tasks, demonstrating the limitations of current LLMs. We conduct a thorough analysis of different failure cases of LLMs, and share our thoughts on the implications of our findings on the future of NLP.

16:00-17:30 (East Foyer / Virtual)
**Hierarchical Catalogue Generation for Literature Review: A Benchmark**
*Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu and Bing Qin*
Scientific literature review generation aims to extract and organize important information from an abundant collection of reference papers and produces corresponding reviews while lacking a clear and logical hierarchy. We observe that a high-quality catalogue-guided generation process can effectively alleviate this problem. Therefore, we present an atomic and challenging task named Hierarchical Catalogue Generation for Literature Review as the first step for review generation, which aims to produce a hierarchical catalogue of a review paper given various references. We construct a novel English Hierarchical Catalogues of Literature Reviews Dataset with 7.6k literature review catalogues and 389k reference papers. To accurately assess the model performance, we design two evaluation metrics for informativeness and similarity to ground truth from semantics and structure. Our extensive analyses verify the high quality of our dataset and the effectiveness of our evaluation metrics. We further benchmark diverse experiments on state-of-the-art summarization models like BART and large language models like ChatGPT to evaluate their capabilities. We further discuss potential directions for this task to motivate future research.

16:00-17:30 (East Foyer / Virtual)
**Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models**
*Haoran Wang and Kai Shu*
Claim verification plays a crucial role in combating misinformation. While existing works on claim verification have shown promising results, a crucial piece of the puzzle that remains unsolved is to understand how to verify claims without relying on human-annotated data, which is expensive to create at a large scale. Additionally, it is important for models to provide comprehensive explanations that can justify their decisions and assist human fact-checkers. This paper presents First-Order-Logic-Guided Knowledge-Grounded (FOLK) Reasoning that can verify complex claims and generate explanations without the need for annotated evidence using Large Language Models (LLMs). FOLK leverages the in-context learning ability of LLMs to translate the claim into a First-Order-Logic (FOL) clause consisting of predicates, each corresponding to a sub-claim that needs to be verified. Then, FOLK performs FOL-Guided reasoning over a set of knowledge-grounded question-and-answer pairs to make veracity predictions and generate explanations to justify its decision-making process. This process makes our model highly explanatory, providing clear explanations of its reasoning process in human-readable form. Our experiment results indicate that FOLK outperforms strong baselines on three datasets encompassing various claim verification challenges. Our code and data are available.

16:00-17:30 (East Foyer / Virtual)
**IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models**
*Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang and Shih-Fu Chang*
The field of vision-and-language (VL) understanding has made unprecedented progress with end-to-end large pre-trained VL models (VLMs). However, they still fall short in zero-shot reasoning tasks that require multi-step inferencing. To achieve this goal, previous works resort to a divide-and-conquer pipeline. In this paper, we argue that previous efforts have several inherent shortcomings: 1) They rely on domain-specific sub-question decomposing models. 2) They force models to predict the final answer even if the sub-questions or sub-answers provide insufficient information. We address these limitations via IdealGPT, a framework that iteratively decomposes VL reasoning using large language models (LLMs). Specifically, IdealGPT utilizes an LLM to generate sub-questions, a VLM to provide corresponding sub-answers, and another LLM to reason to achieve the final answer. These three modules perform the divide-and-conquer procedure iteratively until the model is confident about the final answer to the main question. We evaluate IdealGPT on multiple challenging VL reasoning tasks under a zero-shot setting. In particular, our IdealGPT outperforms the best existing GPT-4-like models by an absolute 10% on VCR and 15% on SNLI-VE. Code is available at https://github.com/Hxyou/IdealGPT.

16:00-17:30 (East Foyer / Virtual)
**Parameter Efficient Multi-task Fine-tuning by Learning to Transfer Token-wise Prompts**
*Muling Wu, Wenhao Liu, Jianhan Xu, Changze Lv, Zixuan Ling, Tianlong Li, Longtao Huang, Xiaoqing Zheng and Xuanjing Huang*

Prompt tuning has been proven to be successful on various tasks by incorporating a small number of trainable parameters while freezing large pre-trained language models (PLMs). However, it is still unsettled how to generate more proper prompts for any individual examples and how to extend prompt tuning to multi-task learning scenarios by leveraging cross-task features. To address these challenges, we propose a token-wise prompt tuning (TPT), in which a bank of finer-grained soft prompt tokens is built for multi-task learning by memory network. The tokens are retrieved from the bank against an input example and assembled to an instance-dependent prompt. Extensive experimental results on 14 datasets demonstrated that the models enhanced by our TPT performed far better than full parameter fine-tuned models and achieved state-of-the-art by tuning only $0.035\%$ parameters.

16:00-17:30 (East Foyer / Virtual)
**Extrapolating Multilingual Understanding Models as Multilingual Generators**
*Bohong Wu, Fei Yuan, Hai Zhao, Lei Li and Jingjing Xu*
Multilingual understanding models (or encoder-based), pre-trained via masked language modeling, have achieved promising results on many language understanding tasks (e.g., mBERT). However, these models are not capable of generating high-quality text compared with decoder-based causal language models. Can we transform a pre-trained language understanding model into an effective language generation model? We propose a Semantic-Guided Alignment-then-Denoising (SGA) approach to adapt a multilingual encoder to a multilingual generator with a small number of additional parameters. Experiments show that the proposed approach is an effective adaption method, outperforming widely-used initialization-based methods with gains of 9.4 BLEU on machine translation, 8.1 Rouge-L on question generation, and 5.5 METEOR on story generation on XLM-R$_{large}$. On the other hand, we observe that XLM-R is still inferior to mBART in supervised settings despite better results on zero-shot settings, indicating that more exploration is required to make understanding models strong generators. Our code is available at https://github.com/chengzhipanpan/XLMR4MT.

16:00-17:30 (East Foyer / Virtual)
**TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks**
*Shubhra Kanti Karmaker Santu and Dongji Feng*
While LLMs have shown great success in understanding and generating text in traditional conversational settings, their potential for performing ill-defined complex tasks is largely under-studied and yet to be benchmarked. However, conducting such benchmarking studies is challenging because of the large variations in LLMs' performance when different prompt types/styles are used and different degrees of detail are provided in the prompts. To address this issue, this paper proposes a general taxonomy that can be used to design prompts with specific properties in order to perform a wide range of complex tasks. This taxonomy will allow future benchmarking studies to report the specific categories of prompts used as part of the study, enabling meaningful comparisons across different studies. Also, by establishing a common standard through this taxonomy, researchers will be able to draw more accurate conclusions about LLMs' performance on a specific complex task.

16:00-17:30 (East Foyer / Virtual)
**HPE: Answering Complex Questions over Text by Hybrid Question Parsing and Execution**
*Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, Shafiq Joty and Yingbo Zhou*
The dominant paradigm of textual question answering systems is based on end-to-end neural networks, which excels at answering natural language questions but falls short on complex ones. This stands in contrast to the broad adaptation of semantic parsing approaches over structured data sources (e.g., relational database, knowledge graphs), that convert natural language questions to logical forms and execute them with query engines. Towards combining the strengths of neural and symbolic methods, we propose a framework of question parsing and execution on textual QA. It comprises two central pillars: (1) We parse the question of varying complexity into an intermediate representation, named H-expression, which is composed of simple questions as the primitives and symbolic operations representing the relationships among them; (2) To execute the resulting H-expressions, we design a hybrid executor, which integrates the deterministic rules to translate the symbolic operations with a drop-in neural reader network to answer each decomposed simple question. Hence, the proposed framework can be viewed as a top-down question parsing followed by a bottom-up answer backtracking. The resulting H-expressions closely guide the execution process, offering higher precision besides better interpretability while still preserving the advantages of the neural readers for resolving its primitive elements. Our extensive experiments on MuSiQue, 2WikiQA, HotpotQA, and NQ show that the proposed parsing and hybrid execution framework outperforms existing approaches in supervised, few-shot, and zero-shot settings, while also effectively exposing its underlying reasoning process.

16:00-17:30 (East Foyer / Virtual)
**End-to-End Autoregressive Retrieval via Bootstrapping for Smart Reply Systems**
*Benjamin Towle and Ke Zhou*
Reply suggestion systems represent a staple component of many instant messaging and email systems. However, the requirement to produce sets of replies, rather than individual replies, makes the task poorly suited for out-of-the-box retrieval architectures, which only consider individual message-reply similarity. As a result, these system often rely on additional post-processing modules to diversify the outputs. However, these approaches are ultimately bottlenecked by the performance of the initial retriever, which in practice struggles to present a sufficiently diverse range of options to the downstream diversification module, leading to the suggestions being less relevant to the user. In this paper, we consider a novel approach that radically simplifies this pipeline through an autoregressive text-to-text retrieval model, that learns the smart reply task end-to-end from a dataset of (message, reply set) pairs obtained via bootstrapping. Empirical results show this method consistently outperforms a range of state-of-the-art baselines across three datasets, corresponding to a 5.1%-17.9% improvement in relevance, and a 0.5%-63.1% improvement in diversity compared to the best baseline approach. We make our code publicly available.

## Industry 3

16:00-17:30 (East Foyer / Virtual)

---

16:00-17:30 (East Foyer / Virtual)
**Automatic Linking of Judgements to UK Supreme Court Hearings**
*Hadeel Saadany and Constantin Orasan*
One the most important archived legal material in the UK is the Supreme Court published judgements and video recordings of court sittings for the decided cases. The impact of Supreme Court published material extends far beyond the parties involved in any given case as it provides landmark rulings on arguable points of law of the greatest public and constitutional importance. However, the recordings of a case are usually very long which makes it both time and effort consuming for legal professionals to study the critical arguments in the legal deliberations. In this research, we summarise the second part of a combined research-industrial project for building an automated tool designed specifically to link segments in the text judgement to semantically relevant timespans in the videos of the hearings. The tool is employed as a User-Interface (UI) platform that provides a better access to justice by bookmarking the timespans in the videos which contributed to the final judgement of

the case. We explain how we employ AI generative technology to retrieve the relevant links and show that the customisation of the GPT text embeddings to our dataset achieves the best accuracy for our automatic linking system.

# Main Conference: Saturday, December 9, 2023

# Session 5: Oral & Poster - 09:00-10:30

## Multilinguality and Linguistic Diversity 1

09:00-10:30 (East Ballroom)

09:00-09:15 (East Ballroom)
**BasahaCorpus: An Expanded Linguistic Resource for Readability Assessment in Central Philippine Languages**
*Joseph Marvin Imperial and Ekaterina Kochmar*
Current research on automatic readability assessment (ARA) has focused on improving the performance of models in high-resource languages such as English. In this work, we introduce and release BasahaCorpus as part of an initiative aimed at expanding available corpora and baseline models for readability assessment in lower resource languages in the Philippines. We compiled a corpus of short fictional narratives written in Hiligaynon, Minasbate, Karay-a, and Rinconada—languages belonging to the Central Philippine family tree subgroup—to train ARA models using surface-level, syllable-pattern, and n-gram overlap features. We also propose a new hierarchical cross-lingual modeling approach that takes advantage of a language's placement in the family tree to increase the amount of available training data. Our study yields encouraging results that support previous work showcasing the efficacy of cross-lingual models in low-resource settings, as well as similarities in highly informative linguistic features for mutually intelligible languages.

09:15-09:30 (East Ballroom)
**Translating away Translationese without Parallel Data**
*Richa Jalota, Koel Dutta Chowdhury, Cristina España-Bonet and Josef van Genabith*
Translated texts exhibit systematic linguistic differences compared to original texts in the same language, and these differences are referred to as translationese. Translationese has effects on various cross-lingual natural language processing tasks, potentially leading to biased results. In this paper, we explore a novel approach to reduce translationese in translated texts: translation-based style transfer. As there are no parallel human-translated and original data in the same language, we use a self-supervised approach that can learn from comparable (rather than parallel) mono-lingual original and translated data. However, even this self-supervised approach requires some parallel data for validation. We show how we can eliminate the need for parallel validation data by combining the self-supervised loss with an unsupervised loss. This unsupervised loss leverages the original language model loss over the style-transferred output and a semantic similarity loss between the input and style-transferred output. We evaluate our approach in terms of original vs. translationese binary classification in addition to measuring content preservation and target-style fluency. The results show that our approach is able to reduce translationese classifier accuracy to a level of a random classifier after style transfer while adequately preserving the content and fluency in the target original style.

09:30-09:45 (East Ballroom)
**Automatic Transcription of Handwritten Old Occitan Language**
*Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann and Matthias ASSenmacher*
While existing neural network-based approaches have shown promising results in Handwritten Text Recognition (HTR) for high-resource languages and standardized/machine-written text, their application to low-resource languages often presents challenges, resulting in reduced effectiveness. In this paper, we propose an innovative HTR approach that leverages the Transformer architecture for recognizing handwritten Old Occitan language. Given the limited availability of data, which comprises only word pairs of graphical variants and lemmas, we develop and rely on elaborate data augmentation techniques for both text and image data. Our model combines a custom-trained Swin image encoder with a BERT text decoder, which we pre-train using a large-scale augmented synthetic data set and fine-tune on the small human-labeled data set. Experimental results reveal that our approach surpasses the performance of current state-of-the-art models for Old Occitan HTR, including open-source Transformer-based models such as a fine-tuned TrOCR and commercial applications like Google Cloud Vision. To nurture further research and development, we make our models, data sets, and code publicly available.

09:45-10:00 (East Ballroom)
**Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs**
*Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi and Grzegorz Kondrak*
Large language models (LLMs) have demonstrated exceptional natural language understanding abilities, and have excelled in a variety of natural language processing (NLP) tasks. Despite the fact that most LLMs are trained predominantly on English, multiple studies have demonstrated their capabilities in a variety of languages. However, fundamental questions persist regarding how LLMs acquire their multilingual abilities and how performance varies across different languages. These inquiries are crucial for the study of LLMs since users and researchers often come from diverse language backgrounds, potentially influencing how they use LLMs and interpret their output. In this work, we propose a systematic way of qualitatively and quantitatively evaluating the multilingual capabilities of LLMs. We investigate the phenomenon of cross-language generalization in LLMs, wherein limited multilingual training data leads to advanced multilingual capabilities. To accomplish this, we employ a novel prompt back-translation method. The results demonstrate that LLMs, such as GPT, can effectively transfer learned knowledge across different languages, yielding relatively consistent results in translation-equivariant tasks, in which the correct output does not depend on the language of the input. However, LLMs struggle to provide accurate results in translation-variant tasks, which lack this property, requiring careful user judgment to evaluate the answers.

10:00-10:15 (East Ballroom)
**Revisiting Machine Translation for Cross-lingual Classification**
*Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan and Luke Zettlemoyer*
Machine Translation (MT) has been widely used for cross-lingual classification, either by translating the test set into English and running inference with a monolingual model (translate-test), or translating the training set into the target languages and finetuning a multilingual model (translate-train). However, most research in the area focuses on the multilingual models rather than the MT component. We show that, by using a stronger MT system and mitigating the mismatch between training on original text and running inference on machine translated text, translate-test can do substantially better than previously assumed. The optimal approach, however, is highly task dependent, as we identify various sources of cross-lingual transfer gap that affect different tasks and approaches differently. Our work calls into question the dominance of multilingual models for cross-lingual classification, and prompts to pay more attention to MT-based baselines.

10:15-10:30 (East Ballroom)
**Language Representation Projection: Can We Transfer Factual Knowledge across Languages in Multilingual Language Models?**

*Shaoyang Xu, Junzhuo Li and Deyi Xiong*
Multilingual pretrained language models serve as repositories of multilingual factual knowledge. Nevertheless, a substantial performance gap of factual knowledge probing exists between high-resource languages and low-resource languages, suggesting limited implicit factual knowledge transfer across languages in multilingual pretrained language models. This paper investigates the feasibility of explicitly transferring relatively rich factual knowledge from English to non-English languages. To accomplish this, we propose two parameter-free **L**anguage **R**epresentation **P**rojection modules (LRP2). The first module converts non-English representations into English-like equivalents, while the second module reverts English-like representations back into representations of the corresponding non-English language. Experimental results on the mLAMA dataset demonstrate that LRP2 significantly improves factual knowledge retrieval accuracy and facilitates knowledge transferability across diverse non-English languages. We further investigate the working mechanism of LRP2 from the perspectives of representation space and cross-lingual knowledge neuron.

## Natural Language Generation 1

09:00-10:30 (Central 1 Ballroom)

09:00-09:15 (Central 1 Ballroom)
**Structure-aware Knowledge Graph-to-text Generation with Planning Selection and Similarity Distinction**
*Feng Zhao, Hongzhi Zou and Cheng Yan*
The knowledge graph-to-text (KG-to-text) generation task aims to synthesize coherent and engaging sentences that accurately convey the complex information derived from an input knowledge graph. One of the primary challenges in this task is bridging the gap between the diverse structures of the KG and the target text, while preserving the details of the input KG. To address this, we propose a novel approach that efficiently integrates graph structure-aware modules with pre-trained language models. Unlike conventional techniques, which only consider direct connections between first-order neighbors, our method delves deeper by incorporating Relative Distance Encoding as a bias within the graph structure-aware module. This enables our model to better capture the intricate topology information present in the KG. To further elevate the fidelity of the generated text, Planning Selection and Similarity Distinction are introduced. Our approach filters the most relevant linearized sequences by employing a planning scorer, while simultaneously distinguishing similar input KGs through contrastive learning techniques. Experiments on two datasets demonstrate the superiority of our model.

09:15-09:30 (Central 1 Ballroom)
**Granularity Matters: Pathological Graph-driven Cross-modal Alignment for Brain CT Report Generation**
*Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu and Ying Liu*
The automatic Brain CT reports generation can improve the efficiency and accuracy of diagnosing cranial diseases. However, current methods are limited by 1) coarse-grained supervision: the training data in image-text format lacks detailed supervision for recognizing subtle abnormalities, and 2) coupled cross-modal alignment: visual-textual alignment may be inevitably coupled in a coarse-grained manner, resulting in tangled feature representation for report generation. In this paper, we propose a novel Pathological Graph-driven Cross-modal Alignment (PGCA) model for accurate and robust Brain CT report generation. Our approach effectively decouples the cross-modal alignment by constructing a Pathological Graph to learn fine-grained visual cues and align them with textual words. This graph comprises heterogeneous nodes representing essential pathological attributes (i.e., tissue and lesion) connected by intra- and inter-attribute edges with prior domain knowledge. Through carefully designed graph embedding and updating modules, our model refines the visual features of subtle tissues and lesions and aligns them with textual words using contrastive learning. Extensive experimental results confirm the viability of our method. We believe that our PGCA model holds the potential to greatly enhance the automatic generation of Brain CT reports and ultimately contribute to improved cranial disease diagnosis.

09:30-09:45 (Central 1 Ballroom)
**Improving Image Captioning via Predicting Structured Concepts**
*Ting Wang, Weidong Chen, Yuanhe Tian, Yan Song and Zhendong Mao*
Having the difficulty of solving the semantic gap between images and texts for the image captioning task, conventional studies in this area paid some attention to treating semantic concepts as a bridge between the two modalities and improved captioning performance accordingly. Although promising results on concept prediction were obtained, the aforementioned studies normally ignore the relationship among concepts, which relies on not only objects in the image, but also word dependencies in the text, so that offers a considerable potential for improving the process of generating good descriptions. In this paper, we propose a structured concept predictor (SCP) to predict concepts and their structures, then we integrate them into captioning, so that enhance the contribution of visual signals in this task via concepts and further use their relations to distinguish cross-modal semantics for better description generation. Particularly, we design weighted graph convolutional networks (W-GCN) to depict concept relations driven by word dependencies, and then learns differentiated contributions from these concepts for following decoding process. Therefore, our approach captures potential relations among concepts and discriminatively learns different concepts, so that effectively facilitates image captioning with inherited information across modalities. Extensive experiments and their results demonstrate the effectiveness of our approach as well as each proposed module in this work.

09:45-10:00 (Central 1 Ballroom)
**CodeFusion: A Pre-trained Diffusion Model for Code Generation**
*Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Suzana Negreanu and Gust Verbruggen*
Imagine a developer who can only change their last line of code—how often would they have to start writing a function from scratch before it is correct? Auto-regressive models for code generation from natural language have a similar limitation: they do not easily allow reconsidering earlier tokens generated. We introduce CodeFusion, a pre-trained diffusion code generation model that addresses this limitation by iteratively denoising a complete program conditioned on the encoded natural language. We evaluate CodeFusion on the task of natural language to code generation for Bash, Python, and Microsoft Excel conditional formatting (CF) rules. Experiments show that CodeFusion (75M parameters) performs on par with state-of-the-art auto-regressive systems (350M-175B parameters) in top-1 accuracy and outperforms them in top-3 and top-5 accuracy due to its better balance in diversity versus quality.

10:00-10:15 (Central 1 Ballroom)
**Look-back Decoding for Open-Ended Text Generation**
*Nan Xu, Chunting Zhou, Asli Celikyilmaz and Xuezhe Ma*
Given a prefix (context), open-ended generation aims to decode texts that are coherent, which do not abruptly drift from previous topics, and informative, which do not suffer from undesired repetitions. In this paper, we propose Look-back, an improved decoding algorithm that leverages the Kullback–Leibler divergence to track the distribution distance between current and historical decoding steps. Thus Look-back can automatically predict potential repetitive phrase and topic drift, and remove tokens that may cause the failure modes, restricting the next token probability distribution within a plausible distance to the history. We perform decoding experiments on document continuation and

story generation, and demonstrate that Look-back is able to generate more fluent and coherent text, outperforming other strong decoding methods significantly in both automatic and human evaluations.

10:15-10:30 (Central 1 Ballroom)
**Measuring Attribution in Natural Language Generation Models**
*Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc and David Reitter*
Large neural models have brought a new challenge to natural language generation (NLG): It has become imperative to ensure the safety and reliability of the output of models that generate freely. To this end, we present an evaluation framework, Attributable to Identified Sources (AIS), stipulating that NLG output pertaining to the external world is to be verified against an independent, provided source. We define AIS and a two-stage annotation pipeline for allowing annotators to evaluate model output according to annotation guidelines. We successfully validate this approach on generation datasets spanning three tasks (two conversational QA datasets, a summarization dataset, and a table-to-text dataset). We provide full annotation guidelines in the appendices and publicly release the annotated data at https://github.com/google-research-datasets/AIS

## NLP Applications 1

09:00-10:30 (Central 3 Ballroom)

---

09:00-09:15 (Central 3 Ballroom)
**Learning Co-Speech Gesture for Multimodal Aphasia Type Detection**
*Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim and Jinyoung Han*
Aphasia, a language disorder resulting from brain damage, requires accurate identification of specific aphasia types, such as Broca's and Wernicke's aphasia, for effective treatment. However, little attention has been paid to developing methods to detect different types of aphasia. Recognizing the importance of analyzing co-speech gestures for distinguish aphasia types, we propose a multimodal graph neural network for aphasia type detection using speech and corresponding gesture patterns. By learning the correlation between the speech and gesture modalities for each aphasia type, our model can generate textual representations sensitive to gesture information, leading to accurate aphasia type detection. Extensive experiments demonstrate the superiority of our approach over existing methods, achieving state-of-the-art results (F1 84.2%). We also show that gesture features outperform acoustic features, highlighting the significance of gesture expression in detecting aphasia types. We provide the codes for reproducibility purposes.

09:15-09:30 (Central 3 Ballroom)
**ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets**
*Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop and Markus Leippold*
Public and private actors struggle to assess the vast amounts of information about sustainability commitments made by various institutions. To address this problem, we create a novel tool for automatically detecting corporate and national net zero and reduction targets in three steps. First, we introduce an expert-annotated data set with 3.5K text samples. Second, we train and release ClimateBERT-NetZero, a natural language classifier to detect whether a text contains a net zero or reduction target. Third, we showcase its analysis potential with two use cases: We first demonstrate how ClimateBERT-NetZero can be combined with conventional question-answering (Q&A) models to analyze the ambitions displayed in net zero and reduction targets. Furthermore, we employ the ClimateBERT-NetZero model on quarterly earning call transcripts and outline how communication patterns evolve over time. Our experiments demonstrate promising pathways for extracting and analyzing net zero and emission reduction targets at scale.

09:30-09:45 (Central 3 Ballroom)
**Advancements in Arabic Grammatical Error Detection and Correction: An Empirical Investigation**
*Bashar Alhafni, Go Inoue, Christian Khairallah and Nizar Habash*
Grammatical error correction (GEC) is a well-explored problem in English with many existing models and datasets. However, research on GEC in morphologically rich languages has been limited due to challenges such as data scarcity and language complexity. In this paper, we present the first results on Arabic GEC using two newly developed Transformer-based pretrained sequence-to-sequence models. We also define the task of multi-class Arabic grammatical error detection (GED) and present the first results on multi-class Arabic GED. We show that using GED information as auxiliary input in GEC models improves GEC performance across three datasets spanning different genres. Moreover, we also investigate the use of contextual morphological preprocessing in aiding GEC systems. Our models achieve SOTA results on two Arabic GEC shared task datasets and establish a strong benchmark on a recently created dataset. We make our code, data, and pretrained models publicly available.

09:45-10:00 (Central 3 Ballroom)
**Detection of Multiple Mental Disorders from Social Media with Two-Stream Psychiatric Experts**
*Siyuan Chen, Zhiling Zhang, Mengyue Wu and Kenny Q. Zhu*
Existing Mental Disease Detection (MDD) research largely studies the detection of a single disorder, overlooking the fact that mental diseases might occur in tandem. Many approaches are not backed by domain knowledge (e.g., psychiatric symptoms) and thus fail to produce interpretable results. To tackle these issues, we propose an MDD framework that is capable of learning the shared clues of all diseases, while also capturing the specificity of each single disease. The two-stream architecture which simultaneously processes text and symptom features can combine the strength of both modalities and offer knowledge-based explainability. Experiments on the detection of 7 diseases show that our model can boost detection performance by more than 10%, especially in relatively rare classes.

10:00-10:15 (Central 3 Ballroom)
**Event-Location Tracking in Narratives: A Case Study on Holocaust Testimonies**
*Eitan Wagner, Renana Keydar and Omri Abend*
This work focuses on the spatial dimension of narrative understanding and presents the task of event-location tracking in narrative texts. The task intends to extract the sequence of locations where the narrative is set through its progression. We present several architectures for the task that seeks to model the global structure of the sequence, with varying levels of context awareness. We compare these methods to several baselines, including the use of strong methods applied over narrow contexts. We also develop methods for the generation of location embeddings and show that learning to predict a sequence of continuous embeddings, rather than a string of locations, is advantageous in terms of performance. We focus on the test case of Holocaust survivor testimonies. We argue for the moral and historical importance of studying this dataset in computational means and that it provides a unique case of a large set of narratives with a relatively restricted set of location trajectories. Our results show that models that are aware of the larger context of the narrative can generate more accurate location chains. We

further corroborate the effectiveness of our methods by showing similar trends from experiments on an additional domain.

10:15-10:30 (Central 3 Ballroom)
**Learning the Visualness of Text Using Large Vision-Language Models**
*Gaurav Verma, Ryan A. Rossi, Christopher Tensmeyer, Jiuxiang Gu and Ani Nenkova*
Visual text evokes an image in a person's mind, while non-visual text fails to do so. A method to automatically detect visualness in text will enable text-to-image retrieval and generation models to augment text with relevant images. This is particularly challenging with long-form text as text-to-image generation and retrieval models are often triggered for text that is designed to be explicitly visual in nature, whereas long-form text could contain many non-visual sentences. To this end, we curate a dataset of 3,620 English sentences and their visualness scores provided by multiple human annotators. We also propose a fine-tuning strategy that adapts large vision-language models like CLIP by modifying the model's contrastive learning objective to map text identified as non-visual to a common NULL image while matching visual text to their corresponding images in the document. We evaluate the proposed approach on its ability to (i) classify visual and non-visual text accurately, and (ii) attend over words that are identified as visual in psycholinguistic studies. Empirical evaluation indicates that our approach performs better than several heuristics and baseline models for the proposed task. Furthermore, to highlight the importance of modeling the visualness of text, we conduct qualitative analyses of text-to-image generation systems like DALL-E.

## Theme Track: Large Language Models and the Future of NLP 1

09:00-10:30 (West 1 Ballroom)

---

09:00-09:15 (West 1 Ballroom)
**Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation**
*Jason S Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi and Dongwon Lee*
Recent ubiquity and disruptive impacts of large language models (LLMs) have raised concerns about their potential to be misused (*.i.e, generating large-scale harmful and misleading content*). To combat this emerging risk of LLMs, we propose a novel "***Fighting Fire with Fire***" (F3) strategy that harnesses modern LLMs' generative and emergent reasoning capabilities to counter human-written and LLM-generated disinformation. First, we leverage GPT-3.5-turbo to synthesize authentic and deceptive LLM-generated content through paraphrase-based and perturbation-based prefix-style prompts, respectively. Second, we apply zero-shot in-context semantic reasoning techniques with cloze-style prompts to discern genuine from deceptive posts and news articles. In our extensive experiments, we observe GPT-3.5-turbo's zero-shot superiority for both in-distribution and out-of-distribution datasets, where GPT-3.5-turbo consistently achieved accuracy at 68-72%, unlike the decline observed in previous customized and fine-tuned disinformation detectors. Our codebase and dataset are available at https://github.com/mickeymst/F3.

09:15-09:30 (West 1 Ballroom)
**Reasoning with Language Model is Planning with World Model**
*Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang and Zhiting Hu*
Large language models (LLMs) have shown remarkable reasoning capabilities, particularly with Chain-of-Thought-style prompts. However, LLMs can still struggle with problems that are easy for humans, such as generating action plans for executing tasks or performing complex math or logical reasoning. This is due to LLMs' absence of an internal world model for predicting world states (e.g., environment status, variable values) and simulating long-term action outcomes of actions. This prevents LLMs from performing deliberate planning akin to human brains, which involves exploring alternative reasoning paths, anticipating future states and rewards, and iteratively refining existing reasoning steps. To overcome the limitations, we propose a new LLM reasoning framework, Reasoning via Planning (RAP). RAP repurposes the LLM as both a world model and a reasoning agent, and incorporates a principled planning algorithm (based on Monte Carlo Tree Search) for strategic exploration in the vast reasoning space. During reasoning, the LLM (as agent) incrementally builds a reasoning tree under the guidance of the LLM (as world model) and task-specific rewards, properly balancing exploration v.s. exploitation to achieve a high-reward reasoning path efficiently. We apply RAP to a variety of challenging reasoning problems, such as plan generation, math reasoning, and logical inference. Empirical results demonstrate the superiority of RAP over various strong baselines, including CoT and least-to-most prompting with self-consistency, e.g., RAP on LLaMA-33B surpasses CoT on GPT-4 with 33% relative improvement in plan generation.

09:30-09:45 (West 1 Ballroom)
**Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks**
*Alon Jacovi, Avi Caciularu, Omer Goldman and Yoav Goldberg*
Data contamination has become prevalent and challenging with the rise of models pretrained on large automatically-crawled corpora. For closed models, the training data becomes a trade secret, and even for open models, it is not trivial to detect contamination. Strategies such as leaderboards with hidden answers, or using test data which is guaranteed to be unseen, are expensive and become fragile with time. Assuming that all relevant actors value clean test data and will cooperate to mitigate data contamination, what can be done? We propose three strategies that can make a difference: (1) Test data made public should be encrypted with a public key and licensed to disallow derivative distribution; (2) demand training exclusion controls from closed API holders, and protect your test data by refusing to evaluate without them; (3) avoid data which appears with its solution on the internet, and release the web-page context of internet-derived data along with the data. These strategies are practical and can be effective in preventing data contamination.

09:45-10:00 (West 1 Ballroom)
**Prompting is not a substitute for probability measurements in large language models**
*Jennifer Hu and Roger P. Levy*
Prompting is now a dominant method for evaluating the linguistic knowledge of large language models (LLMs). While other methods directly read out models' probability distributions over strings, prompting requires models to access this internal information by processing linguistic input, thereby implicitly testing a new type of emergent ability: metalinguistic judgment. In this study, we compare metalinguistic prompting and direct probability measurements as ways of measuring models' linguistic knowledge. Broadly, we find that LLMs' metalinguistic judgments are inferior to quantities directly derived from representations. Furthermore, consistency gets worse as the prompt query diverges from direct measurements of next-word probabilities. Our findings suggest that negative results relying on metalinguistic prompts cannot be taken as conclusive evidence that an LLM lacks a particular linguistic generalization. Our results also highlight the value that is lost with the move to closed APIs where access to probability distributions is limited.

10:00-10:15 (West 1 Ballroom)
**FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions**
*Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi and Maarten Sap*
Theory of mind (ToM) evaluations currently focus on testing models using passive narratives that inherently lack interactivity. We introduce

FANToM, a new benchmark designed to stress-test ToM within information-asymmetric conversational contexts via question answering. Our benchmark draws upon important theoretical requisites from psychology and necessary empirical considerations when evaluating large language models (LLMs). In particular, we formulate multiple types of questions that demand the same underlying reasoning to identify illusory or false sense of ToM capabilities in LLMs. We show that FANToM is challenging for state-of-the-art LLMs, which perform significantly worse than humans even with chain-of-thought reasoning or fine-tuning.

10:15-10:30 (West 1 Ballroom)
**LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models**
*Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang and Lili Qiu*
Large language models (LLMs) have been applied in various applications due to their astonishing capabilities. With advancements in technologies such as chain-of-thought (CoT) prompting and in-context learning (ICL), the prompts fed to LLMs are becoming increasingly lengthy, even exceeding tens of thousands of tokens. To accelerate model inference and reduce cost, this paper presents LLMLingua, a coarse-to-fine prompt compression method that involves a budget controller to maintain semantic integrity under high compression ratios, a token-level iterative compression algorithm to better model the interdependence between compressed contents, and an instruction tuning based method for distribution alignment between language models. We conduct experiments and analysis over four datasets from different scenarios, i.e., GSM8K, BBH, ShareGPT, and Arxiv-March23; showing that the proposed approach yields state-of-the-art performance and allows for up to 20x compression with little performance loss.

## Efficient Methods for NLP 2

09:00-10:30 (West 2 Ballroom)

---

09:00-09:15 (West 2 Ballroom)
**APrompt: Attention Prompt Tuning for Efficient Adaptation of Pre-trained Language Models**
*Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu and Dongfang Liu*
With the continuous growth of large language models, the process of fine-tuning these models for new tasks has become increasingly parameter-intensive. Prompt tuning, a method that involves tuning a small set of soft prompts, has emerged as an effective and efficient approach for adapting large pre-trained language models. However, most existing prompt tuning approaches only introduce prompts at the input layer, limiting their performance and leaving large rooms for improvement. In this work, we propose a novel Attention Prompt tuning method, namely APrompt, for efficient adaptation of pre-trained language models. We first demonstrate that existing prompt tuning can be considered as a special case of attention prompt tuning. We then formally introduce APrompt, which incorporates query, key, and value prompts into the attention layer to guide the attention computation during fine-tuning. Experimental results on the SuperGLUE benchmark consistently demonstrate that our proposed approach outperforms state-of-the-art baselines and full fine-tuning method with pre-trained models at different scales. In addition, a comprehensive set of ablation studies validate the effectiveness of the prompt design, as well as the efficiency of our approach.

09:15-09:30 (West 2 Ballroom)
**Parameter-Efficient Language Model Tuning with Active Learning in Low-Resource Settings**
*Josip Jukić and Jan Snajder*
Pre-trained language models (PLMs) have ignited a surge in demand for effective fine-tuning techniques, particularly in low-resource domains and languages. Active learning (AL), a set of algorithms designed to decrease labeling costs by minimizing label complexity, has shown promise in confronting the labeling bottleneck. In parallel, adapter modules designed for parameter-efficient fine-tuning (PEFT) have demonstrated notable potential in low-resource settings. However, the interplay between AL and adapter-based PEFT remains unexplored. We present an empirical study of PEFT behavior with AL in low-resource settings for text classification tasks. Our findings affirm the superiority of PEFT over full-fine tuning (FFT) in low-resource settings and demonstrate that this advantage persists in AL setups. We further examine the properties of PEFT and FFT through the lens of forgetting dynamics and instance-level representations, where we find that PEFT yields more stable representations of early and middle layers compared to FFT. Our research underscores the synergistic potential of AL and PEFT in low-resource settings, paving the way for advancements in efficient and effective fine-tuning.

09:30-09:45 (West 2 Ballroom)
**Merging Experts into One: Improving Computational Efficiency of Mixture of Experts**
*Shwai He, Run-Ze Fan, Liang Ding, Li Shen, Tianyi Zhou and Dacheng Tao*
Scaling the size of language models usually leads to remarkable advancements in NLP tasks. But it often comes with a price of growing computational cost. Although a sparse Mixture of Experts (MoE) can reduce the cost by activating a small subset of parameters (e.g., one expert) for each input, its computation escalates significantly if increasing the number of activated experts, limiting its practical utility. Can we retain the advantages of adding more experts without substantially increasing the computational costs? In this paper, we first demonstrate the superiority of selecting multiple experts and then propose a computation-efficient approach called **`Merging Experts into One`** (MEO), which reduces the computation cost to that of a single expert. Extensive experiments show that MEO significantly improves computational efficiency, e.g., FLOPS drops from 72.0G of vanilla MoE to 28.6G (MEO). Moreover, we propose a token-level attention block that further enhances the efficiency and performance of token-level MEO, e.g., 83.3% (MEO) vs. 82.6% (vanilla MoE) average score on the GLUE benchmark. Our code will be released upon acceptance. Code will be released at: `https://github.com/Shwai-He/MEO`.

09:45-10:00 (West 2 Ballroom)
**Selective Labeling: How to Radically Lower Data-Labeling Costs for Document Extraction Models**
*Yichao Zhou, James Bradley Wendt, Navneet Potti, Jing Xie and Sandeep Tata*
Building automatic extraction models for visually rich documents like invoices, receipts, bills, tax forms, etc. has received significant attention lately. A key bottleneck in developing extraction models for new document types is the cost of acquiring the several thousand high-quality labeled documents that are needed to train a model with acceptable accuracy. In this paper, we propose selective labeling as a solution to this problem. The key insight is to simplify the labeling task to provide "yes/no" labels for candidate extractions predicted by a model trained on partially labeled documents. We combine this with a custom active learning strategy to find the predictions that the model is most uncertain about. We show through experiments on document types drawn from 3 different domains that selective labeling can reduce the cost of acquiring labeled data by $10\times$ with a negligible loss in accuracy.

10:00-10:15 (West 2 Ballroom)
**Focus Your Attention (with Adaptive IIR Filters)**
*Shahar Lutati, Itamar Zimerman and Lior Wolf*

We present a new layer in which dynamic (i.e., input-dependent) Infinite Impulse Response (IIR) filters of order two are used to process the input sequence prior to applying conventional attention. The input is split into chunks, and the coefficients of these filters are determined based on previous chunks to maintain causality. Despite their relatively low order, the causal adaptive filters are shown to focus attention on the relevant sequence elements. The new layer is grounded in control theory, and is shown to generalize diagonal state-space layers. The layer performs on-par with state-of-the-art networks, with a fraction of their parameters and with time complexity that is sub-quadratic with input size. The obtained layer is favorable to layers such as Heyna, GPT2, and Mega, both with respect to the number of parameters and the obtained level of performance on multiple long-range sequence problems.

10:15-10:30 (West 2 Ballroom)
**Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs**
*Pranjal Aggarwal, Aman Madaan, Yiming Yang and Mausam*
A popular approach for improving the correctness of output from large language models (LLMs) is Self-Consistency - poll the LLM multiple times and output the most frequent solution. Existing Self-Consistency techniques always generate a constant number of samples per question, where a better approach will be to non-uniformly distribute the available budget based on the amount of agreement in the samples generated so far. In response, we introduce Adaptive-Consistency, a cost-efficient, model-agnostic technique that dynamically adjusts the number of samples per question using a lightweight stopping criterion. Our experiments over 17 reasoning and code generation datasets and three LLMs demonstrate that Adaptive-Consistency reduces sample budget by up to 7.9 times with an average accuracy drop of less than 0.1%

## Human-Centered NLP

09:00-10:30 (West 3 Ballroom)

---

09:00-09:15 (West 3 Ballroom)
**Modeling Empathic Similarity in Personal Narratives**
*Jocelyn J Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Won Park and Cynthia Breazeal*
The most meaningful connections between people are often fostered through expression of shared vulnerability and emotional experiences in personal narratives. We introduce a new task of identifying similarity in personal stories based on empathic resonance, i.e., the extent to which two people empathize with each others' experiences, as opposed to raw semantic or lexical similarity, as has predominantly been studied in NLP. Using insights from social psychology, we craft a framework that operationalizes empathic similarity in terms of three key features of stories: main events, emotional trajectories, and overall morals or takeaways. We create EmpathicStories, a dataset of 1,500 personal stories annotated with our empathic similarity features, and 2,000 pairs of stories annotated with empathic similarity scores. Using our dataset, we fine-tune a model to compute empathic similarity of story pairs, and show that this outperforms semantic similarity models on automated correlation and retrieval metrics. Through a user study with 150 participants, we also assess the effect our model has on retrieving stories that users empathize with, compared to naive semantic similarity-based retrieval, and find that participants empathized significantly more with stories retrieved by our model. Our work has strong implications for the use of empathy-aware models to foster human connection and empathy between people.

09:15-09:30 (West 3 Ballroom)
**A Diachronic Perspective on User Trust in AI under Uncertainty**
*Shehzaad Zuzar Dhuliawala, Vilém Zouhar, Mennatallah El-Assady and Mrinmaya Sachan*
In human-AI collaboration, users typically form a mental model of the AI system, which captures the user's beliefs about when the system performs well and when it does not. The construction of this mental model is guided by both the system's veracity as well as the system output presented to the user e.g., the system's confidence and an explanation for the prediction. However, modern NLP systems are seldom calibrated and are often confidently incorrect about their predictions, which violates users' mental model and erodes their trust. In this work, we design a study where users bet on the correctness of an NLP system, and use it to study the evolution of user trust as a response to these trust-eroding events and how the user trust is rebuilt as a function of time after these events. We find that even a few highly inaccurate confidence estimation instances are enough to damage users' trust in the system and performance, which does not easily recover over time. We further find that users are more forgiving to the NLP system if it is unconfidently correct rather than confidently incorrect, even though, from a game-theoretic perspective, their payoff is equivalent. Finally, we find that each user can entertain multiple mental models of the system based on the type of the question. These results highlight the importance of confidence calibration in developing user-centered NLP applications to avoid damaging user trust and compromising the collaboration performance.

09:30-09:45 (West 3 Ballroom)
**Dr ChatGPT tell me what I want to hear: How different prompts impact health answer correctness**
*Bevan Koopman and Guido Zuccon*
This paper investigates the significant impact different prompts have on the behaviour of ChatGPT when used for health information seeking. As people more and more depend on generative large language models (LLMs) like ChatGPT, it is critical to understand model behaviour under different conditions, especially for domains where incorrect answers can have serious consequences such as health. Using the TREC Misinformation dataset, we empirically evaluate ChatGPT to show not just its effectiveness but reveal that knowledge passed in the prompt can bias the model to the detriment of answer correctness. We show this occurs both for retrieve-then-generate pipelines and based on how a user phrases their question as well as the question type. This work has important implications for the development of more robust and transparent question-answering systems based on generative large language models. Prompts, raw result files and manual analysis are made publicly available at https://github.com/ielab/drchatgpt-health_prompting.

09:45-10:00 (West 3 Ballroom)
**Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding**
*Shuwen Deng, Paul Prasse, David Robert Reich, Tobias Scheffer and Lena Ann Jäger*
Human gaze data offer cognitive information that reflects natural language comprehension. Indeed, augmenting language models with human scanpaths has proven beneficial for a range of NLP tasks, including language understanding. However, the applicability of this approach is hampered because the abundance of text corpora is contrasted by a scarcity of gaze data. Although models for the generation of human-like scanpaths during reading have been developed, the potential of synthetic gaze data across NLP tasks remains largely unexplored. We develop a model that integrates synthetic scanpath generation with a scanpath-augmented language model, eliminating the need for human gaze data. Since the model's error gradient can be propagated throughout all parts of the model, the scanpath generator can be fine-tuned to downstream tasks. We find that the proposed model not only outperforms the underlying language model, but achieves a performance that is comparable to a language model augmented with real human gaze data. Our code is publicly available.

10:00-10:15 (West 3 Ballroom)

**Generating and Evaluating Tests for K-12 Students with Language Model Simulations: A Case Study on Sentence Reading Efficiency**
*Eric Zelikman, Wanjing Anya Ma, Jasmine Elizabeth Tran, Diyi Yang, Jason D Yeatman and Nick Haber*
Developing an educational test can be expensive and time-consuming, as each item must be written by experts and then evaluated by collecting hundreds of student responses. Moreover, many tests require multiple distinct sets of questions administered throughout the school year to closely monitor students' progress, known as parallel tests. In this study, we focus on tests of silent sentence reading efficiency, used to assess students' reading ability over time. To generate high-quality parallel tests, we propose to fine-tune large language models (LLMs) to simulate how previous students would have responded to unseen items. With these simulated responses, we can estimate each item's difficulty and ambiguity. We first use GPT-4 to generate new test items following a list of expert-developed rules and then apply a fine-tuned LLM to filter the items based on criteria from psychological measurements. We also propose an optimal-transport-inspired technique for generating parallel tests and show the generated tests closely correspond to the original test's difficulty and reliability based on crowdworker responses. Our evaluation of a generated test with 234 students from grades 2 to 8 produces test scores highly correlated (r=0.93) to those of a standard test form written by human experts and evaluated across thousands of K-12 students.

10:15-10:30 (West 3 Ballroom)
**"Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text**
*Rishav Hada, Agrima Seth, Harshita Diddee and Kalika Bali*
Language serves as a powerful tool for the manifestation of societal belief systems. In doing so, it also perpetuates the prevalent biases in our society. Gender bias is one of the most pervasive biases in our society and is seen in online and offline discourses. With LLMs increasingly gaining human-like fluency in text generation, gaining a nuanced understanding of the biases these systems can generate is imperative. Prior work often treats gender bias as a binary classification task. However, acknowledging that bias must be perceived at a relative scale; we investigate the generation and consequent receptivity of manual annotators to bias of varying degrees. Specifically, we create the first dataset of GPT-generated English text with normative ratings of gender bias. Ratings were obtained using Best–Worst Scaling – an efficient comparative annotation framework. Next, we systematically analyze the variation of themes of gender biases in the observed ranking and show that identity-attack is most closely related to gender bias. Finally, we show the performance of existing automated models trained on related concepts on our dataset.

## Demo session 4

09:00-10:30 (East Foyer / Virtual)

09:00-10:30 (East Foyer / Virtual)
**LM-Polygraph: Uncertainty Estimation for Language Models**
*Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin and Artem Shelmanov*
Recent advancements in the capabilities of large language models (LLMs) have paved the way for a myriad of groundbreaking applications in various fields. However, a significant challenge arises as these models often "hallucinate", i.e., fabricate facts without providing users an apparent means to discern the veracity of their statements. Uncertainty estimation (UE) methods are one path to safer, more responsible, and more effective use of LLMs. However, to date, research on UE methods for LLMs has been focused primarily on theoretical rather than engineering contributions. In this work, we tackle this issue by introducing LM-Polygraph, a framework with implementations of a battery of state-of-the-art UE methods for LLMs in text generation tasks, with unified program interfaces in Python. Additionally, it introduces an extendable benchmark for consistent evaluation of UE techniques by researchers, and a demo web application that enriches the standard chat dialog with confidence scores, empowering end-users to discern unreliable responses. LM-Polygraph is compatible with the most recent LLMs, including BLOOMz, LLaMA-2, ChatGPT, and GPT-4, and is designed to support future releases of similarly-styled LMs.

09:00-10:30 (East Foyer / Virtual)
**Prompterator: Iterate Efficiently towards More Effective Prompts**
*Samuel Sučík, Daniel Skala, Andrej Švec, Peter Hraška and Marek Šuppa*
With the advent of Large Language Models (LLMs) the process known as prompting, which entices the LLM to solve an arbitrary language processing task without the need for finetuning, has risen to prominence. Finding well-performing prompts, however, is a non-trivial task which requires experimentation in order to arrive at a prompt that solves a specific task. When a given task does not readily reduce to one that can be easily measured with well established metrics, human evaluation of the results obtained by prompting is often necessary. In this work we present prompterator, a tool that helps the user interactively iterate over various potential prompts and choose the best performing one based on human feedback. It is distributed as an open source package with out-of-the-box support for various LLM providers and was designed to be easily extensible.

09:00-10:30 (East Foyer / Virtual)
**PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents**
*Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chee Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey and Luca Soldaini*
Despite growing interest in applying natural language processing (NLP) and computer vision (CV) models to the scholarly domain, scientific documents remain challenging to work with. They're often in difficult-to-use PDF formats, and the ecosystem of models to process them is fragmented and incomplete. We introduce PaperMage, an open-source Python toolkit for analyzing and processing visually-rich, structured scientific documents. PaperMage offers clean and intuitive abstractions for seamlessly representing and manipulating both textual and visual document elements. PaperMage achieves this by integrating disparate state-of-the-art NLP and CV models into a unified framework, and provides turn-key recipes for common scientific document processing use-cases. PaperMage has powered multiple research prototypes of AI applications over scientific documents, along with Semantic Scholar's large-scale production system for processing millions of PDFs. GitHub: https://github.com/allenai/papermage

09:00-10:30 (East Foyer / Virtual)
**OmniEvent: A Comprehensive, Fair, and Easy-to-Use Toolkit for Event Understanding**
*Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou and Juanzi Li*
Event understanding aims at understanding the content and relationship of events within texts, which covers multiple complicated information extraction tasks: event detection, event argument extraction, and event relation extraction. To facilitate related research and application, we present an event understanding toolkit OmniEvent, which features three desiderata: (1) Comprehensive. OmniEvent supports mainstream modeling paradigms of all the event understanding tasks and the processing of 15 widely-used English and Chinese datasets. (2) Fair. OmniEvent carefully handles the inconspicuous evaluation pitfalls reported in Peng et al. (2023), which ensures fair comparisons between different models. (3) Easy-to-use. OmniEvent is designed to be easily used by users with varying needs. We provide off-the-shelf models that

can be directly deployed as web services. The modular framework also enables users to easily implement and evaluate new event understanding models with OmniEvent. The toolkit is publicly released along with the demonstration website and video.

## Poster session 4

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**#1 SeqXGPT: Sentence-Level AI-Generated Text Detection**
*Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang and Xipeng Qiu*
Widely applied large language models (LLMs) can generate human-like content, raising concerns about the abuse of LLMs. Therefore, it is important to build strong AI-generated text (AIGT) detectors. Current works only consider document-level AIGT detection, therefore, in this paper, we first introduce a sentence-level detection challenge by synthesizing a dataset that contains documents that are polished with LLMs, that is, the documents contain sentences written by humans and sentences modified by LLMs. Then we propose **Seq**uence **X** (Check) **GPT**, a novel method that utilizes log probability lists from white-box LLMs as features for sentence-level AIGT detection. These features are composed like *waves* in speech processing and cannot be studied by LLMs. Therefore, we build SeqXGPT based on convolution and self-attention networks. We test it in both sentence and document-level detection challenges. Experimental results show that previous methods struggle in solving sentence-level AIGT detection, while our method not only significantly surpasses baseline methods in both sentence and document-level detection challenges but also exhibits strong generalization capabilities.

09:00-10:30 (East Foyer / Virtual)
**#2 Character-LLM: A Trainable Agent for Role-Playing**
*Yunfan Shao, Linyang Li, Junqi Dai and Xipeng Qiu*
Large language models (LLMs) can be used to serve as agents to simulate human behaviors, given the powerful ability to understand human instructions and provide high-quality generated texts. Such ability stimulates us to wonder whether LLMs can simulate a person in a higher form than simple human behaviors. Therefore, we aim to train an agent with the profile, experience, and emotional states of a specific person instead of using limited prompts to instruct ChatGPT API. In this work, we introduce Character-LLM that teach LLMs to act as specific people such as Beethoven, Queen Cleopatra, Julius Caesar, etc. Our method focuses on editing profiles as experiences of a certain character and training models to be personal simulacra with these experiences. To assess the effectiveness of our approach, we build a test playground that interviews trained agents and evaluates whether the agents *memorize* their characters and experiences. Experimental results show interesting observations that help build future simulacra of humankind.

09:00-10:30 (East Foyer / Virtual)
**#3 Human Raters Cannot Distinguish English Translations from Original English Texts**
*Shira Wein*
The term translationese describes the set of linguistic features unique to translated texts, which appear regardless of translation quality. Though automatic classifiers designed to distinguish translated texts achieve high accuracy and prior work has identified common hallmarks of translationese, human accuracy of identifying translated text is understudied. In this work, we perform a human evaluation of English original/translated texts in order to explore raters' ability to classify texts as being original or translated English and the features that lead a rater to judge text as being translated. Ultimately, we find that, regardless of the annotators' native language or the source language of the text, annotators are unable to distinguish translations from original English texts and also have low agreement. Our results provide critical insight into work in translation studies and context for assessments of translationese classifiers.

09:00-10:30 (East Foyer / Virtual)
**#4 SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models**
*Potsawee Manakul, Adian Liusie and Mark Gales*
Generative Large Language Models (LLMs) such as GPT-3 are capable of generating highly fluent responses to a wide variety of user prompts. However, LLMs are known to hallucinate facts and make non-factual statements which can undermine trust in their output. Existing fact-checking approaches either require access to the output probability distribution (which may not be available for systems such as ChatGPT) or external databases that are interfaced via separate, often complex, modules. In this work, we propose "SelfCheckGPT", a simple sampling-based approach that can be used to fact-check the responses of black-box models in a zero-resource fashion, i.e. without an external database. SelfCheckGPT leverages the simple idea that if an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts. However, for hallucinated facts, stochastically sampled responses are likely to diverge and contradict one another. We investigate this approach by using GPT-3 to generate passages about individuals from the WikiBio dataset, and manually annotate the factuality of the generated passages. We demonstrate that SelfCheckGPT can: i) detect non-factual and factual sentences; and ii) rank passages in terms of factuality. We compare our approach to several baselines and show that our approach has considerably higher AUC-PR scores in sentence-level hallucination detection and higher correlation scores in passage-level factuality assessment compared to grey-box methods.

09:00-10:30 (East Foyer / Virtual)
**#5 Large Language Models are Temporal and Causal Reasoners for Video Question Answering**
*Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh and Hyunwoo J. Kim*
Large Language Models (LLMs) have shown remarkable performances on a wide range of natural language understanding and generation tasks. We observe that the LLMs provide effective priors in exploiting *linguistic shortcuts* for temporal and causal reasoning in Video Question Answering (VideoQA). However, such priors often cause suboptimal results on VideoQA by leading the model to over-rely on questions, *i.e.*, *linguistic bias*, while ignoring visual content. This is also known as 'ungrounded guesses' or 'hallucinations'. To address this problem while leveraging LLMs' prior on VideoQA, we propose a novel framework, Flipped-VQA, encouraging the model to predict all the combinations of ⟨V, Q, A⟩ triplet by flipping the source pair and the target label to understand their complex relationships, *i.e.*, predict A, Q, and V given a VQ, VA, and QA pairs, respectively. In this paper, we develop LLaMA-VQA by applying Flipped-VQA to LLaMA, and it outperforms both LLMs-based and non-LLMs-based models on five challenging VideoQA benchmarks. Furthermore, our Flipped-VQA is a general framework that is applicable to various LLMs (OPT and GPT-J) and consistently improves their performances. We empirically demonstrate that Flipped-VQA not only enhances the exploitation of linguistic shortcuts but also mitigates the linguistic bias, which causes incorrect answers over-relying on the question. Code is available at https://github.com/mlvlab/Flipped-VQA.

09:00-10:30 (East Foyer / Virtual)
**#6 Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations**
*Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu and Ming Yin*
The collection and curation of high-quality training data is crucial for developing text classification models with superior performance, but

it is often associated with significant costs and time investment. Researchers have recently explored using large language models (LLMs) to generate synthetic datasets as an alternative approach. However, the effectiveness of the LLM-generated synthetic data in supporting model training is inconsistent across different classification tasks. To better understand factors that moderate the effectiveness of the LLM-generated synthetic data, in this study, we look into how the performance of models trained on these synthetic data may vary with the *subjectivity* of classification. Our results indicate that subjectivity, at both the task level and instance level, is negatively associated with the performance of the model trained on synthetic data. We conclude by discussing the implications of our work on the potential and limitations of leveraging LLM for synthetic data generation.

09:00-10:30 (East Foyer / Virtual)
**#7 Skill-Based Few-Shot Selection for In-Context Learning**
*Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen and Jian-Guang Lou*
*In-context learning* is the paradigm that adapts large language models to downstream tasks by providing a few examples. *Few-shot selection*—selecting appropriate examples for each test instance separately—is important for in-context learning. In this paper, we propose **Skill-KNN**, a skill-based few-shot selection method for in-context learning. The key advantages of Skill-KNN include: (1) it addresses the problem that existing methods based on pre-trained embeddings can be easily biased by surface natural language features that are not important for the target task; (2) it does not require training or fine-tuning of any models, making it suitable for frequently expanding or changing example banks. The key insight is to optimize the inputs fed into the embedding model, rather than tuning the model itself. Technically, Skill-KNN generates the skill-based descriptions for each test case and candidate example by utilizing a pre-processing few-shot prompting, thus eliminating unimportant surface features. Experimental results across five cross-domain semantic parsing datasets and six backbone models show that Skill-KNN significantly outperforms existing methods.

09:00-10:30 (East Foyer / Virtual)
**#8 UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation**
*Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng and Qi Zhang*
Large Language Models (LLMs) are popular for their impressive abilities, but the need for model-specific fine-tuning or task-specific prompt engineering can hinder their generalization. We propose UPRISE (Universal Prompt Retrieval for Improving zero-Shot Evaluation), which tunes a lightweight and versatile retriever that automatically retrieves prompts for a given zero-shot task input. Specifically, we demonstrate universality in a cross-task and cross-model scenario: the retriever is tuned on diverse tasks, but tested on unseen task types; we use a small frozen LLM, GPT-Neo-2.7B, for tuning the retriever, but test the retriever on different LLMs of much larger scales, such as BLOOM-7.1B, OPT-66B and GPT3-175B. Additionally, we show that UPRISE mitigates the hallucination problem in our experiments with ChatGPT, suggesting its potential to improve even the strongest LLMs. Our model and code are available at https://github.com/microsoft/LMOps.

09:00-10:30 (East Foyer / Virtual)
**#9 MoT: Memory-of-Thought Enables ChatGPT to Self-Improve**
*Xiaonan Li and Xipeng Qiu*
Large Language Models (LLMs) have shown impressive abilities on various tasks. However, fundamentally improving them depends on high-quality datasets or computationally expensive fine-tuning. On the contrary, humans can easily improve themselves by self-thinking and memory, without external resources. In this paper, we propose a framework, **MoT**, to let the LLM self-improve through **M**emory **o**f **T**houghts, without annotated datasets and parameter updates. Specifically, MoT is divided into two stages: 1. before the test stage, the LLM pre-thinks on the unlabeled dataset and saves the high-confidence thoughts as external memory; 2. During the test stage, given a test question, the LLM recalls relevant memory to help itself reason and answer it. Experimental results show that MoT can help ChatGPT significantly improve its abilities in arithmetic reasoning, commonsense reasoning, factual reasoning, and natural language inference. Further analyses show that each component contributes critically to the improvements and MoT can lead to consistent improvements across various CoT methods and LLMs.

09:00-10:30 (East Foyer / Virtual)
**#10 How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances**
*Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad and Jun Wang*
Although large language models (LLMs) are impressive in solving various tasks, they can quickly be outdated after deployment. Maintaining their up-to-date status is a pressing concern in the current era. This paper provides a comprehensive review of recent advances in aligning deployed LLMs with the ever-changing world knowledge. We categorize research works systemically and provide in-depth comparisons and discussions. We also discuss existing challenges and highlight future directions to facilitate research in this field.

09:00-10:30 (East Foyer / Virtual)
**#11 Explore-Instruct: Enhancing Domain-Specific Instruction Coverage through Active Exploration**
*Fanqi Wan, Xinting Huang, Tao Yang, Xiaojun Quan, Wei Bi and Shuming Shi*
Instruction-tuning can be substantially optimized through enhanced diversity, resulting in models capable of handling a broader spectrum of tasks. However, existing data employed for such tuning often exhibit an inadequate coverage of individual domains, limiting the scope for nuanced comprehension and interactions within these areas. To address this deficiency, we propose Explore-Instruct, a novel approach to enhance the data coverage to be used in domain-specific instruction-tuning through active exploration via Large Language Models (LLMs). Built upon representative domain use cases, Explore-Instruct explores a multitude of variations or possibilities by implementing a search algorithm to obtain diversified and domain-focused instruction-tuning data. Our data-centric analysis validates the effectiveness of this proposed approach in improving domain-specific instruction coverage. Moreover, our model's performance demonstrates considerable advancements over multiple baselines, including those utilizing domain-specific data enhancement. Our findings offer a promising opportunity to improve instruction coverage, especially in domain-specific contexts, thereby advancing the development of adaptable language models. Our code, model weights, and data are public at `https://github.com/fanqiwan/Explore-Instruct`.

09:00-10:30 (East Foyer / Virtual)
**#12 On the Benefits of Learning to Route in Mixture-of-Experts Models**
*Nishanth Dikkala, Nikhil Ghosh, Raghu Meka, Rina Panigrahy, Nikhil Vyas and Xin Wang*
Mixture-of-Expert (MoE) Transformer models, such as the Switch Transformer, allow us to successfully scale up model sizes while keeping the amount of compute time fixed. Prior work has established the computational efficiency benefits of using these models. A core component of these models is a router that routes input tokens to different experts in a layer. We show theoretical and empirical evidence that the router's ability to route tokens intelligently confers a significant advantage to MoE models. We study synthetic settings where the input data is distributed in clusters and show theoretically and empirically that the router learns to route the inputs according to these clusters. Then we perform experiments on real data using the T5X library, where we observe that a trainable router confers a non-trivial benefit instead of a non-trainable router.

09:00-10:30 (East Foyer / Virtual)

**#13 EasyQuant: An Efficient Data-free Quantization Algorithm for LLMs**
*Hanlin Tang, Yifu Sun, Decheng Wu, Kai Liu, Jianchen Zhu and Zhanhui Kang*
Large language models (LLMs) have proven to be very superior to conventional methods in various tasks. However, their expensive computations and high memory requirements are prohibitive for deployment. Model quantization is an effective method for reducing this overhead. The problem is that in most previous works, the quantized model was calibrated using few samples from the training data, which might affect the generalization of the quantized LLMs to unknown cases and tasks. Hence in this work, we explore an important question: Can we design a data-independent quantization method for LLMs to guarantee its generalization performance? In this work, we propose EasyQuant, a training-free and data-independent weight-only quantization algorithm for LLMs. Our observation indicates that two factors: outliers in the weight and quantization ranges, are essential for reducing the quantization error. Therefore, in EasyQuant, we leave the outliers (less than 1%) unchanged and optimize the quantization range to reduce the reconstruction error. With these methods, we surprisingly find that EasyQuant achieves comparable performance to the original model. Since EasyQuant does not depend on any training data, the generalization performance of quantized LLMs is safely guaranteed. Moreover, EasyQuant can be implemented in parallel so that the quantized model could be attained in a few minutes even for LLMs over 100B. To our best knowledge, we are the first work that achieves almost lossless quantization performance for LLMs under a data-independent setting and our algorithm runs over 10 times faster than the data-dependent methods.

09:00-10:30 (East Foyer / Virtual)
**#14 Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models**
*Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A. Smith and Yulia Tsvetkov*
Language models have graduated from being research prototypes to commercialized products offered as web APIs, and recent works have highlighted the multilingual capabilities of these products. The API vendors charge their users based on usage, more specifically on the number of "tokens" processed or generated by the underlying language models. What constitutes a token, however, is training data and model dependent with a large variance in the number of tokens required to convey the same information in different languages. In this work, we analyze the effect of this non-uniformity on the fairness of an API's pricing policy across languages. We conduct a systematic analysis of the cost and utility of OpenAI's language model API on multilingual benchmarks in 22 typologically diverse languages. We show evidence that speakers of a large number of the supported languages are overcharged while obtaining poorer results. These speakers tend to also come from regions where the APIs are less affordable, to begin with. Through these analyses, we aim to increase transparency around language model APIs' pricing policies and encourage the vendors to make them more equitable.

09:00-10:30 (East Foyer / Virtual)
**#15 Editing Large Language Models: Problems, Methods, and Opportunities**
*Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen and Ningyu Zhang*
Despite the ability to train capable LLMs, the methodology for maintaining their relevancy and rectifying errors remains elusive. To this end, the past few years have witnessed a surge in techniques for editing LLMs, the objective of which is to alter the behavior of LLMs **efficiently** within a specific domain without negatively impacting performance across other inputs. This paper embarks on a deep exploration of the problems, methods, and opportunities related to model editing for LLMs. In particular, we provide an exhaustive overview of the task definition and challenges associated with model editing, along with an in-depth empirical analysis of the most progressive methods currently at our disposal. We also build a new benchmark dataset to facilitate a more robust evaluation and pinpoint enduring issues intrinsic to existing techniques. Our objective is to provide valuable insights into the effectiveness and feasibility of each editing technique, thereby assisting the community in making informed decisions on the selection of the most appropriate method for a specific task or context.

09:00-10:30 (East Foyer / Virtual)
**#16 Knowledge Rumination for Pre-trained Language Models**
*Yunzhi Yao, Peng Wang, Shengyu Mao, Chuanqi Tan, Fei Huang, Huajun Chen and Ningyu Zhang*
Previous studies have revealed that vanilla pre-trained language models (PLMs) lack the capacity to handle knowledge-intensive NLP tasks alone; thus, several works have attempted to integrate external knowledge into PLMs. However, despite the promising outcome, we empirically observe that PLMs may have already encoded rich knowledge in their pre-trained parameters but fails to fully utilize them when applying to knowledge-intensive tasks. In this paper, we propose a new paradigm dubbed **Knowledge Rumination** to help the pre-trained language model utilize that related latent knowledge without retrieving them from the external corpus. By simply adding a prompt like *"As far as I know"* to the PLMs, we try to review related latent knowledge and inject them back into the model for knowledge consolidation. We apply the proposed knowledge rumination to various language models, including RoBERTa, DeBERTa, and GPT-3. Experimental results on six commonsense reasoning tasks and GLUE benchmarks demonstrate the effectiveness of our proposed approach, which proves that the knowledge stored in PLMs can be better exploited to enhance performance.

09:00-10:30 (East Foyer / Virtual)
**#17 Revisiting Automated Topic Model Evaluation with Large Language Models**
*Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan and Elliott Ash*
Topic models help us make sense of large text collections. Automatically evaluating their output and determining the optimal number of topics are both longstanding challenges, with no effective automated solutions to date. This paper proposes using large language models (LLMs) for these tasks. We find that LLMs appropriately assess the resulting topics, correlating more strongly with human judgments than existing automated metrics. However, the setup of the evaluation task is crucial — LLMs perform better on coherence ratings of word sets than on intrustion detection. We find that LLMs can also assist us in guiding us towards a reasonable number of topics. In actual applications, topic models are typically used to answer a research question related to a collection of texts. We can incorporate this research question in the prompt to the LLM, which helps estimating the optimal number of topics.

09:00-10:30 (East Foyer / Virtual)
**#18 Prompting with Pseudo-Code Instructions**
*Mayank Mishra, Prince Kumar, Riyaz Ahmad Bhat, Rudra Murthy, Danish Contractor and Srikanth G. Tamilselvam*
Prompting with natural language instructions has recently emerged as a popular method of harnessing the capabilities of large language models (LLM). Given the inherent ambiguity present in natural language, it is intuitive to consider the possible advantages of prompting with less ambiguous prompt styles, like pseudo-code. In this paper, we explore if prompting via pseudo-code instructions helps improve the performance of pre-trained language models. We manually create a dataset of pseudo-code prompts for 132 different tasks spanning classification, QA, and generative language tasks, sourced from the Super-NaturalInstructions dataset. Using these prompts along with their counterparts in natural language, we study their performance on two LLM families - BLOOM, CodeGen. Our experiments show that using pseudo-code instructions leads to better results, with an average increase (absolute) of 7-16 points in F1 scores for classification tasks and an improvement (relative) of 12-38% in aggregate ROUGE-L scores across all tasks. We include detailed ablation studies which indicate that code comments, docstrings, and the structural clues encoded in pseudo-code all contribute towards the improvement in performance. To the best of our knowledge, our work is the first to demonstrate how pseudo-code prompts can be helpful in improving the performance of pre-trained LMs.

09:00-10:30 (East Foyer / Virtual)

**#19 Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks**
*Andrea Sottana, Bin Liang, Kai Zou and Zheng Yuan*
Large Language Models (LLMs) evaluation is a patchy and inconsistent landscape, and it is becoming clear that the quality of automatic evaluation metrics is not keeping up with the pace of development of generative models. We aim to improve the understanding of current models' performance by providing a preliminary and hybrid evaluation on a range of open and closed-source generative LLMs on three NLP benchmarks: text summarisation, text simplification and grammatical error correction (GEC), using both automatic and human evaluation. We also explore the potential of the recently released GPT-4 to act as an evaluator. We find that ChatGPT consistently outperforms many other popular models according to human reviewers on the majority of metrics, while scoring much more poorly when using classic automatic evaluation metrics. We also find that human reviewers rate the gold reference as much worse than the best models' outputs, indicating the poor quality of many popular benchmarks. Finally, we find that GPT-4 is capable of ranking models' outputs in a way which aligns reasonably closely to human judgement despite task-specific variations, with a lower alignment in the GEC task.

09:00-10:30 (East Foyer / Virtual)
**#20 IAG: Induction-Augmented Generation Framework for Answering Reasoning Questions**
*Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai and Zhao Cao*
Retrieval-Augmented Generation (RAG), by incorporating external knowledge with parametric memory of language models, has become the state-of-the-art architecture for open-domain QA tasks. However, common knowledge bases are inherently constrained by limited coverage and noisy information, making retrieval-based approaches inadequate to answer implicit reasoning questions. In this paper, we propose an Induction-Augmented Generation (IAG) framework that utilizes inductive knowledge along with the retrieved documents for implicit reasoning. We leverage large language models (LLMs) for deriving such knowledge via a novel prompting method based on inductive reasoning patterns. On top of this, we implement two versions of IAG named IAG-GPT and IAG-Student, respectively. IAG-GPT directly utilizes the knowledge generated by GPT-3 for answer prediction, while IAG-Student gets rid of dependencies on GPT service at inference time by incorporating a student inductor model. The inductor is firstly trained via knowledge distillation and further optimized by back-propagating the generator feedback via differentiable beam scores. Experimental results show that IAG outperforms RAG baselines as well as ChatGPT on two Open-Domain QA tasks. Notably, our best models have won the first place in the official leaderboards of CSQA2.0 (since Nov 1, 2022) and StrategyQA (since Jan 8, 2023).

09:00-10:30 (East Foyer / Virtual)
**#21 Large Language Models Meet Open-World Intent Discovery and Recognition: An Evaluation of ChatGPT**
*Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai and Weiran Xu*
The tasks of out-of-domain (OOD) intent discovery and generalized intent discovery (GID) aim to extend a closed intent classifier to open-world intent sets, which is crucial to task-oriented dialogue (TOD) systems. Previous methods address them by fine-tuning discriminative models. Recently, although some studies has been exploring the application of large language models (LLMs) represented by ChatGPT to various downstream tasks, it is still unclear for the ability of ChatGPT to discover and incrementally extent OOD intents. In this paper, we comprehensively evaluate ChatGPT on OOD intent discovery and GID, and then outline the strengths and weaknesses of ChatGPT. Overall, ChatGPT exhibits consistent advantages under zero-shot settings, but is still at a disadvantage compared to fine-tuned models. More deeply, through a series of analytical experiments, we summarize and discuss the challenges faced by LLMs including clustering, domain-specific understanding, and cross-domain in-context learning scenarios. Finally, we provide empirical guidance for future directions to address these challenges.

09:00-10:30 (East Foyer / Virtual)
**#22 TheoremQA: A Theorem-driven Question Answering Dataset**
*Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang and Tony Xia*
The recent LLMs like GPT-4 and PaLM-2 have made tremendous progress in solving fundamental math problems like GSM8K by achieving over 90% accuracy. However, their capabilities to solve more challenging math problems which require domain-specific knowledge (i.e. theorem) have yet to be investigated. In this paper, we introduce TheoremQA, the first theorem-driven question-answering dataset designed to evaluate AI models' capabilities to apply theorems to solve challenging science problems. TheoremQA is curated by domain experts containing 800 high-quality questions covering 350 theorems from Math, Physics, EE&CS, and Finance. We evaluate a wide spectrum of 16 large language and code models with different prompting strategies like Chain-of-Thoughts and Program-of-Thoughts. We found that GPT-4's capabilities to solve these problems are unparalleled, achieving an accuracy of 51% with Program-of-Thoughts Prompting. All the existing open-sourced models are below 15%, barely surpassing the random-guess baseline. Given the diversity and broad coverage of TheoremQA, we believe it can be used as a better benchmark to evaluate LLMs' capabilities to solve challenging science problems.

09:00-10:30 (East Foyer / Virtual)
**#23 TLM: Token-Level Masking for Transformers**
*Yangjun Wu, Kebin Fang, Dongxiang Zhang, Han Wang, Hao Zhang and Gang Chen*
Structured dropout approaches, such as attention dropout and DropHead, have been investigated to regularize the multi-head attention mechanism in Transformers. In this paper, we propose a new regularization scheme based on token-level rather than structure-level to reduce overfitting. Specifically, we devise a novel Token-Level Masking (TLM) training strategy for Transformers to regularize the connections of self-attention, which consists of two masking techniques that are effective and easy to implement. The underlying idea is to manipulate the connections between tokens in the multi-head attention via masking, where the networks are forced to exploit partial neighbors' information to produce a meaningful representation. The generality and effectiveness of TLM are thoroughly evaluated via extensive experiments on 4 diversified NLP tasks across 18 datasets, including natural language understanding benchmark GLUE, ChineseGLUE, Chinese Grammatical Error Correction, and data-to-text generation. The results indicate that TLM can consistently outperform attention dropout and DropHead, e.g., it increases by 0.5 points relative to DropHead with BERT-large on GLUE. Moreover, TLM can establish a new record on the data-to-text benchmark Rotowire (18.93 BLEU). Our code will be publicly available at https://github.com/Young1993/tlm.

09:00-10:30 (East Foyer / Virtual)
**#24 Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus**
*Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang and Luoyi Fu*
Large Language Models (LLMs) have gained significant popularity for their impressive performance across diverse fields. However, LLMs are prone to hallucinate untruthful or nonsensical outputs that fail to meet user expectations in many real-world applications. Existing works for detecting hallucinations in LLMs either rely on external knowledge for reference retrieval or require sampling multiple responses from the LLM for consistency verification, making these methods costly and inefficient. In this paper, we propose a novel reference-free, uncertainty-based method for detecting hallucinations in LLMs. Our approach imitates human focus in factuality checking from three aspects: 1) focus on the most informative and important keywords in the given text; 2) focus on the unreliable tokens in historical context which may lead to a cascade of hallucinations; and 3) focus on the token properties such as token type and token frequency. Experimental results on relevant datasets demonstrate the effectiveness of our proposed method, which achieves state-of-the-art performance across all the evaluation metrics and eliminates the need for additional information.

09:00-10:30 (East Foyer / Virtual)
**#25 MoPe: Model Perturbation based Privacy Attacks on Language Models**
*Marvin Li, Jason Wang, Jeffrey George Wang and Seth Neel*
Recent work has shown that Large Language Models (LLMs) can unintentionally leak sensitive information present in their training data. In this paper, we present Model Perturbations (MoPe), a new method to identify with high confidence if a given text is in the training data of a pre-trained language model, given white-box access to the models parameters. MoPe adds noise to the model in parameter space and measures the drop in log-likelihood at a given point $x$, a statistic we show approximates the trace of the Hessian matrix with respect to model parameters. Across language models ranging from $70\text{M}$ to $12\text{B}$ parameters, we show that MoPe is more effective than existing loss-based attacks and recently proposed perturbation-based methods. We also examine the role of training point order and model size in attack success, and empirically demonstrate that MoPe accurately approximate the trace of the Hessian in practice. Our results show that the loss of a point alone is insufficient to determine extractability—there are training points we can recover using our method that have average loss. This casts some doubt on prior works that use the loss of a point as evidence of memorization or unlearning.

09:00-10:30 (East Foyer / Virtual)
**#26 The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations**
*Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth and Amitava Das*
The recent advancements in Large Language Models (LLMs) have garnered widespread acclaim for their remarkable emerging capabilities. However, the issue of hallucination has parallelly emerged as a by-product, posing significant concerns. While some recent endeavors have been made to identify and mitigate different types of hallucination, there has been a limited emphasis on the nuanced categorization of hallucination and associated mitigation methods. To address this gap, we offer a fine-grained discourse on profiling hallucination based on its degree, orientation, and category, along with offering strategies for alleviation. As such, we define two overarching orientations of hallucination: (i) factual mirage (FM) and (ii) silver lining (SL). To provide a more comprehensive understanding, both orientations are further sub-categorized into intrinsic and extrinsic, with three degrees of severity - (i) mild, (ii) moderate, and (iii) alarming. We also meticulously categorize hallucination into six types: (i) acronym ambiguity, (ii) numeric nuisance, (iii) generated golem, (iv) virtual voice, (v) geographic erratum, and (vi) time wrap. Furthermore, we curate HallucInation eLiciTation (HILT), a publicly available dataset comprising of 75,000 samples generated using 15 contemporary LLMs along with human annotations for the aforementioned categories. Finally, to establish a method for quantifying and to offer a comparative spectrum that allows us to evaluate and rank LLMs based on their vulnerability to producing hallucinations, we propose Hallucination Vulnerability Index (HVI). Amidst the extensive deliberations on policy-making for regulating AI development, it is of utmost importance to assess and measure which LLM is more vulnerable towards hallucination. We firmly believe that HVI holds significant value as a tool for the wider NLP community, with the potential to serve as a rubric in AI-related policy-making. In conclusion, we propose two solution strategies for mitigating hallucinations.

09:00-10:30 (East Foyer / Virtual)
**#27 Simple and Effective Input Reformulations for Translation**
*Brian Yu, Hansen Lillemark and Kurt Keutzer*
Foundation language models learn from their finetuning input context in different ways. In this paper, we reformulate inputs during finetuning for challenging translation tasks, leveraging model strengths from pretraining in novel ways to improve downstream performance. These reformulations are simple data level modifications, require no additional collection of training data or modification of data at inference time. They can be applied either on single language pair translation tasks or massively multilingual translation tasks. Experiments with these techniques demonstrate significant performance improvements up to **3.5 chrF++ on the Flores200 translation benchmark**. We hope our research accessibly improves finetuning data efficiency, enabling more effective training to scalably improve state-of-the-art performance. Our code is released here.

09:00-10:30 (East Foyer / Virtual)
**#28 Revisiting De-Identification of Electronic Medical Records: Evaluation of Within- and Cross-Hospital Generalization**
*Yiyang Liu, Jinpeng Li and Enwei Zhu*
The de-identification task aims to detect and remove the protected health information from electronic medical records (EMRs). Previous studies generally focus on the within-hospital setting and achieve great successes, while the cross-hospital setting has been overlooked. This study introduces a new de-identification dataset comprising EMRs from three hospitals in China, creating a benchmark for evaluating both within- and cross-hospital generalization. We find significant domain discrepancy between hospitals. A model with almost perfect within-hospital performance struggles when transferred across hospitals. Further experiments show that pretrained language models and some domain generalization methods can alleviate this problem. We believe that our data and findings will encourage investigations on the generalization of medical NLP models.

09:00-10:30 (East Foyer / Virtual)
**#29 Federated Learning of Large Language Models with Parameter-Efficient Prompt Tuning and Adaptive Optimization**
*Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor S. Sheng, Huaiyu Dai and Dejing Dou*
Federated learning (FL) is a promising paradigm to enable collaborative model training with decentralized data. However, the training process of Large Language Models (LLMs) generally incurs the update of significant parameters, which limits the applicability of FL techniques to tackle the LLMs in real scenarios. Prompt tuning can significantly reduce the number of parameters to update, but it either incurs performance degradation or low training efficiency. The straightforward utilization of prompt tuning in the FL often raises non-trivial communication costs and dramatically degrades performance. In addition, the decentralized data is generally non-Independent and Identically Distributed (non-IID), which brings client drift problems and thus poor performance. This paper proposes a Parameter-efficient prompt Tuning approach with Adaptive Optimization, i.e., FedPepTAO, to enable efficient and effective FL of LLMs. First, an efficient partial prompt tuning approach is proposed to improve performance and efficiency simultaneously. Second, a novel adaptive optimization method is developed to address the client drift problems on both the device and server sides to enhance performance further. Extensive experiments based on 10 datasets demonstrate the superb performance (up to 60.8% in terms of accuracy) and efficiency (up to 97.59% in terms of training time) of FedPepTAO compared with 9 baseline approaches. Our code is available at https://github.com/llm-eff/FedPepTAO.

09:00-10:30 (East Foyer / Virtual)
**#30 API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs**
*Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang and Yongbin Li*
Recent research has demonstrated that Large Language Models (LLMs) can enhance their capabilities by utilizing external tools. However, three pivotal questions remain unanswered: (1) How effective are current LLMs in utilizing tools? (2) How can we enhance LLMs' ability to utilize tools? (3) What obstacles need to be overcome to leverage tools? To address these questions, we introduce API-Bank, a groundbreaking benchmark, specifically designed for tool-augmented LLMs. For the first question, we develop a runnable evaluation system consisting of 73

API tools. We annotate 314 tool-use dialogues with 753 API calls to assess the existing LLMs' capabilities in planning, retrieving, and calling APIs. For the second question, we construct a comprehensive training set containing 1,888 tool-use dialogues from 2,138 APIs spanning 1,000 distinct domains. Using this dataset, we train Lynx, a tool-augmented LLM initialized from Alpaca. Experimental results demonstrate that GPT-3.5 exhibits improved tool utilization compared to GPT-3, while GPT-4 excels in planning. However, there is still significant potential for further improvement. Moreover, Lynx surpasses Alpaca's tool utilization performance by more than 26 pts and approaches the effectiveness of GPT-3.5. Through error analysis, we highlight the key challenges for future research in this field to answer the third question.

09:00-10:30 (East Foyer / Virtual)
**#31 Document-Level Machine Translation with Large Language Models**
*Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi and Zhaopeng Tu*
Large language models (LLMs) such as ChatGPT can produce coherent, cohesive, relevant, and fluent answers for various natural language processing (NLP) tasks. Taking document-level machine translation (MT) as a testbed, this paper provides an in-depth evaluation of LLMs' ability on discourse modeling. The study focuses on three aspects: 1) Effects of Context-Aware Prompts, where we investigate the impact of different prompts on document-level translation quality and discourse phenomena; 2) Comparison of Translation Models, where we compare the translation performance of ChatGPT with commercial MT systems and advanced document-level MT methods; 3) Analysis of Discourse Modelling Abilities, where we further probe discourse knowledge encoded in LLMs and shed light on impacts of training techniques on discourse modeling. By evaluating on a number of benchmarks, we surprisingly find that LLMs have demonstrated superior performance and show potential to become a new paradigm for document-level translation: 1) leveraging their powerful long-text modeling capabilities, GPT-3.5 and GPT-4 outperform commercial MT systems in terms of human evaluation; 2) GPT-4 demonstrates a stronger ability for probing linguistic knowledge than GPT-3.5. This work highlights the challenges and opportunities of LLMs for MT, which we hope can inspire the future design and evaluation of LLMs (We release our data and annotations at https://github.com/longyuewangdcu/Document-MT-LLM).

09:00-10:30 (East Foyer / Virtual)
**#32 Self-Evolution Learning for Mixup: Enhance Data Augmentation on Few-Shot Text Classification Tasks**
*Haoqi Zheng, Qihuang Zhong, Liang Ding, Zhiliang Tian, Xin Niu, Changjian Wang, Dongsheng Li and Dacheng Tao*
Text classification tasks often encounter few-shot scenarios with limited labeled data, and addressing data scarcity is crucial. Data augmentation with mixup merges sample pairs to generate new pseudos, which can relieve the data deficiency issue in text classification. However, the quality of pseudo-samples generated by mixup exhibits significant variations. Most of the mixup methods fail to consider the varying degree of learning difficulty in different stages of training. And mixup generates new samples with one-hot labels, which encourages the model to produce a high prediction score for the correct class that is much larger than other classes, resulting in the model's over-confidence. In this paper, we propose a self-evolution learning (SE) based mixup approach for data augmentation in text classification, which can generate more adaptive and model-friendly pseudo samples for the model training. SE caters to the growth of the model learning ability and adapts to the ability when generating training samples. To alleviate the model over-confidence, we introduce an instance-specific label smoothing regularization approach, which linearly interpolates the model's output and one-hot labels of the original samples to generate new soft labels for label mixing up. Through experimental analysis, experiments show that our SE brings consistent and significant improvements upon different mixup methods. In-depth analyses demonstrate that SE enhances the model's generalization ability.

09:00-10:30 (East Foyer / Virtual)
**#33 Argue with Me Tersely: Towards Sentence-Level Counter-Argument Generation**
*Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang and zhongyu wei*
Counter-argument generation—a captivating area in computational linguistics—seeks to craft statements that offer opposing views. While most research has ventured into paragraph-level generation, sentence-level counter-argument generation beckons with its unique constraints and brevity-focused challenges. Furthermore, the diverse nature of counter-arguments poses challenges for evaluating model performance solely based on n-gram-based metrics. In this paper, we present the ArgTersely benchmark for sentence-level counter-argument generation, drawing from a manually annotated dataset from the ChangeMyView debate forum. We also propose Arg-LlaMA for generating high-quality counter-argument. For better evaluation, we trained a BERT-based evaluator Arg-Judge with human preference data. We conducted comparative experiments involving various baselines such as LlaMA, Alpaca, GPT-3, and others. The results show the competitiveness of our proposed framework and evaluator in counter-argument generation tasks. Code and data are available at https://github.com/amazingljy1206/ArgTersely.

09:00-10:30 (East Foyer / Virtual)
**#34 DIVE: Towards Descriptive and Diverse Visual Commonsense Generation**
*Jun-Hyung Park, Hyuntae Park, Youjin Kang, Eojin Jeon and SangKeun Lee*
Towards human-level visual understanding, visual commonsense generation has been introduced to generate commonsense inferences beyond images. However, current research on visual commonsense generation has overlooked an important human cognitive ability: generating descriptive and diverse inferences. In this work, we propose a novel visual commonsense generation framework, called DIVE, which aims to improve the descriptiveness and diversity of generated inferences. DIVE involves two methods, generic inference filtering and contrastive retrieval learning, which address the limitations of existing visual commonsense resources and training objectives. Experimental results verify that DIVE outperforms state-of-the-art models for visual commonsense generation in terms of both descriptiveness and diversity, while showing a superior quality in generating unique and novel inferences. Notably, DIVE achieves human-level descriptiveness and diversity on Visual Commonsense Graphs. Furthermore, human evaluations confirm that DIVE aligns closely with human judgments on descriptiveness and diversity.

09:00-10:30 (East Foyer / Virtual)
**#35 SUT: Active Defects Probing for Transcompiler Models**
*Mengnan Qi, Yufan Huang, Maoquan Wang, Yongqiang Yao, Zihan Liu, Bin Gu, Colin Clement and Neel Sundaresan*
Automatic Program translation has enormous application value and hence has been attracting significant interest from AI researchers. However, we observe that current program translation models still make elementary syntax errors, particularly, when the target language does not have syntax elements in the source language. Metrics like BLUE, CodeBLUE and computation accuracy may not expose these issues. In this paper we introduce a new metrics for programming language translation and these metrics address these basic syntax errors. We develop a novel active defects probing suite called Syntactic Unit Tests (SUT) which includes a highly interpretable evaluation harness for accuracy and test scoring. Experiments have shown that even powerful models like ChatGPT still make mistakes on these basic unit tests. Specifically, compared to previous program translation task evaluation dataset, its pass rate on our unit tests has decreased by 26.15%. Further our evaluation harness reveal syntactic element errors in which these models exhibit deficiencies.

09:00-10:30 (East Foyer / Virtual)
**#36 Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study**
*Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar and Bryan Catanzaro*
Large decoder-only language models (LMs) can be largely improved in terms of perplexity by retrieval (e.g., RETRO), but its impact on text

generation quality and downstream task accuracy is unclear. Thus, it is still an open question: shall we pretrain large autoregressive LMs with retrieval? To answer it, we perform a comprehensive study on a scalable pre-trained retrieval-augmented LM (i.e., RETRO) compared with standard GPT and retrieval-augmented GPT incorporated at fine-tuning or inference stages. We first provide the recipe to reproduce RETRO up to 9.5B parameters while retrieving a text corpus with 330B tokens. Based on that, we have the following novel findings: i) RETRO outperforms GPT on text generation with much less degeneration (i.e., repetition), moderately higher factual accuracy, and slightly lower toxicity with a nontoxic retrieval database. ii) On the LM Evaluation Harness benchmark, RETRO largely outperforms GPT on knowledge-intensive tasks, but is on par with GPT on other tasks. Furthermore, we introduce a simple variant of the model, RETRO++, which largely improves open-domain QA results of original RETRO (e.g., EM score +8.6 on Natural Question) and significantly outperforms retrieval-augmented GPT across different model sizes. Our findings highlight the promising direction of pretraining autoregressive LMs with retrieval as future foundation models. We release our implementation at: https://github.com/NVIDIA/Megatron-LM/tree/main/tools/retro.

09:00-10:30 (East Foyer / Virtual)
**#37 clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents**
*Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler and David Schlangen*
Recent work has proposed a methodology for the systematic evaluation of "Situated Language Understanding Agents" — agents that operate in rich linguistic and non-linguistic contexts — through testing them in carefully constructed interactive settings. Other recent work has argued that Large Language Models (LLMs), if suitably set up, can be understood as (simulators of) such agents. A connection suggests itself, which this paper explores: Can LLMs be evaluated meaningfully by exposing them to constrained game-like settings that are built to challenge specific capabilities? As a proof of concept, this paper investigates five interaction settings, showing that current chat-optimised LLMs are, to an extent, capable of following game-play instructions. Both this capability and the quality of the game play, measured by how well the objectives of the different games are met, follows the development cycle, with newer models generally performing better. The metrics even for the comparatively simple example games are far from being saturated, suggesting that the proposed instrument will remain to have diagnostic value.

09:00-10:30 (East Foyer / Virtual)
**#38 Task-Level Thinking Steps Help Large Language Models for Challenging Classification Task**
*Chunhui Du, Jidong Tian, Haoran Liao, Jindou Chen, Hao He and Yaohui Jin*
Large language models (LLMs) have shown incredible performance on many tasks such as dialogue generation, commonsense reasoning and question answering. In-context learning (ICL) is an important paradigm for adapting LLMs to the downstream tasks by prompting few demonstrations. However, the distribution of demonstrations can severely affect the performance, especially for challenging classification tasks. In this paper, we propose the concept of task-level thinking steps that can eliminate bias introduced by demonstrations. Further, to help LLMs distinguish confusing classes, we design a progressive revision framework, which can improve the thinking steps by correcting hard demonstrations. Experimental results prove the superiority of our proposed method, achieving best performance on three kinds of challenging classification tasks in the zero-shot and few-shot settings. Besides, with task-level thinking steps, automatically generated chain-of-thoughts (CoTs) bring more competitive performance.

09:00-10:30 (East Foyer / Virtual)
**#39 Reduce Human Labor On Evaluating Conversational Information Retrieval System: A Human-Machine Collaboration Approach**
*Chen Huang, Peixin Qin, Wenqiang Lei and Jiancheng Lv*
Evaluating conversational information retrieval (CIR) systems is a challenging task that requires a significant amount of human labor for annotation. It is imperative to invest significant effort into researching more labor-effective methods for evaluating CIR systems. To touch upon this challenge, we take the first step to involve active testing in CIR evaluation and propose a novel method, called HomCoE. It strategically selects a few data for human annotation, then calibrates the evaluation results to eliminate evaluation biases. As such, it makes an accurate evaluation of the CIR system at low human labor. We experimentally reveal that it consumes less than 1% of human labor and achieves a consistency rate of 95%-99% with human evaluation results. This emphasizes the superiority of our method over other baselines.

09:00-10:30 (East Foyer / Virtual)
**#40 DiNeR: A Large Realistic Dataset for Evaluating Compositional Generalization**
*Chengang Hu, Xiao Liu and Yansong Feng*
Most of the existing compositional generalization datasets are synthetically-generated, resulting in a lack of natural language variation. While there have been recent attempts to introduce non-synthetic datasets for compositional generalization, they suffer from either limited data scale or a lack of diversity in the forms of combinations. To better investigate compositional generalization with more linguistic phenomena and compositional diversity, we propose the DIsh NamE Recognition (DiNeR) task and create a large realistic Chinese dataset. Given a recipe instruction, models are required to recognize the dish name composed of diverse combinations of food, actions, and flavors. Our dataset consists of 3,811 dishes and 228,114 recipes, and involves plenty of linguistic phenomena such as anaphora, omission and ambiguity. We provide two strong baselines based on T5 and large language models (LLMs). This work contributes a challenging task, baseline methods to tackle the task, and insights into compositional generalization in the context of dish name recognition.

09:00-10:30 (East Foyer / Virtual)
**#41 Active Instruction Tuning: Improving Cross-Task Generalization by Training on Prompt Sensitive Tasks**
*Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang and Nanyun Peng*
Instruction tuning (IT) achieves impressive zero-shot generalization results by training large language models (LLMs) on a massive amount of diverse tasks with instructions. However, how to select new tasks to improve the performance and generalizability of IT models remains an open question. Training on all existing tasks is impractical due to prohibiting computation requirements, and randomly selecting tasks can lead to suboptimal performance. In this work, we propose active instruction tuning based on prompt uncertainty, a novel framework to identify informative tasks, and then actively tune the models on the selected tasks. We represent the informativeness of new tasks with the disagreement of the current model outputs over perturbed prompts. Our experiments on NIV2 and Self-Instruct datasets demonstrate that our method consistently outperforms other baseline strategies for task selection, achieving better out-of-distribution generalization with fewer training tasks. Additionally, we introduce a task map that categorizes and diagnoses tasks based on prompt uncertainty and prediction probability. We discover that training on ambiguous (prompt-uncertain) tasks improves generalization while training on difficult (prompt-certain and low-probability) tasks offers no benefit, underscoring the importance of task selection for instruction tuning.

09:00-10:30 (East Foyer / Virtual)
**#42 On Evaluation of Bangla Word Analogies**
*Mousumi Akter, Souvika Sarkar and Shubhra Kanti Karmaker Santu*
This paper presents a benchmark dataset of Bangla word analogies for evaluating the quality of existing Bangla word embeddings. Despite being the 7th largest spoken language in the world, Bangla is still a low-resource language and popular NLP models often struggle to perform well on Bangla data sets. Therefore, developing a robust evaluation set is crucial for benchmarking and guiding future research on improving

Bangla word embeddings, which is currently missing. To address this issue, we introduce a new evaluation set of 16,678 unique word analogies in Bangla as well as a translated and curated version of the original Mikolov dataset (10,594 samples) in Bangla. Our experiments with different state-of-the-art embedding models reveal that current Bangla word embeddings struggle to achieve high accuracy on both data sets, demonstrating a significant gap in multilingual NLP research.

09:00-10:30 (East Foyer / Virtual)
**#43 Establishing Trustworthiness: Rethinking Tasks and Model Evaluation**
*Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel and Barbara Plank*
Language understanding is a multi-faceted cognitive capability, which the Natural Language Processing (NLP) community has striven to model computationally for decades. Traditionally, facets of linguistic intelligence have been compartmentalized into tasks with specialized model architectures and corresponding evaluation protocols. With the advent of large language models (LLMs) the community has witnessed a dramatic shift towards general purpose, task-agnostic approaches powered by generative models. As a consequence, the traditional compartmentalized notion of language tasks is breaking down, followed by an increasing challenge for evaluation and analysis. At the same time, LLMs are being deployed in more real-world scenarios, including previously unforeseen zero-shot setups, increasing the need for trustworthy and reliable systems. Therefore, we argue that it is time to rethink what constitutes tasks and model evaluation in NLP, and pursue a more holistic view on language, placing trustworthiness at the center. Towards this goal, we review existing compartmentalized approaches for understanding the origins of a model's functional capacity, and provide recommendations for more multi-faceted evaluation protocols.

09:00-10:30 (East Foyer / Virtual)
**#44 Democratizing Reasoning Ability: Tailored Learning from Large Language Model**
*Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun and Qi Zhang*
Large language models (LLMs) exhibit impressive emergent abilities in natural language processing, but their democratization is hindered due to huge computation requirements and closed-source nature. Recent research on advancing open-source smaller LMs by distilling knowledge from black-box LLMs has obtained promising results in the instruction-following ability. However, the reasoning ability which is more challenging to foster, is relatively rarely explored. In this paper, we propose a tailored learning approach to distill such reasoning ability to smaller LMs to facilitate the democratization of the exclusive reasoning ability. In contrast to merely employing LLM as a data annotator, we exploit the potential of LLM as a reasoning teacher by building an interactive multi-round learning paradigm. This paradigm enables the student to expose its deficiencies to the black-box teacher who then can provide customized training data in return. Further, to exploit the reasoning potential of the smaller LM, we propose self-reflection learning to motivate the student to learn from self-made mistakes. The learning from self-reflection and LLM are all tailored to the student's learning status, thanks to the seamless integration with the multi-round learning paradigm. Comprehensive experiments and analysis on mathematical and commonsense reasoning tasks demonstrate the effectiveness of our method. The code will be available at https://github.com/Raibows/Learn-to-Reason.

09:00-10:30 (East Foyer / Virtual)
**#45 SMoP: Towards Efficient and Effective Prompt Tuning with Sparse Mixture-of-Prompts**
*Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok and SangKeun Lee*
Prompt tuning has emerged as a successful parameter-efficient alternative to the full fine-tuning of language models. However, prior works on prompt tuning often utilize long soft prompts of up to 100 tokens to improve performance, overlooking the inefficiency associated with extended inputs. In this paper, we propose a novel prompt tuning method $SMoP$ ($S$parse $M$ixture-$o$f-$P$rompts) that utilizes short soft prompts for efficient training and inference while maintaining performance gains typically induced from longer soft prompts. To achieve this, $SMoP$ employs a gating mechanism to train multiple short soft prompts specialized in handling different subsets of the data, providing an alternative to relying on a single long soft prompt to cover the entire data. Experimental results demonstrate that $SMoP$ outperforms baseline methods while reducing training and inference costs. We release our code at https://github.com/jyjohnchoi/SMoP.

09:00-10:30 (East Foyer / Virtual)
**#46 CRAB: Assessing the Strength of Causal Relationships Between Real-world Events**
*Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer and Antoine Bosselut*
Understanding narratives requires reasoning about the cause-and-effect relationships between events mentioned in the text. While existing foundation models yield impressive results in many NLP tasks requiring reasoning, it is unclear whether they understand the complexity of the underlying network of causal relationships of events in narratives. In this work, we present CRAB, a new Causal Reasoning Assessment Benchmark designed to evaluate causal understanding of events in real-world narratives. CRAB contains fine-grained, contextual causality annotations for $\sim$2.7K pairs of real-world events that describe various newsworthy event timelines (e.g., the acquisition of Twitter by Elon Musk). Using CRAB, we measure the performance of several large language models, demonstrating that most systems achieve poor performance on the task. Motivated by classical causal principles, we also analyze the causal structures of groups of events in CRAB, and find that models perform worse on causal reasoning when events are derived from complex causal structures compared to simple linear causal chains. We make our dataset and code available to the research community.

09:00-10:30 (East Foyer / Virtual)
**#47 Chinese Lexical Substitution: Dataset and Method**
*Jipeng Qiang, Kang Liu, Ying Li, Yun Li, Yi Zhu, Yun-Hao Yuan, Xiaocheng Hu and Xiaoye Ouyang*
Existing lexical substitution (LS) benchmarks were collected by asking human annotators to think of substitutes from memory, resulting in benchmarks with limited coverage and relatively small scales. To overcome this problem, we propose a novel annotation method to construct an LS dataset based on human and machine collaboration. Based on our annotation method, we construct the first Chinese LS dataset CHNLS which consists of 33,695 instances and 144,708 substitutes, covering three text genres (News, Novel, and Wikipedia). Specifically, we first combine four unsupervised LS methods as an ensemble method to generate the candidate substitutes, and then let human annotators judge these candidates or add new ones. This collaborative process combines the diversity of machine-generated substitutes with the expertise of human annotators. Experimental results that the ensemble method outperforms other LS methods. To our best knowledge, this is the first study for the Chinese LS task.

09:00-10:30 (East Foyer / Virtual)
**#48 Somali Information Retrieval Corpus: Bridging the Gap between Query Translation and Dedicated Language Resources**
*Abdisalam Mahamed Badel, Ting Zhong, Wenxin Tai and Fan Zhou*
Despite the growing use of the Somali language in various online domains, research on Somali language information retrieval remains limited and primarily relies on query translation due to the lack of a dedicated corpus. To address this problem, we collaborated with language experts and natural language processing (NLP) researchers to create an annotated corpus for Somali information retrieval. This corpus comprises 2335 documents collected from various well-known online sites, such as hiiraan online, dhacdo net, and Somali poetry books. We explain how the corpus was constructed, and develop a Somali language information retrieval system using a pseudo-relevance feedback (PRF) query expansion technique on the corpus. Note that collecting such a data set for the low-resourced Somali language can help overcome NLP

barriers, such as the lack of electronically available data sets. Which, if available, can enable the development of various NLP tools and applications such as question-answering and text classification. It also provides researchers with a valuable resource for investigating and developing new techniques and approaches for Somali.

09:00-10:30 (East Foyer / Virtual)
**#49 Countering Misinformation via Emotional Response Generation**
*Daniel Russo, Shane Peter Kaszefski-Yaschuk, Jacopo Staiano and Marco Guerini*
The proliferation of misinformation on social media platforms (SMPs) poses a significant danger to public health, social cohesion and ultimately democracy. Previous research has shown how social correction can be an effective way to curb misinformation, by engaging directly in a constructive dialogue with users who spread – often in good faith – misleading messages. Although professional fact-checkers are crucial to debunking viral claims, they usually do not engage in conversations on social media. Thereby, significant effort has been made to automate the use of fact-checker material in social correction; however, no previous work has tried to integrate it with the style and pragmatics that are commonly employed in social media communication. To fill this gap, we present VerMouth, the first large-scale dataset comprising roughly 12 thousand claim-response pairs (linked to debunking articles), accounting for both SMP-style and basic emotions, two factors which have a significant role in misinformation credibility and spreading. To collect this dataset we used a technique based on an author-reviewer pipeline, which efficiently combines LLMs and human annotators to obtain high-quality data. We also provide comprehensive experiments showing how models trained on our proposed dataset have significant improvements in terms of output quality and generalization capabilities.

09:00-10:30 (East Foyer / Virtual)
**#50 Influence Scores at Scale for Efficient Language Data Sampling**
*Nikhil Anand, Joshua Tan and Maria Minakova*
Modern ML systems ingest data aggregated from diverse sources, such as synthetic, human-annotated, and live customer traffic. Understanding *which* examples are important to the performance of a learning algorithm is crucial for efficient model training. Recently, a growing body of literature has given rise to various "influence scores," which use training artifacts such as model confidence or checkpointed gradients to identify important subsets of data. However, these methods have primarily been developed in computer vision settings, and it remains unclear how well they generalize to language-based tasks using pretrained models. In this paper, we explore the applicability of influence scores in language classification tasks. We evaluate a diverse subset of these scores on the SNLI dataset by quantifying accuracy changes in response to pruning training data through random and influence-score-based sampling. We then stress-test one of the scores – "variance of gradients" (VoG) from Agarwal and Hooker (2022) – in an NLU model stack that was exposed to dynamic user speech patterns in a voice assistant type of setting. Our experiments demonstrate that in many cases, encoder-based language models can be fine-tuned on roughly 50% of the original data without degradation in performance metrics. Along the way, we summarize lessons learned from applying out-of-the-box implementations of influence scores, quantify the effects of noisy and class-imbalanced data, and offer recommendations on score-based sampling for better accuracy and training efficiency.

09:00-10:30 (East Foyer / Virtual)
**#51 Optimizing Retrieval-augmented Reader Models via Token Elimination**
*Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan and Moshe Wasserblat*
Fusion-in-Decoder (FiD) is an effective retrieval-augmented language model applied across a variety of open-domain tasks, such as question answering, fact checking, etc. In FiD, supporting passages are first retrieved and then processed using a generative model (Reader), which can cause a significant bottleneck in decoding time, particularly with long outputs. In this work, we analyze the contribution and necessity of all the retrieved passages to the performance of reader models, and propose eliminating some of the retrieved information, at the token level, that might not contribute essential information to the answer generation process. We demonstrate that our method can reduce run-time by up to 62.2%, with only a 2% reduction in performance, and in some cases, even improve the performance results.

09:00-10:30 (East Foyer / Virtual)
**#52 SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation**
*Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das and Ankur P Parikh*
Reliable automatic evaluation of summarization systems is challenging due to the multifaceted and subjective nature of the task. This is especially the case for languages other than English, where human evaluations are scarce. In this work, we introduce SEAHORSE, a dataset for multilingual, multifaceted summarization evaluation. SEAHORSE consists of 96K summaries with human ratings along 6 dimensions of text quality: comprehensibility, repetition, grammar, attribution, main ideas, and conciseness, covering 6 languages, 9 systems, and 4 datasets. As a result of its size and scope, SEAHORSE can serve both as a benchmark to evaluate learnt metrics, as well as a large-scale resource for training such metrics. We show that metrics trained with SEAHORSE achieve strong performance on the out-of-domain meta-evaluation benchmarks TRUE (Honovich et al., 2022) and mFACE (Aharoni et al., 2022). We make the SEAHORSE dataset and metrics publicly available for future research on multilingual and multifaceted summarization evaluation.

09:00-10:30 (East Foyer / Virtual)
**#53 A Self-enhancement Multitask Framework for Unsupervised Aspect Category Detection**
*Thi-Nhung Nguyen, Hoang Ngo, Kiem-Hieu Nguyen and Tuan-Dung Cao*
Our work addresses the problem of unsupervised Aspect Category Detection using a small set of seed words. Recent works have focused on learning embedding spaces for seed words and sentences to establish similarities between sentences and aspects. However, aspect representations are limited by the quality of initial seed words, and model performances are compromised by noise. To mitigate this limitation, we propose a simple framework that automatically enhances the quality of initial seed words and selects high-quality sentences for training instead of using the entire dataset. Our main concepts are to add a number of seed words to the initial set and to treat the task of noise resolution as a task of augmenting data for a low-resource task. In addition, we jointly train Aspect Category Detection with Aspect Term Extraction and Aspect Term Polarity to further enhance performance. This approach facilitates shared representation learning, allowing Aspect Category Detection to benefit from the additional guidance offered by other tasks. Extensive experiments demonstrate that our framework surpasses strong baselines on standard datasets.

09:00-10:30 (East Foyer / Virtual)
**#54 Large Language Models are biased to overestimate profoundness**
*Eugenio Herrera-Berg, Tomás Vergara Browne, Pablo León-Villagrá, Marc-Lluís Vives and Cristian Buc Calderon*
Recent advancements in natural language processing by large language models (LLMs), such as GPT-4, have been suggested to approach Artificial General Intelligence. And yet, it is still under dispute whether LLMs possess similar reasoning abilities to humans. This study evaluates GPT-4 and various other LLMs in judging the profoundness of mundane, motivational, and pseudo-profound statements. We found a significant statement-to-statement correlation between the LLMs and humans, irrespective of the type of statements and the prompting technique used. However, LLMs systematically overestimate the profoundness of nonsensical statements, with the exception of Tk-instruct, which uniquely underestimates the profoundness of statements. Only few-shot learning prompts, as opposed to chain-of-thought prompting,

draw LLMs ratings closer to humans. Furthermore, this work provides insights into the potential biases induced by Reinforcement Learning from Human Feedback (RLHF), inducing an increase in the bias to overestimate the profoundness of statements.

09:00-10:30 (East Foyer / Virtual)
**#55 NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation**
*Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang and Smaranda Muresan*
Social norms fundamentally shape interpersonal communication. We present NormDial, a high-quality dyadic dialogue dataset with turn-by-turn annotations of social norm adherences and violations for Chinese and American cultures. Introducing the task of social norm observance detection, our dataset is synthetically generated in both Chinese and English using a human-in-the-loop pipeline by prompting large language models with a small collection of expert-annotated social norms. We show that our generated dialogues are of high quality through human evaluation and further evaluate the performance of existing large language models on this task. Our findings point towards new directions for understanding the nuances of social norms as they manifest in conversational contexts that span across languages and cultures.

09:00-10:30 (East Foyer / Virtual)
**#56 CLAIR: Evaluating Image Captions with Large Language Models**
*David Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell and John Canny*
The evaluation of machine-generated image captions poses an interesting yet persistent challenge. Effective evaluation measures must consider numerous dimensions of similarity, including semantic relevance, visual structure, object interactions, caption diversity, and specificity. Existing highly-engineered measures attempt to capture specific aspects, but fall short in providing a holistic score that aligns closely with human judgments. Here, we propose CLAIR, a novel method that leverages the zero-shot language modeling capabilities of large language models (LLMs) to evaluate candidate captions. In our evaluations, CLAIR demonstrates a stronger correlation with human judgments of caption quality compared to existing measures. Notably, on Flickr8K-Expert, CLAIR achieves relative correlation improvements over SPICE of 39.6% and over image-augmented methods such as RefCLIP-S of 18.3%. Moreover, CLAIR provides noisily interpretable results by allowing the language model to identify the underlying reasoning behind its assigned score.

09:00-10:30 (East Foyer / Virtual)
**#57 AMR Parsing is Far from Solved: GrAPES, the Granular AMR Parsing Evaluation Suite**
*Jonas Groschwitz, Shay B Cohen, Lucia Donatelli and Meaghan Fowlie*
We present the Granular AMR Parsing Evaluation Suite (GrAPES), a challenge set for Abstract Meaning Representation (AMR) parsing with accompanying evaluation metrics. AMR parsers now obtain high scores on the standard AMR evaluation metric Smatch, close to or even above reported inter-annotator agreement. But that does not mean that AMR parsing is solved; in fact, human evaluation in previous work indicates that current parsers still quite frequently make errors on node labels or graph structure that substantially distort sentence meaning. Here, we provide an evaluation suite that tests AMR parsers on a range of phenomena of practical, technical, and linguistic interest. Our 36 categories range from seen and unseen labels, to structural generalization, to coreference. GrAPES reveals in depth the abilities and shortcomings of current AMR parsers.

09:00-10:30 (East Foyer / Virtual)
**#58 From Heuristic to Analytic: Cognitively Motivated Strategies for Coherent Physical Commonsense Reasoning**
*Zheyuan Zhang, Shane Storks, Fengyuan Hu, Sungryull Sohn, Moontae Lee, Honglak Lee and Joyce Chai*
Pre-trained language models (PLMs) have shown impressive performance in various language tasks. However, they are prone to spurious correlations, and often generate illusory information. In real-world applications, PLMs should justify decisions with formalized, coherent reasoning chains, but this challenge remains under-explored. Cognitive psychology theorizes that humans are capable of utilizing fast and intuitive *heuristic* thinking to make decisions based on past experience, then rationalizing the decisions through slower and deliberative *analytic* reasoning. We incorporate these interlinked dual processes in fine-tuning and in-context learning with PLMs, applying them to two language understanding tasks that require coherent physical commonsense reasoning. We show that our proposed Heuristic-Analytic Reasoning (HAR) strategies drastically improve the coherence of rationalizations for model decisions, yielding state-of-the-art results on Tiered Reasoning for Intuitive Physics (TRIP). We also find that this improved coherence is a direct result of more faithful attention to relevant language context in each step of reasoning. Our findings suggest that human-like reasoning strategies can effectively improve the coherence and reliability of PLM reasoning.

09:00-10:30 (East Foyer / Virtual)
**#59 MedEval: A Multi-Level, Multi-Task, and Multi-Domain Medical Benchmark for Language Model Evaluation**
*Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y Chang, Amilcare Gentili, Julian McAuley and Chun-Nan Hsu*
Curated datasets for healthcare are often limited due to the need of human annotations from experts. In this paper, we present MedEval, a multi-level, multi-task, and multi-domain medical benchmark to facilitate the development of language models for healthcare. MedEval is comprehensive and consists of data from several healthcare systems and spans 35 human body regions from 8 examination modalities. With 22,779 collected sentences and 21,228 reports, we provide expert annotations at multiple levels, offering a granular potential usage of the data and supporting a wide range of tasks. Moreover, we systematically evaluated 10 generic and domain-specific language models under zero-shot and finetuning settings, from domain-adapted baselines in healthcare to general-purposed state-of-the-art large language models (e.g., ChatGPT). Our evaluations reveal varying effectiveness of the two categories of language models across different tasks, from which we notice the importance of instruction tuning for few-shot usage of large language models. Our investigation paves the way toward benchmarking language models for healthcare and provides valuable insights into the strengths and limitations of adopting large language models in medical domains, informing their practical applications and future advancements.

09:00-10:30 (East Foyer / Virtual)
**#60 Calc-X and Calcformers: Empowering Arithmetical Chain-of-Thought through Interaction with Symbolic Systems**
*Marek Kadlčík, Michal Štefánik, Ondrej Sotolar and Vlastimil Martinek*
Despite outstanding performance in many tasks, language models are notoriously inclined to make factual errors in tasks requiring arithmetic computation. We address this deficiency by creating Calc-X, a collection of datasets that demonstrates the appropriate use of a calculator in reasoning chains. Calc-X is suitable for teaching language models to offload computations to a symbolic system. We survey and unify several existing chain-of-thought datasets into a proposed format, resulting in a standard collection of over 300,000 samples requiring arithmetic reasoning. Finally, we use the new Calc-X collection to train open-source calculator-using models and show that these models approximately double the accuracy of generating correct results compared to vanilla language model baselines.

09:00-10:30 (East Foyer / Virtual)
**#61 Efficient Classification of Long Documents via State-Space Models**
*Peng Lu, Suyuchen Wang, Mehdi Rezagholizadeh, Bang Liu and Ivan Kobyzev*
Transformer-based models have achieved state-of-the-art performance on numerous NLP applications. However, long documents which are prevalent in real-world scenarios cannot be efficiently processed by transformers with the vanilla self-attention module due to their quadratic

computation complexity and limited length extrapolation ability. Instead of tackling the computation difficulty for self-attention with sparse or hierarchical structures, in this paper, we investigate the use of State-Space Models (SSMs) for long document classification tasks. We conducted extensive experiments on six long document classification datasets, including binary, multi-class, and multi-label classification, comparing SSMs (with and without pre-training) to self-attention-based models. We also introduce the SSM-pooler model and demonstrate that it achieves comparable performance while being on average 36% more efficient. Additionally our method exhibits higher robustness to the input noise even in the extreme scenario of 40%.

09:00-10:30 (East Foyer / Virtual)
**#62 Larger Probes Tell a Different Story: Extending Psycholinguistic Datasets Via In-Context Learning**
*Namrata Shivagunde, Vladislav Lialin and Anna Rumshisky*
Language model probing is often used to test specific capabilities of models. However, conclusions from such studies may be limited when the probing benchmarks are small and lack statistical power. In this work, we introduce new, larger datasets for negation (NEG-1500-SIMP) and role reversal (ROLE-1500) inspired by psycholinguistic studies. We dramatically extend existing NEG-136 and ROLE-88 benchmarks using GPT3, increasing their size from 18 and 44 sentence pairs to 750 each. We also create another version of extended negation dataset (NEG-1500-SIMP-TEMP), created using template-based generation. It consists of 770 sentence pairs. We evaluate 22 models on the extended datasets, seeing model performance dip 20-57% compared to the original smaller benchmarks. We observe high levels of negation sensitivity in models like BERT and ALBERT demonstrating that previous findings might have been skewed due to smaller test sets. Finally, we observe that while GPT3 has generated all the examples in ROLE-1500 is only able to solve 24.6% of them during probing. The datasets and code are available on Github.

09:00-10:30 (East Foyer / Virtual)
**#63 ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness**
*Archiki Prasad, Swarnadeep Saha, Xiang Zhou and Mohit Bansal*
Multi-step reasoning ability is fundamental to many natural language tasks, yet it is unclear what constitutes a good reasoning chain and how to evaluate them. Most existing methods focus solely on whether the reasoning chain leads to the correct conclusion, but this answer-oriented view may confound reasoning quality with other spurious shortcuts to predict the answer. To bridge this gap, we evaluate reasoning chains by viewing them as informal proofs that derive the final answer. Specifically, we propose ReCEval (Reasoning Chain Evaluation), a framework that evaluates reasoning chains via two key properties: (1) correctness, i.e., each step makes a valid inference based on information contained within the step, preceding steps, and input context, and (2) informativeness, i.e., each step provides new information that is helpful towards deriving the generated answer. We evaluate these properties by developing metrics using natural language inference models and $\mathcal{V}$-Information. On multiple datasets, we show that ReCEval effectively identifies various error types and yields notable improvements compared to prior methods. We analyze the impact of step boundaries, and previous steps on evaluating correctness and demonstrate that our informativeness metric captures the expected flow of information in high-quality reasoning chains. Finally, we show that scoring reasoning chains based on ReCEval improves downstream task performance.

09:00-10:30 (East Foyer / Virtual)
**#64 TATA: Stance Detection via Topic-Agnostic and Topic-Aware Embeddings**
*Hans William Alexander Hanley and Zakir Durumeric*
Stance detection is important for understanding different attitudes and beliefs on the Internet. However, given that a passage's stance toward a given topic is often highly dependent on that topic, building a stance detection model that generalizes to unseen topics is difficult. In this work, we propose using contrastive learning as well as an unlabeled dataset of news articles that cover a variety of different topics to train topic-agnostic/TAG and topic-aware/TAW embeddings for use in downstream stance detection. Combining these embeddings in our full TATA model, we achieve state-of-the-art performance across several public stance detection datasets (0.771 $F_1$-score on the Zero-shot VAST dataset). We release our code and data at https://github.com/hanshanley/tata.

09:00-10:30 (East Foyer / Virtual)
**#65 Fast and Robust Early-Exiting Framework for Autoregressive Language Models with Synchronized Parallel Decoding**
*Sangmin Bae, Jongwoo Ko, Hwanjun Song and Se-Young Yun*
To tackle the high inference latency exhibited by autoregressive language models, previous studies have proposed an early-exiting framework that allocates adaptive computation paths for each token based on the complexity of generating the subsequent token. However, we observed several shortcomings, including performance degradation caused by a state copying mechanism or numerous exit paths, and sensitivity to exit confidence thresholds. Consequently, we propose a Fast and Robust Early-Exiting (FREE) framework, which incorporates a shallow-deep module and a synchronized parallel decoding. Our framework enables faster inference by synchronizing the decoding process of the current token with previously stacked early-exited tokens. Furthermore, as parallel decoding allows us to observe predictions from both shallow and deep models, we present a novel adaptive threshold estimator that exploits a Beta mixture model to determine suitable confidence thresholds. We empirically demonstrated the superiority of our proposed framework on extensive generation tasks.

09:00-10:30 (East Foyer / Virtual)
**#66 Simple Temporal Adaptation to Changing Label Sets: Hashtag Prediction via Dense KNN**
*Niloofar Mireshghallah, Nikolai Vogler, Junxian He, Omar Florez, Ahmed El-Kishky and Taylor Berg-Kirkpatrick*
User-generated social media data is constantly changing as new trends influence online discussion and personal information is deleted due to privacy concerns. However, traditional NLP models rely on fixed training datasets, which means they are unable to adapt to temporal change—both test distribution shift and deleted training data—without frequent, costly re-training. In this paper, we study temporal adaptation through the task of longitudinal hashtag prediction and propose a non-parametric dense retrieval technique, which does not require re-training, as a simple but effective solution. In experiments on a newly collected, publicly available, year-long Twitter dataset exhibiting temporal distribution shift, our method improves by 64% over the best static parametric baseline while avoiding costly gradient-based re-training. Our approach is also particularly well-suited to dynamically deleted user data in line with data privacy laws, with negligible computational cost/performance loss.

09:00-10:30 (East Foyer / Virtual)
**#67 DUnE: Dataset for Unified Editing**
*Afra Feyza Akyürek, Eric L Pan, Garry Kuwanto and Derry Wijaya*
Even the most advanced language models remain susceptible to errors necessitating to modify these models without initiating a comprehensive retraining process. Model editing refers to the modification of a model's knowledge or representations in a manner that produces the desired outcomes. Prior research primarily centered around editing factual data e.g. "Messi plays for Inter Miami" confining the definition of an edit to a knowledge triplet i.e. (subject, object, relation). However, as the applications of language models expand, so do the diverse ways in which we wish to edit and refine their outputs. In this study, we broaden the scope of the editing problem to include an array of editing cases such as debiasing and rectifying reasoning errors and define an edit as any natural language expression that solicits a change in the model's outputs. We are introducing DUnE, an editing benchmark where edits are natural language sentences and propose that DUnE presents a challenging yet

relevant task. To substantiate this claim, we conduct an extensive series of experiments testing various editing approaches to address DUnE, demonstrating their respective strengths and weaknesses. We argue that retrieval-augmented language modeling can outperform specialized editing techniques and neither set of approaches has fully solved the generalized editing problem covered by our benchmark.

09:00-10:30 (East Foyer / Virtual)
**#68 Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA**
*David Heineman, Yao Dou, Mounica Maddela and Wei Xu*
Large language models (e.g., GPT-4) are uniquely capable of producing highly rated text simplification, yet current human evaluation methods fail to provide a clear understanding of systems' specific strengths and weaknesses. To address this limitation, we introduce SALSA, an edit-based human annotation framework that enables holistic and fine-grained text simplification evaluation. We develop twenty one linguistically grounded edit types, covering the full spectrum of success and failure across dimensions of conceptual, syntactic and lexical simplicity. Using SALSA, we collect 19K edit annotations on 840 simplifications, revealing discrepancies in the distribution of simplification strategies performed by fine-tuned models, prompted LLMs and humans, and find GPT-3.5 performs more quality edits than humans, but still exhibits frequent errors. Using our fine-grained annotations, we develop LENS-SALSA, a reference-free automatic simplification metric, trained to predict sentence- and word-level quality simultaneously. Additionally, we introduce word-level quality estimation for simplification and report promising baseline results. Our data, new metric, and annotation toolkit are available at https://salsa-eval.com.

09:00-10:30 (East Foyer / Virtual)
**#69 Token Prediction as Implicit Classification to Identify LLM-Generated Text**
*Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh and Bhiksha Raj*
This paper introduces a novel approach for identifying the possible large language models (LLMs) involved in text generation. Instead of adding an additional classification layer to a base LM, we reframe the classification task as a next-token prediction task and directly fine-tune the base LM to perform it. We utilize the Text-to-Text Transfer Transformer (T5) model as the backbone for our experiments. We compared our approach to the more direct approach of utilizing hidden states for classification. Evaluation shows the exceptional performance of our method in the text classification task, highlighting its simplicity and efficiency. Furthermore, interpretability studies on the features extracted by our model reveal its ability to differentiate distinctive writing styles among various LLMs even in the absence of an explicit classifier. We also collected a dataset named OpenLLMText, containing approximately 340k text samples from human and LLMs, including GPT3.5, PaLM, LLaMA, and GPT2.

09:00-10:30 (East Foyer / Virtual)
**#70 Are All Steps Equally Important? Benchmarking Essentiality Detection in Event Processes**
*Haoyu Wang, Hongming Zhang, Yueguan Wang, Yuqian Deng, Muhao Chen and Dan Roth*
Natural language often describes events in different granularities, such that more coarse-grained (goal) events can often be decomposed into fine-grained sequences of (step) events. A critical but overlooked challenge in understanding an event process lies in the fact that the step events are not equally important to the central goal. In this paper, we seek to fill this gap by studying how well current models can understand the essentiality of different step events towards a goal event. As discussed by cognitive studies, such an ability enables the machine to mimic human's commonsense reasoning about preconditions and necessary efforts of daily-life tasks. Our work contributes with a high-quality corpus of (goal, step) pairs from a community guideline website WikiHow, where the steps are manually annotated with their essentiality w.r.t. the goal. The high IAA indicates that humans have a consistent understanding of the events. Despite evaluating various statistical and massive pre-trained NLU models, we observe that existing SOTA models all perform drastically behind humans, indicating the need for future investigation of this crucial yet challenging task.

09:00-10:30 (East Foyer / Virtual)
**#71 TCFLE-8: a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring**
*Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny and Thomas François*
Automated Essay Scoring (AES) aims to automatically assess the quality of essays. Automation enables large-scale assessment, improvements in consistency, reliability, and standardization. Those characteristics are of particular relevance in the context of language certification exams. However, a major bottleneck in the development of AES systems is the availability of corpora, which, unfortunately, are scarce, especially for languages other than English. In this paper, we aim to foster the development of AES for French by providing the TCFLE-8 corpus, a corpus of 6.5k essays collected in the context of the *Test de Connaissance du Français* (TCF - French Knowledge Test) certification exam. We report the strict quality procedure that led to the scoring of each essay by at least two raters according to the CEFR levels and to the creation of a balanced corpus. In addition, we describe how linguistic properties of the essays relate to the learners' proficiency in TCFLE-8. We also advance the state-of-the-art performance for the AES task in French by experimenting with two strong baselines (i.e. RoBERTa and feature-based). Finally, we discuss the challenges of AES using TCFLE-8.

09:00-10:30 (East Foyer / Virtual)
**#72 Editing Common Sense in Transformers**
*Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li, Sarah Wiegreffe and Niket Tandon*
Editing model parameters directly in Transformers makes updating open-source transformer-based models possible without re-training. However, these editing methods have only been evaluated on statements about encyclopedic knowledge with a single correct answer. Commonsense knowledge with multiple correct answers, e.g., an apple can be green or red but not transparent, has not been studied but is as essential for enhancing transformers' reliability and usefulness. In this paper, we investigate whether commonsense judgments are causally associated with localized, editable parameters in Transformers, and we provide an affirmative answer. We find that directly applying the MEMIT editing algorithm results in sub-par performance and improve it for the commonsense domain by varying edit tokens and improving the layer selection strategy, i.e., $MEMIT_{CSK}$. GPT-2 Large and XL models edited using $MEMIT_{CSK}$ outperform best-fine-tuned baselines by 10.97% and 10.73% F1 scores on PEP3k and 20Q datasets. In addition, we propose a novel evaluation dataset, $PROBE\ SET$, that contains unaffected and affected neighborhoods, affected paraphrases, and affected reasoning challenges. $MEMIT_{CSK}$ performs well across the metrics while fine-tuning baselines show significant trade-offs between unaffected and affected metrics. These results suggest a compelling future direction for incorporating feedback about common sense into Transformers through direct model editing.

09:00-10:30 (East Foyer / Virtual)
**#73 Hi Guys or Hi Folks? Benchmarking Gender-Neutral Machine Translation with the GeNTE Corpus**
*Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri and Luisa Bentivogli*
Gender inequality is embedded in our communication practices and perpetuated in translation technologies. This becomes particularly apparent when translating into grammatical gender languages, where machine translation (MT) often defaults to masculine and stereotypical representations by making undue binary gender assumptions. Our work addresses the rising demand for inclusive language by focusing head-on on gender-neutral translation from English to Italian. We start from the essentials: proposing a dedicated benchmark and exploring automated evaluation methods. First, we introduce GeNTE, a natural, bilingual test set for gender-neutral translation, whose creation was

informed by a survey on the perception and use of neutral language. Based on GeNTE, we then overview existing reference-based evaluation approaches, highlight their limits, and propose a reference-free method more suitable to assess gender-neutral translation.

09:00-10:30 (East Foyer / Virtual)
**#74 Using Artificial French Data to Understand the Emergence of Gender Bias in Transformer Language Models**
*Lina Conti and Guillaume Wisniewski*
Numerous studies have demonstrated the ability of neural language models to learn various linguistic properties without direct supervision. This work takes an initial step towards exploring the less researched topic of how neural models discover linguistic properties of words, such as gender, as well as the rules governing their usage. We propose to use an artificial corpus generated by a PCFG based on French to precisely control the gender distribution in the training data and determine under which conditions a model correctly captures gender information or, on the contrary, appears gender-biased.

09:00-10:30 (East Foyer / Virtual)
**#75 Construction Artifacts in Metaphor Identification Datasets**
*Joanne Boisson, Luis Espinosa-Anke and Jose Camacho-Collados*
Metaphor identification aims at understanding whether a given expression is used figuratively in context. However, in this paper we show how existing metaphor identification datasets can be gamed by fully ignoring the potential metaphorical expression or the context in which it occurs. We test this hypothesis in a variety of datasets and settings, and show that metaphor identification systems based on language models without complete information can be competitive with those using the full context. This is due to the construction procedures to build such datasets, which introduce unwanted biases for positive and negative classes. Finally, we test the same hypothesis on datasets that are carefully sampled from natural corpora and where this bias is not present, making these datasets more challenging and reliable.

09:00-10:30 (East Foyer / Virtual)
**#76 Combining Denoising Autoencoders with Contrastive Learning to fine-tune Transformer Models**
*Alejo Lopez-Avila and Víctor Suárez-Paniagua*
Recently, using large pre-trained Transformer models for transfer learning tasks has evolved to the point where they have become one of the flagship trends in the Natural Language Processing (NLP) community, giving rise to various outlooks such as prompt-based, adapters, or combinations with unsupervised approaches, among many others. In this work, we propose a 3-Phase technique to adjust a base model for a classification task. First, we adapt the model's signal to the data distribution by performing further training with a Denoising Autoencoder (DAE). Second, we adjust the representation space of the output to the corresponding classes by clustering through a Contrastive Learning (CL) method. In addition, we introduce a new data augmentation approach for Supervised Contrastive Learning to correct the unbalanced datasets. Third, we apply fine-tuning to delimit the predefined categories. These different phases provide relevant and complementary knowledge to the model to learn the final task. We supply extensive experimental results on several datasets to demonstrate these claims. Moreover, we include an ablation study and compare the proposed method against other ways of combining these techniques.

09:00-10:30 (East Foyer / Virtual)
**#77 Sparse Universal Transformer**
*Shawn Tan, Yikang Shen, Zhenfang Chen, Aaron Courville and Chuang Gan*
The Universal Transformer (UT) is a variant of the Transformer that shares parameters across its layers and is Turing-complete under certain assumptions. Empirical evidence also shows that UTs have better compositional generalization than Vanilla Transformers (VTs) in formal language tasks. The parameter-sharing also affords it better parameter efficiency than VTs. Despite its many advantages, most state-of-the-art NLP systems use VTs as their backbone model instead of UTs. This is mainly because scaling UT parameters is more compute and memory intensive than scaling up a VT. This paper proposes the Sparse Universal Transformer (SUT), which leverages Sparse Mixture of Experts (SMoE) to reduce UT's computation complexity while retaining its parameter efficiency and generalization ability. Experiments show that SUT combines the best of both worlds, achieving strong generalization results on formal language tasks (Logical inference and CFQ) and impressive parameter and computation efficiency on standard natural language benchmarks like WMT'14.

09:00-10:30 (East Foyer / Virtual)
**#78 Holistic Inter-Annotator Agreement and Corpus Coherence Estimation in a Large-scale Multilingual Annotation Campaign**
*Nicolas Stefanovitch and Jakub Piskorski*
In this paper we report on the complexity of persuasion technique annotation in the context of a large multilingual annotation campaign involving 6 languages and approximately 40 annotators. We highlight the techniques that appear to be difficult for humans to annotate and elaborate on our findings on the causes of this phenomenon. We introduce Holistic IAA, a new word embedding-based annotator agreement metric and we report on various experiments using this metric and its correlation with the traditional Inter Annotator Agreement (IAA) metrics. However, given somewhat limited and loose interaction between annotators, i.e., only a few annotators annotate the same document subsets, we try to devise a way to assess the coherence of the entire dataset and strive to find a good proxy for IAA between annotators tasked to annotate different documents and in different languages, for which classical IAA metrics can not be applied.

09:00-10:30 (East Foyer / Virtual)
**#79 Efficient Algorithms for Recognizing Weighted Tree-Adjoining Languages**
*Alexandra Butoi, Tim Vieira, Ryan Cotterell and David Chiang*
The class of tree-adjoining languages can be characterized by various two-level formalisms, consisting of a context-free grammar (CFG) or pushdown automaton (PDA) controlling another CFG or PDA. These four formalisms are equivalent to tree-adjoining grammars (TAG), linear indexed grammars (LIG), pushdown-adjoining automata (PAA), and embedded pushdown automata (EPDA). We define semiring-weighted versions of the above two-level formalisms, and we design new algorithms for computing their stringsums (the weight of all derivations of a string) and allsums (the weight of all derivations). From these, we also immediately obtain stringsum and allsum algorithms for TAG, LIG, PAA, and EPDA. For LIG, our algorithm is more time-efficient by a factor of $\mathcal{O}(n|\mathcal{N}|)$ (where $n$ is the string length and $|\mathcal{N}|$ is the size of the nonterminal set) and more space-efficient by a factor of $\mathcal{O}(|\Gamma|)$ (where $\Gamma$ is the size of the stack alphabet) than the algorithm of Vijay-Shanker and Weir (1989). For EPDA, our algorithm is both more space-efficient and time-efficient than the algorithm of Alonso et al. (2001) by factors of $\mathcal{O}(|\Gamma|^2)$ and $\mathcal{O}(|\Gamma|^3)$, respectively. Finally, we give the first PAA stringsum and allsum algorithms.

09:00-10:30 (East Foyer / Virtual)
**#80 Towards Robust Pruning: An Adaptive Knowledge-Retention Pruning Strategy for Language Models**
*Jianwei Li, Qi Lei, Wei Cheng and Dongkuan Xu*
The pruning objective has recently extended beyond accuracy and sparsity to robustness in language models. Despite this, existing methods struggle to enhance robustness against adversarial attacks when continually increasing model sparsity and require a retraining process. As humans step into the era of large language models, these issues become increasingly prominent. This paper proposes that the robustness of language models is proportional to the extent of pre-trained knowledge they encompass. Accordingly, we introduce a post-training pruning strategy designed to faithfully replicate the embedding space and feature space of dense language models, aiming to conserve more pre-trained

knowledge during the pruning process. In this setup, each layer's reconstruction error not only originates from itself but also includes cumulative error from preceding layers, followed by an adaptive rectification. Compared to other state-of-art baselines, our approach demonstrates a superior balance between accuracy, sparsity, robustness, and pruning cost with BERT on datasets SST2, IMDB, and AGNews, marking a significant stride towards robust pruning in language models.

09:00-10:30 (East Foyer / Virtual)
**#81 Memory-Based Invariance Learning for Out-of-Domain Text Classification**
*Chen Jia and Yue Zhang*
We investigate the task of out-of-domain (OOD) text classification with the aim of extending a classification model, trained on multiple source domains, to an unseen target domain. Recent studies have shown that learning invariant representations can enhance the performance of OOD generalization. However, the inherent disparity in data distribution across different domains poses challenges for achieving effective invariance learning. This study addresses this issue by employing memory augmentations. Specifically, we augment the original feature space using key-value memory and employ a meta-learning-based approach to enhance the quality of the invariant representations. Experimental results on sentiment analysis and natural language inference tasks show the effectiveness of memory-based method for invariance learning, leading to state-of-the-art performance on six datasets.

09:00-10:30 (East Foyer / Virtual)
**#82 Increasing Coverage and Precision of Textual Information in Multilingual Knowledge Graphs**
*Simone Conia, Min Li, Daniel Lee, Umar Farooq Minhas, Ihab Ilyas and Yunyao Li*
Recent work in Natural Language Processing and Computer Vision has been using textual information – e.g., entity names and descriptions – available in knowledge graphs to ground neural models to high-quality structured data. However, when it comes to non-English languages, the quantity and quality of textual information are comparatively scarce. To address this issue, we introduce the novel task of automatic Knowledge Graph Completion (KGE) and perform a thorough investigation on bridging the gap in both the quantity and quality of textual information between English and non-English languages. More specifically, we: i) bring to light the problem of increasing multilingual coverage and precision of entity names and descriptions in Wikidata; ii) demonstrate that state-of-the-art methods, namely, Machine Translation (MT), Web Search (WS), and Large Language Models (LLMs), struggle with this task; iii) present M-NTA, a novel unsupervised approach that combines MT, WS, and LLMs to generate high-quality textual information; and, iv) study the impact of increasing multilingual coverage and precision of non-English textual information in Entity Linking, Knowledge Graph Completion, and Question Answering. As part of our effort towards better multilingual knowledge graphs, we also introduce WikiKGE-10, the first human-curated benchmark to evaluate KGE approaches in 10 languages across 7 language families.

09:00-10:30 (East Foyer / Virtual)
**#83 Bootstrapping Small & High Performance Language Models with Unmasking-Removal Training Policy**
*Yahan Yang, Elior Sulem, Insup Lee and Dan Roth*
BabyBERTa, a language model trained on small-scale child-directed speech while none of the words are unmasked during training, has been shown to achieve a level of grammaticality comparable to that of RoBERTa-base, which is trained on 6,000 times more words and 15 times more parameters. Relying on this promising result, we explore in this paper the performance of BabyBERTa-based models in downstream tasks, focusing on Semantic Role Labeling (SRL) and two Extractive Question Answering tasks, with the aim of building more efficient systems that rely on less data and smaller models. We investigate the influence of these models both alone and as a starting point to larger pre-trained models, separately examining the contribution of the pre-training data, the vocabulary, and the masking policy on the downstream task performance. Our results show that BabyBERTa trained with unmasking-removal policy is a much stronger starting point for downstream tasks compared to the use of RoBERTa masking policy when 10M words are used for training and that this tendency persists, although to a lesser extent, when adding more training data.

09:00-10:30 (East Foyer / Virtual)
**#84 Introducing Rhetorical Parallelism Detection: A New Task with Datasets, Metrics, and Baselines**
*Stephen Bothwell, Justin DeBenedetto, Theresa Crnkovich, Hildegund Muller and David Chiang*
Rhetoric, both spoken and written, involves not only content but also style. One common stylistic tool is *parallelism*: the juxtaposition of phrases which have the same sequence of linguistic (*e.g.*, phonological, syntactic, semantic) features. Despite the ubiquity of parallelism, the field of natural language processing has seldom investigated it, missing a chance to better understand the nature of the structure, meaning, and intent that humans convey. To address this, we introduce the task of *rhetorical parallelism detection*. We construct a formal definition of it; we provide one new Latin dataset and one adapted Chinese dataset for it; we establish a family of metrics to evaluate performance on it; and, lastly, we create baseline systems and novel sequence labeling schemes to capture it. On our strictest metric, we attain $F_1$ scores of $0.40$ and $0.43$ on our Latin and Chinese datasets, respectively.

09:00-10:30 (East Foyer / Virtual)
**#85 ALCUNA: Large Language Models Meet New Knowledge**
*Xunjian Yin, Baizhou Huang and Xiaojun Wan*
With the rapid development of NLP, large-scale language models (LLMs) excel in various tasks across multiple domains now. However, existing benchmarks may not adequately measure these models' capabilities, especially when faced with new knowledge. In this paper, we address the lack of benchmarks to evaluate LLMs' ability to handle new knowledge, an important and challenging aspect in the rapidly evolving world. We propose an approach called KnowGen that generates new knowledge by altering existing entity attributes and relationships, resulting in artificial entities that are distinct from real-world entities. With KnowGen, we introduce a benchmark named ALCUNA to assess LLMs' abilities in knowledge understanding, differentiation, and association. We benchmark several LLMs, reveals that their performance in face of new knowledge is not satisfactory, particularly in reasoning between new and internal knowledge. We also explore the impact of entity similarity on the model's understanding of entity knowledge and the influence of contextual entities. We appeal to the need for caution when using LLMs in new scenarios or with new knowledge, and hope that our benchmarks can help drive the development of LLMs in face of new knowledge.

09:00-10:30 (East Foyer / Virtual)
**#86 Tagging-Assisted Generation Model with Encoder and Decoder Supervision for Aspect Sentiment Triplet Extraction**
*Luo Xianlong, Meng Yang and Yihao Wang*
ASTE (Aspect Sentiment Triplet Extraction) has gained increasing attention. Recent advancements in the ASTE task have been primarily driven by Natural Language Generation-based (NLG) approaches. However, most NLG methods overlook the supervision of the encoder-decoder hidden representations and fail to fully utilize the semantic information provided by the labels to enhance supervision. These limitations can hinder the extraction of implicit aspects and opinions. To address these challenges, we propose a tagging-assisted generation model with encoder and decoder supervision (TAGS), which enhances the supervision of the encoder and decoder through multiple-perspective tagging assistance and label semantic representations. Specifically, TAGS enhances the generation task by integrating an additional sequence tagging task, which improves the encoder's capability to distinguish the words of triplets. Moreover, it utilizes sequence tagging probabilities

to guide the decoder, improving the generated content's quality. Furthermore, TAGS employs a self-decoding process for labels to acquire the semantic representations of the labels and aligns the decoder's hidden states with these semantic representations, thereby achieving enhanced semantic supervision for the decoder's hidden states. Extensive experiments on various public benchmarks demonstrate that TAGS achieves state-of-the-art performance.

09:00-10:30 (East Foyer / Virtual)
**#87 CRT-QA: A Dataset of Complex Reasoning Question Answering over Tabular Data**
*Zhehao Zhang, Xitao Li, Yan Gao and Jian-Guang Lou*
Large language models (LLMs) show powerful reasoning abilities on various text-based tasks. However, their reasoning capability on structured data such as tables has not been systematically explored. In this work, we first establish a comprehensive taxonomy of reasoning and operation types for tabular data analysis. Then, we construct a complex reasoning QA dataset over tabular data, named CRT-QA dataset (Complex Reasoning QA over Tabular data), with the following unique features: (1) it is the first Table QA dataset with multi-step operation and informal reasoning; (2) it contains fine-grained annotations on questions' directness, composition types of sub-questions, and human reasoning paths which can be used to conduct a thorough investigation on LLMs' reasoning ability; (3) it contains a collection of unanswerable and indeterminate questions that commonly arise in real-world situations. We further introduce an efficient and effective tool-augmented method, named ARC (Auto-exemplar-guided Reasoning with Code), to use external tools such as Pandas to solve table reasoning tasks without handcrafted demonstrations. The experiment results show that CRT-QA presents a strong challenge for baseline methods and ARC achieves the best result.

09:00-10:30 (East Foyer / Virtual)
**#88 EtiCor: Corpus for Analyzing LLMs for Etiquettes**
*Ashutosh Dwivedi, Pradhyumna Lavania and Ashutosh Modi*
Etiquettes are an essential ingredient of day-to-day interactions among people. Moreover, etiquettes are region-specific, and etiquettes in one region might contradict those in other regions. In this paper, we propose EtiCor, an Etiquettes Corpus, having texts about social norms from five different regions across the globe. The corpus provides a test bed for evaluating LLMs for knowledge and understanding of region-specific etiquettes. Additionally, we propose the task of Etiquette Sensitivity. We experiment with state-of-the-art LLMs (Delphi, Falcon40B, and GPT-3.5). Initial results indicate that LLMs, mostly fail to understand etiquettes from regions from non-Western world.

09:00-10:30 (East Foyer / Virtual)
**#89 BRAINTEASER: Lateral Thinking Puzzles for Large Language Models**
*Yifan Jiang, Filip Ilievski, Kaixin Ma and Zhivar Sourati*
The success of language models has inspired the NLP community to attend to tasks that require implicit and complex reasoning, relying on human-like commonsense mechanisms. While such vertical thinking tasks have been relatively popular, lateral thinking puzzles have received little attention. To bridge this gap, we devise BrainTeaser: a multiple-choice Question Answering task designed to test the model's ability to exhibit lateral thinking and defy default commonsense associations. We design a three-step procedure for creating the first lateral thinking benchmark, consisting of data collection, distractor generation, and generation of adversarial examples, leading to 1,100 puzzles with high-quality annotations. To assess the consistency of lateral reasoning by models, we enrich BrainTeaser based on a semantic and contextual reconstruction of its questions. Our experiments with state-of-the-art instruction- and commonsense language models reveal a significant gap between human and model performance, which is further widened when consistency across adversarial formats is considered. We make all of our code and data available to stimulate work on developing and evaluating lateral thinking models.

09:00-10:30 (East Foyer / Virtual)
**#90 SciRepEval: A Multi-Format Benchmark for Scientific Document Representations**
*Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey and Sergey Feldman*
Learned representations of scientific documents can serve as valuable input features for downstream tasks without further fine-tuning. However, existing benchmarks for evaluating these representations fail to capture the diversity of relevant tasks. In response, we introduce SciRepEval, the first comprehensive benchmark for training and evaluating scientific document representations. It includes 24 challenging and realistic tasks, 8 of which are new, across four formats: classification, regression, ranking and search. We then use this benchmark to study and improve the generalization ability of scientific document representation models. We show how state-of-the-art models like SPECTER and SciNCL struggle to generalize across the task formats, and that simple multi-task training fails to improve them. However, a new approach that learns multiple embeddings per document, each tailored to a different format, can improve performance. We experiment with task-format-specific control codes and adapters and find they outperform the existing single-embedding state-of-the-art by over 2 points absolute. We release the resulting family of multi-format models, called SPECTER2, for the community to use and build on.

09:00-10:30 (East Foyer / Virtual)
**#91 Pushdown Layers: Encoding Recursive Structure in Transformer Language Models**
*Shikhar Murty, Pratyusha Sharma, Jacob Andreas and Christopher D Manning*
Recursion is a prominent feature of human language, and fundamentally challenging for self-attention due to the lack of an explicit recursive-state tracking mechanism. Consequently, Transformer language models poorly capture long-tail recursive structure and exhibit sample-inefficient syntactic generalization. This work introduces Pushdown Layers, a new self-attention layer that models recursive state via a stack tape that tracks estimated depths of every token in an incremental parse of the observed prefix. Transformer LMs with Pushdown Layers are syntactic language models that autoregressively and synchronously update this stack tape as they predict new tokens, in turn using the stack tape to softly modulate attention over tokens—for instance, learning to "skip" over closed constituents. When trained on a corpus of strings annotated with silver constituency parses, Transformers equipped with Pushdown Layers achieve dramatically better and 3-5x more sample-efficient syntactic generalization, while maintaining similar perplexities. Pushdown Layers are a drop-in replacement for standard self-attention. We illustrate this by finetuning GPT2-medium with Pushdown Layers on an automatically parsed WikiText-103, leading to improvements on several GLUE text classification tasks.

09:00-10:30 (East Foyer / Virtual)
**#92 LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers**
*Theo X. Olausson, Alex Gu, Ben Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum and Roger P. Levy*
Logical reasoning, i.e., deductively inferring the truth value of a conclusion from a set of premises, is an important task for artificial intelligence with wide potential impacts on science, mathematics, and society. While many prompting-based strategies have been proposed to enable Large Language Models (LLMs) to do such reasoning more effectively, they still appear unsatisfactory, often failing in subtle and unpredictable ways. In this work, we investigate the validity of instead reformulating such tasks as modular neurosymbolic programming, which we call LINC: Logical Inference via Neurosymbolic Computation. In LINC, the LLM acts as a semantic parser, translating premises and conclusions from natural language to expressions in first-order logic. These expressions are then offloaded to an external theorem prover, which symbolically performs deductive inference. Leveraging this approach, we observe significant performance gains on FOLIO and a balanced subset of ProofWriter for three different models in nearly all experimental conditions we evaluate. On ProofWriter, augmenting the

comparatively small open-source StarCoder+ (15.5B parameters) with LINC even outperforms GPT-3.5 and GPT-4 with Chain-of-Thought (CoT) prompting by an absolute 38% and 10%, respectively. When used with GPT-4, LINC scores 26% higher than CoT on ProofWriter while performing comparatively on FOLIO. Further analysis reveals that although both methods on average succeed roughly equally often on this dataset, they exhibit distinct and complementary failure modes. We thus provide promising evidence for how logical reasoning over natural language can be tackled through jointly leveraging LLMs alongside symbolic provers. All corresponding code is publicly available.

09:00-10:30 (East Foyer / Virtual)
**#93 Enhancing Computation Efficiency in Large Language Models through Weight and Activation Quantization**
*Janghwan Lee, Minsoo Kim, Seungcheol Baek, Seok Joong Hwang, Wonyong Sung and Jungwook Choi*
Large Language Models (LLMs) are proficient in natural language processing tasks, but their deployment is often restricted by extensive parameter sizes and computational demands. This paper focuses on post-training quantization (PTQ) in LLMs, specifically 4-bit weight and 8-bit activation (W4A8) quantization, to enhance computational efficiency—a topic less explored compared to weight-only quantization. We present two innovative techniques: activation-quantization-aware scaling (AQAS) and sequence-length-aware calibration (SLAC) to enhance PTQ by considering the combined effects on weights and activations and aligning calibration sequence lengths to target tasks. Moreover, we introduce dINT, a hybrid data format combining integer and denormal representations, to address the underflow issue in W4A8 quantization, where small values are rounded to zero. Through rigorous evaluations of LLMs, including OPT and LLaMA, we demonstrate that our techniques significantly boost task accuracies to levels comparable with full-precision models. By developing arithmetic units compatible with dINT, we further confirm that our methods yield a 2$\times$ hardware efficiency improvement compared to 8-bit integer MAC unit.

09:00-10:30 (East Foyer / Virtual)
**#94 M2DF: Multi-grained Multi-curriculum Denoising Framework for Multimodal Aspect-based Sentiment Analysis**
*Fei Zhao, Chunhui Li, Zhen Wu, Yawen Ouyang, Jianbing Zhang and Xinyu Dai*
Multimodal Aspect-based Sentiment Analysis (MABSA) is a fine-grained Sentiment Analysis task, which has attracted growing research interests recently. Existing work mainly utilizes image information to improve the performance of MABSA task. However, most of the studies overestimate the importance of images since there are many noise images unrelated to the text in the dataset, which will have a negative impact on model learning. Although some work attempts to filter low-quality noise images by setting thresholds, relying on thresholds will inevitably filter out a lot of useful image information. Therefore, in this work, we focus on whether the negative impact of noisy images can be reduced without modifying the data. To achieve this goal, we borrow the idea of Curriculum Learning and propose a Multi-grained Multi-curriculum Denoising Framework (M2DF), which can achieve denoising by adjusting the order of training data. Extensive experimental results show that our framework consistently outperforms state-of-the-art work on three sub-tasks of MABSA.

09:00-10:30 (East Foyer / Virtual)
**#95 Sparse Low-rank Adaptation of Pre-trained Language Models**
*Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu and Maosong Sun*
Fine-tuning pre-trained large language models in a parameter-efficient manner is widely studied for its effectiveness and efficiency. The popular method of low-rank adaptation (LoRA) offers a notable approach, hypothesizing that the adaptation process is intrinsically low-dimensional. Although LoRA has demonstrated commendable performance, it is implemented with a fixed and unalterable intrinsic rank that might not always be the ideal choice. Recognizing the need for more flexible adaptation, we extend the methodology of LoRA to an innovative approach we call sparse low-rank adaptation (SoRA) that enables dynamic adjustments to the intrinsic rank during the adaptation process. We achieve this through the incorporation of a gate unit optimized with proximal gradient method in the training stage, controlling the cardinality of rank under the sparsity of the gate. In the subsequent inference stage, we eliminate the parameter blocks corresponding to the zeroed-out ranks, to reduce each SoRA module back to a concise yet rank-optimal LoRA. Our approach strengthens the representation power of LoRA by initializing it with a higher rank, while efficiently taming a temporarily increased number of parameters via updating in a sparse way. We further introduce a sparsifying scheduler for SoRA, aiming to examine the impact of the number of non-zero parameters on the model's memorization and generalization. Our experimental results demonstrate that SoRA can outperform other baselines even with 70% retained parameters and 70% training time.

09:00-10:30 (East Foyer / Virtual)
**#96 VECHR: A Dataset for Explainable and Robust Classification of Vulnerability Type in the European Court of Human Rights**
*Shanshan Xu, Leon Staufer, Santosh T.Y.S.S, Oana Ichim, Corina Heri and Matthias Grabmair*
Recognizing vulnerability is crucial for understanding and implementing targeted support to empower individuals in need. This is especially important at the European Court of Human Rights (ECtHR), where the court adapts Convention standards to meet actual individual needs and thus to ensure effective human rights protection. However, the concept of vulnerability remains elusive at the ECtHR and no prior NLP research has dealt with it. To enable future research in this area, we present VECHR, a novel expert-annotated multi-label dataset comprising of vulnerability type classification and explanation rationale. We benchmark the performance of state-of-the-art models on VECHR from both prediction and explainability perspective. Our results demonstrate the challenging nature of task with lower prediction performance and limited agreement between models and experts. Further, we analyze the robustness of these models in dealing with out-of-domain (OOD) data and observe overall limited performance. Our dataset poses unique challenges offering a significant room for improvement regarding performance, explainability and robustness.

09:00-10:30 (East Foyer / Virtual)
**#97 An Exploration of Left-Corner Transformations**
*Andreas Opedal, Eleftheria Tsipidi, Tiago Pimentel, Ryan Cotterell and Tim Vieira*
The left-corner transformation (Rosenkrantz and Lewis, 1970) is used to remove left recursion from context-free grammars, which is an important step towards making the grammar parsable top-down with simple techniques. This paper generalizes prior left-corner transformations to support semiring-weighted production rules and to provide finer-grained control over which left corners may be moved. Our generalized left-corner transformation (GLCT) arose from unifying the left-corner transformation and speculation transformation (Eisner and Blatz, 2007), originally for logic programming. Our new transformation and speculation define equivalent weighted languages. Yet, their derivation trees are structurally different in an important way: GLCT replaces left recursion with right recursion, and speculation does not. We also provide several technical results regarding the formal relationships between the outputs of GLCT, speculation, and the original grammar. Lastly, we empirically investigate the efficiency of GLCT for left-recursion elimination from grammars of nine languages. Code: https://github.com/rycolab/left-corner

09:00-10:30 (East Foyer / Virtual)
**#98 Is ChatGPT a General-Purpose Natural Language Processing Task Solver?**
*Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga and Diyi Yang*
Spurred by advancements in scale, large language models (LLMs) have demonstrated the ability to perform a variety of natural language processing (NLP) tasks zero-shot—i.e., without adaptation on downstream data. Recently, the debut of ChatGPT has drawn a great deal of attention from the natural language processing (NLP) community due to the fact that it can generate high-quality responses to human input

and self-correct previous mistakes based on subsequent conversations. However, it is not yet known whether ChatGPT can serve as a generalist model that can perform many NLP tasks zero-shot. In this work, we empirically analyze the zero-shot learning ability of ChatGPT by evaluating it on 20 popular NLP datasets covering 7 representative task categories. With extensive empirical studies, we demonstrate both the effectiveness and limitations of the current version of ChatGPT. We find that ChatGPT performs well on many tasks favoring reasoning capabilities (e.g., arithmetic reasoning) while it still faces challenges when solving specific tasks such as sequence tagging. We additionally provide in-depth analysis through qualitative case studies.

09:00-10:30 (East Foyer / Virtual)
**#99 Adapt in Contexts: Retrieval-Augmented Domain Adaptation via In-Context Learning**
*Quanyu Long, Wenya Wang and Sinno Jialin Pan*
Large language models (LLMs) have showcased their capability with few-shot inference known as in-context learning. However, in-domain demonstrations are not always readily available in real scenarios, leading to cross-domain in-context learning. Besides, LLMs are still facing challenges in long-tail knowledge in unseen and unfamiliar domains. The above limitations demonstrate the necessity of Unsupervised Domain Adaptation (UDA). In this paper, we study the UDA problem under an in-context learning setting to adapt language models from the source domain to the target domain without any target labels. The core idea is to retrieve a subset of cross-domain elements that are the most similar to the query, and elicit language model to adapt in an in-context manner by learning both target domain distribution and the discriminative task signal simultaneously with the augmented cross-domain in-context examples. We devise different prompting and training strategies, accounting for different LM architectures to learn the target distribution via language modeling. With extensive experiments on Sentiment Analysis (SA) and Named Entity Recognition (NER) tasks, we thoroughly study the effectiveness of ICL for domain transfer and demonstrate significant improvements over baseline models.

09:00-10:30 (East Foyer / Virtual)
**#100 SOUL: Towards Sentiment and Opinion Understanding of Language**
*Yue Deng, Wenxuan Zhang, Sinno Jialin Pan and Lidong Bing*
Sentiment analysis is a well-established natural language processing task, with sentiment polarity classification being one of its most popular and representative tasks. However, despite the success of pre-trained language models in this area, they often fall short of capturing the broader complexities of sentiment analysis. To address this issue, we propose a new task called Sentiment and Opinion Understanding of Language (SOUL). SOUL aims to evaluate sentiment understanding through two subtasks: Review Comprehension (RC) and Justification Generation (JG). RC seeks to validate statements that focus on subjective information based on a review text, while JG requires models to provide explanations for their sentiment predictions. To enable comprehensive evaluation, we annotate a new dataset comprising 15,028 statements from 3,638 reviews. Experimental results indicate that SOUL is a challenging task for both small and large language models, with a performance gap of up to 27% when compared to human performance. Furthermore, evaluations conducted with both human experts and GPT-4 highlight the limitations of the small language model in generating reasoning-based justifications. These findings underscore the challenging nature of the SOUL task for existing models, emphasizing the need for further advancements in sentiment analysis to address its complexities. The new dataset and code are available at https://github.com/DAMO-NLP-SG/SOUL.

09:00-10:30 (East Foyer / Virtual)
**#101 Standardizing Distress Analysis: Emotion-Driven Distress Identification and Cause Extraction (DICE) in Multimodal Online Posts**
*Gopendra Vikram Singh, Soumitra Ghosh, Atul Verma, Chetna Painkra and Asif Ekbal*
Due to its growing impact on public opinion, hate speech on social media has garnered increased attention. While automated methods for identifying hate speech have been presented in the past, they have mostly been limited to analyzing textual content. The interpretability of such models has received very little attention, despite the social and legal consequences of erroneous predictions. In this work, we present a novel problem of *Distress Identification and Cause Extraction (DICE)* from multimodal online posts. We develop a multi-task deep framework for the simultaneous detection of distress content and identify connected causal phrases from the text using emotional information. The emotional information is incorporated into the training process using a zero-shot strategy, and a novel mechanism is devised to fuse the features from the multimodal inputs. Furthermore, we introduce the first-of-its-kind *Distress and Cause annotated Multimodal (DCaM)* dataset of 20,764 social media posts. We thoroughly evaluate our proposed method by comparing it to several existing benchmarks. Empirical assessment and comprehensive qualitative analysis demonstrate that our proposed method works well on distress detection and cause extraction tasks, improving F1 and ROS scores by 1.95% and 3%, respectively, relative to the best-performing baseline. The code and the dataset can be accessed from the following link: https://www.iitp.ac.in/~ai-nlp-ml/resources.html#DICE.

09:00-10:30 (East Foyer / Virtual)
**#102 Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language**
*Michael Wiegand, Jana Kampfmeier, Elisabeth Eder and Josef Ruppenhofer*
We address the task of identifying euphemistic abuse (e.g. "You inspire me to fall asleep") paraphrasing simple explicitly abusive utterances (e.g. "You are boring"). For this task, we introduce a novel dataset that has been created via crowdsourcing. Special attention has been paid to the generation of appropriate negative (non-abusive) data. We report on classification experiments showing that classifiers trained on previous datasets are less capable of detecting such abuse. Best automatic results are obtained by a classifier that augments training data from our new dataset with automatically-generated GPT-3 completions. We also present a classifier that combines a few manually extracted features that exemplify the major linguistic phenomena constituting euphemistic abuse.

09:00-10:30 (East Foyer / Virtual)
**#103 More Than Spoken Words: Nonverbal Message Extraction and Generation**
*Dian Yu, Xiaoyang Wang, Wanshun Chen, Nan Du, Longyue Wang, Haitao Mi and Dong Yu*
Nonverbal messages (NM) such as speakers' facial expressions and speed of speech are essential for face-to-face communication, and they can be regarded as implicit knowledge as they are usually not included in existing dialogue understanding or generation tasks. This paper introduces the task of extracting NMs in written text and generating NMs for spoken text. Previous studies merely focus on extracting NMs from relatively small-scale well-structured corpora such as movie scripts wherein NMs are enclosed in parentheses by scriptwriters, which greatly decreases the difficulty of extraction. To enable extracting NMs from unstructured corpora, we annotate the first NM extraction dataset for Chinese based on novels and develop three baselines to extract single-span or multi-span NM of a target utterance from its surrounding context. Furthermore, we use the extractors to extract 749K (context, utterance, NM) triples from Chinese novels and investigate whether we can use them to improve NM generation via semi-supervised learning. Experimental results demonstrate that the automatically extracted triples can serve as high-quality augmentation data of clean triples extracted from scripts to generate more relevant, fluent, valid, and factually consistent NMs than the purely supervised generator, and the resulting generator can in turn help Chinese dialogue understanding tasks such as dialogue machine reading comprehension and emotion classification by simply adding the predicted "unspoken" NM to each utterance or narrative in inputs.

09:00-10:30 (East Foyer / Virtual)

**#104 MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding**
*Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang and Dan Hendrycks*
Reading comprehension of legal text can be a particularly challenging task due to the length and complexity of legal clauses and a shortage of expert-annotated datasets. To address this challenge, we introduce the Merger Agreement Understanding Dataset (MAUD), an expert-annotated reading comprehension dataset based on the American Bar Association's 2021 Public Target Deal Points Study, with over 39,000 examples and over 47,000 total annotations. Our fine-tuned Transformer baselines show promising results, with models performing well above random on most questions. However, on a large subset of questions, there is still room for significant improvement. As the only expert-annotated merger agreement dataset, MAUD is valuable as a benchmark for both the legal profession and the NLP community.

09:00-10:30 (East Foyer / Virtual)
**#105 Do Differences in Values Influence Disagreements in Online Discussions?**
*Michiel van der Meer, Piek Vossen, Catholijn M Jonker and Pradeep Kumar Murukannaiah*
Disagreements are common in online discussions. Disagreement may foster collaboration and improve the quality of a discussion under some conditions. Although there exist methods for recognizing disagreement, a deeper understanding of factors that influence disagreement is lacking in the literature. We investigate a hypothesis that differences in *personal values* are indicative of disagreement in online discussions. We show how state-of-the-art models can be used for estimating values in online discussions and how the estimated values can be aggregated into value profiles. We evaluate the estimated value profiles based on human-annotated agreement labels. We find that the dissimilarity of value profiles correlates with disagreement in specific cases. We also find that including value information in agreement prediction improves performance.

09:00-10:30 (East Foyer / Virtual)
**#106 A Benchmark for Reasoning with Spatial Prepositions**
*Iulia Maria Comsa and Srini Narayanan*
Spatial reasoning is a fundamental building block of human cognition, used in representing, grounding, and reasoning about physical and abstract concepts. We propose a novel benchmark focused on assessing inferential properties of statements with spatial prepositions. The benchmark includes original datasets in English and Romanian and aims to probe the limits of reasoning about spatial relations in large language models. We use prompt engineering to study the performance of two families of large language models, PaLM and GPT-3, on our benchmark. Our results show considerable variability in the performance of smaller and larger models, as well as across prompts and languages. However, none of the models reaches human performance.

09:00-10:30 (East Foyer / Virtual)
**#107 Elevating Code-mixed Text Handling through Auditory Information of Words**
*Mamta Mamta, Zishan Ahmad and Asif Ekbal*
With the growing popularity of code-mixed data, there is an increasing need for better handling of this type of data, which poses a number of challenges, such as dealing with spelling variations, multiple languages, different scripts, and a lack of resources. Current language models face difficulty in effectively handling code-mixed data as they primarily focus on the semantic representation of words and ignore the auditory phonetic features. This leads to difficulties in handling spelling variations in code-mixed text. In this paper, we propose an effective approach for creating language models for handling code-mixed textual data using auditory information of words from SOUNDEX. Our approach includes a pre-training step based on masked-language-modelling, which includes SOUNDEX representations (SAMLM) and a new method of providing input data to the pre-trained model. Through experimentation on various code-mixed datasets (of different languages) for sentiment, offensive and aggression classification tasks, we establish that our novel language modeling approach (SAMLM) results in improved robustness towards adversarial attacks on code-mixed classification tasks. Additionally, our SAMLM based approach also results in better classification results over the popular baselines for code-mixed tasks. We use the explainability technique, SHAP (SHapley Additive exPlanations) to explain how the auditory features incorporated through SAMLM assist the model to handle the code-mixed text effectively and increase robustness against adversarial attacks.

09:00-10:30 (East Foyer / Virtual)
**#108 Supervised Gradual Machine Learning for Aspect-Term Sentiment Analysis**
*Yanyan Wang, Qun Chen, Murtadha Ahmed, Zhaoqiang Chen, Jing Su, Wei Pan and Zhanhuai Li*
Recent work has shown that Aspect-Term Sentiment Analysis (ATSA) can be effectively performed by Gradual Machine Learning (GML). However, the performance of the current unsupervised solution is limited by inaccurate and insufficient knowledge conveyance. In this paper, we propose a supervised GML approach for ATSA, which can effectively exploit labeled training data to improve knowledge conveyance. It leverages binary polarity relations between instances, which can be either similar or opposite, to enable supervised knowledge conveyance. Besides the explicit polarity relations indicated by discourse structures, it also separately supervises a polarity classification DNN and a binary siamese network to extract implicit polarity relations. The proposed approach fulfills knowledge conveyance by modeling detected relations as binary features in a factor graph. Our extensive experiments on real benchmark data show that it achieves the state-of-the-art performance across all the test workloads. Our work demonstrates clearly that, in collaboration with DNN for feature extraction, GML outperforms pure DNN solutions.

09:00-10:30 (East Foyer / Virtual)
**#109 PASTA: A Dataset for Modeling PArticipant STAtes in Narratives**
*Sayontan Ghosh, Mahnaz Koupaee, Isabella Chen, Francis Ferraro, Nathanael Chambers and Niranjan Balasubramanian*
The events in a narrative are understood as a coherent whole via the underlying states of their participants. Often, these participant states are not explicitly mentioned, instead left to be inferred by the reader. A model that understands narratives should likewise infer these implicit states, and even reason about the impact of changes to these states on the narrative. To facilitate this goal, we introduce a new crowdsourced English-language, Participant States dataset, PASTA. This dataset contains inferable participant states; a counterfactual perturbation to each state; and the changes to the story that would be necessary if the counterfactual were true. We introduce three state-based reasoning tasks that test for the ability to infer when a state is entailed by a story, to revise a story conditioned on a counterfactual state, and to explain the most likely state change given a revised story. Experiments show that today's LLMs can reason about states to some degree, but there is large room for improvement, especially in problems requiring access and ability to reason with diverse types of knowledge (e.g. physical, numerical, factual).

09:00-10:30 (East Foyer / Virtual)
**#110 Cross-functional Analysis of Generalisation in Behavioural Learning**
*Pedro Henrique Luz de Araujo and Benjamin Roth*
In behavioural testing, system functionalities underrepresented in the standard evaluation setting (with a held-out test set) are validated through controlled input-output pairs. Optimising performance on the behavioural tests during training (behavioural learning) would improve coverage of phenomena not sufficiently represented in the i.i.d. data and could lead to seemingly more robust models. However, there is the

risk that the model narrowly captures spurious correlations from the behavioural test suite, leading to overestimation and misrepresentation of model performance one of the original pitfalls of traditional evaluation. In this work, we introduce BeLUGA, an analysis method for evaluating behavioural learning considering generalisation across dimensions of different granularity levels. We optimise behaviour-specific loss functions and evaluate models on several partitions of the behavioural test suite controlled to leave out specific phenomena. An aggregate score measures generalisation to unseen functionalities (or overfitting). We use BeLUGA to examine three representative NLP tasks (sentiment analysis, paraphrase identification and reading comprehension) and compare the impact of a diverse set of regularisation and domain generalisation methods on generalisation performance.

09:00-10:30 (East Foyer / Virtual)
**#111 Benchmarking the Generation of Fact Checking Explanations**
*Daniel Russo, Serra Sinem Tekiroglu and Marco Guerini*
Fighting misinformation is a challenging, yet crucial, task. Despite the growing number of experts being involved in manual fact-checking, this activity is time-consuming and cannot keep up with the ever-increasing amount of Fake News produced daily. Hence, automating this process is necessary to help curb misinformation. Thus far, researchers have mainly focused on claim veracity classification. In this paper, instead, we address the generation of justifications (textual explanation of why a claim is classified as either true or false) and benchmark it with novel datasets and advanced baselines. In particular, we focus on summarization approaches over unstructured knowledge (i.e. news articles) and we experiment with several extractive and abstractive strategies. We employed two datasets with different styles and structures, in order to assess the generalizability of our findings. Results show that in justification production summarization benefits from the claim information, and, in particular, that a claim-driven extractive step improves abstractive summarization performances. Finally, we show that although cross-dataset experiments suffer from performance degradation, a unique model trained on a combination of the two datasets is able to retain style information in an efficient manner.

09:00-10:30 (East Foyer / Virtual)
**#112 Benchmarking Large Language Models for News Summarization**
*Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown and Tatsunori Hashimoto*
Large language models (LLMs) have shown promise for automatic summarization but the reasons behind their successes are poorly understood. By conducting a human evaluation on ten LLMs across different pretraining methods, prompts, and model scales, we make two important observations. First, we find instruction tuning, and not model size, is the key to the LLM's zero-shot summarization capability. Second, existing studies have been limited by low-quality references, leading to underestimates of human performance and lower few-shot and finetuning performance. To better evaluate LLMs, we perform human evaluation over high-quality summaries we collect from freelance writers. Despite major stylistic differences such as the amount of paraphrasing, we find that LMM summaries are judged to be on par with human written summaries.

09:00-10:30 (East Foyer / Virtual)
**#113 DMDD: A Large-Scale Dataset for Dataset Mentions Detection**
*Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut and Longin Jan Lateckit*
The recognition of dataset names is a critical task for automatic information extraction in scientific literature, enabling researchers to understand and identify research opportunities. However, existing corpora for dataset mention detection are limited in size and naming diversity. In this paper, we introduce the Dataset Mentions Detection Dataset (DMDD), the largest publicly available corpus for this task. DMDD consists of the DMDD main corpus, comprising 31,219 scientific articles with over 449,000 dataset mentions weakly annotated in the format of in-text spans, and an evaluation set, which comprises of 450 scientific articles manually annotated for evaluation purposes. We use DMDD to establish baseline performance for dataset mention detection and linking. By analyzing the performance of various models on DMDD, we are able to identify open problems in dataset mention detection. We invite the community to use our dataset as a challenge to develop novel dataset mention detection models.

09:00-10:30 (East Foyer / Virtual)
**#114 Learning to Paraphrase Sentences to Different Complexity Levels**
*Alison Hanyi Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee and Jason S. Chang*
While sentence simplification is an active research topic in NLP, its adjacent tasks of sentence complexification and same-level paraphrasing are not. To train models on all three tasks, we present two new unsupervised datasets. We compare these datasets, one labeled by a weak classifier and the other by a rule-based approach, with a single supervised dataset. Using these three datasets for training, we perform extensive experiments on both multitasking and prompting strategies. Compared to other systems trained on unsupervised parallel data, models trained on our weak classifier labeled dataset achieve state-of-the-art performance on the ASSET simplification benchmark.

09:00-10:30 (East Foyer / Virtual)
**#115 AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR**
*Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F.P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori and Clinton Mbataku*
Africa has a very low doctor-to-patient ratio. At very busy clinics, doctors could see 30+ patients per day– a heavy patient burden compared with developed countries– but productivity tools such as clinical automatic speech recognition (ASR) are lacking for these overworked clinicians. However, clinical ASR is mature, even ubiquitous, in developed nations, and clinician-reported performance of commercial clinical ASR systems is generally satisfactory. Furthermore, the recent performance of general domain ASR is approaching human accuracy. However, several gaps exist. Several publications have highlighted racial bias with speech-to-text algorithms and performance on minority accents lags significantly. To our knowledge, there is no publicly available research or benchmark on accented African clinical ASR, and speech data is non-existent for the majority of African accents. We release AfriSpeech, 200hrs of Pan-African English speech, 67,577 clips from 2,463 unique speakers across 120 indigenous accents from 13 countries for clinical and general domain ASR, a benchmark test set, with publicly available pre-trained models with SOTA performance on the AfriSpeech benchmark.

09:00-10:30 (East Foyer / Virtual)
**#116 An Efficient Self-Supervised Cross-View Training For Sentence Embedding**
*Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich and Sarana Nutanong*
Self-supervised sentence representation learning is the task of constructing an embedding space for sentences without relying on human annotation efforts. One straightforward approach is to finetune a pretrained language model (PLM) with a representation learning method such as contrastive learning. While this approach achieves impressive performance on larger PLMs, the performance rapidly degrades as the number of parameters decreases. In this paper, we propose a framework called Self-supervised Cross-View Training (SCT) to narrow the performance gap between large and small PLMs. To evaluate the effectiveness of SCT, we compare it to 5 baseline and state-of-the-art competitors on seven Semantic Textual Similarity (STS) benchmarks using 5 PLMs with the number of parameters ranging from 4M to 340M. The experimental results show that STC outperforms the competitors for PLMs with less than 100M parameters in 18 of 21 cases.

09:00-10:30 (East Foyer / Virtual)
**#117 Discover, Explain, Improve: An Automatic Slice Detection Benchmark for Natural Language Processing**
*Wenyue Hua, Lifeng Jin, Linfeng Song, Haitao Mi, Yongfeng Zhang and Dong Yu*
Pretrained natural language processing(NLP) models have achieved high overall performance, but they still make systematic errors. Instead of manual error analysis, research on slice detection models (SDM), which automatically identify underperforming groups of datapoints, has caught escalated attention in Computer Vision for both understanding model behaviors and providing insights for future model training and designing. However, little research on SDM and quantitative evaluation of their effectiveness have been conducted on NLP tasks. Our paper fills the gap by proposing a benchmark named Discover, Explain, Improve (DEIM) for classification NLP tasks along with a new SDM Edisa. Edisadiscovers coherent and underperforming groups of datapoints; DEIM then unites them under human-understandable concepts and provides comprehensive evaluation tasks and corresponding quantitative metrics. The evaluation in DEIM shows that Edisa can accurately select error-prone datapoints with informative semantic features that summarize error patterns. Detecting difficult datapoints directly boosts model performance without tuning any original model parameters, showing that discovered slices are actionable for users.

09:00-10:30 (East Foyer / Virtual)
**#118 AmbiFC: Fact-Checking Ambiguous Claims with Evidence**
*Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos and Iryna Gurevych*
Automated fact-checking systems verify claims against evidence to predict their veracity. In real-world scenarios, the retrieved evidence may not unambiguously support or refute the claim and yield conflicting but valid interpretations. Existing fact-checking datasets assume that the models developed with them predict a single veracity label for each claim, thus discouraging the handling of such ambiguity. To address this issue we present AmbiFC, a fact-checking dataset with 10k claims derived from real-world information needs. It contains fine-grained evidence annotations of 50k passages from 5k Wikipedia pages. We analyze the disagreements arising from ambiguity when comparing claims against evidence in AmbiFC, observing a strong correlation of annotator disagreement with linguistic phenomena such as underspecification and probabilistic reasoning. We develop models for predicting veracity handling this ambiguity via soft labels and find that a pipeline that learns the label distribution for sentence-level evidence selection and veracity prediction yields the best performance. We compare models trained on different subsets of AmbiFC and show that models trained on the ambiguous instances perform better when faced with the identified linguistic phenomena.

09:00-10:30 (East Foyer / Virtual)
**#119 Multi3WOZ: A Multilingual, Multi-Domain, Multi-Parallel Dataset for Training and Evaluating Culturally Adapted Task-Oriented Dialog Systems**
*Ivan Vulić, Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci and Anna Korhonen*
Creating high-quality annotated data for task-oriented dialog (ToD) is known to be notoriously difficult, and the challenges are amplified when the goal is to create equitable, culturally adapted, and large-scale ToD datasets for multiple languages. Therefore, the current datasets are still very scarce and suffer from limitations such as translation-based non-native dialogs with translation artefacts, small scale, or lack of cultural adaptation, among others. In this work, we first take stock of the current landscape of multilingual ToD datasets, offering a systematic overview of their properties and limitations. Aiming to reduce all the detected limitations, we then introduce Multi3WOZ, a novel multilingual, multi-domain, multi-parallel ToD dataset. It is large-scale and offers culturally adapted dialogs in 4 languages to enable training and evaluation of multilingual and cross-lingual ToD systems. We describe a complex bottom-up data collection process that yielded the final dataset, and offer the first sets of baseline scores across different ToD-related tasks for future reference, also highlighting its challenging nature.

## Findings 4

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**Always the Best Fit: Adaptive Domain Gap Filling from Causal Perspective for Few-Shot Relation Extraction**
*Ge Bai, Chenji Lu, Jiaxiang Geng, Shilong Li, Yidong Shi, Xiyan Liu, Ying Liu, Zhang Zhang and Ruifang Liu*
Cross-domain Relation Extraction aims to transfer knowledge from a source domain to a different target domain to address low-resource challenges. However, the semantic gap caused by data bias between domains is a major challenge, especially in few-shot scenarios. Previous work has mainly focused on transferring knowledge between domains through shared feature representations without analyzing the impact of each factor that may produce data bias based on the characteristics of each domain. This work takes a causal perspective and proposes a new framework CausalGF. By constructing a unified structural causal model, we estimating the causal effects of factors such as syntactic structure, label distribution,and entities on the outcome. CausalGF calculates the causal effects among the factors and adjusts them dynamically based on domain characteristics, enabling adaptive gap filling. Our experiments show that our approach better fills the domain gap, yielding significantly better results on the cross-domain few-shot relation extraction task.

09:00-10:30 (East Foyer / Virtual)
**C2D2 Dataset: A Resource for the Cognitive Distortion Analysis and Its Impact on Mental Health**
*Bichen Wang, Pengfei Deng, Yanyan Zhao and Bing Qin*
Cognitive distortions refer to patterns of irrational thinking that can lead to distorted perceptions of reality and mental health problems in individuals. Despite previous attempts to detect cognitive distortion through language, progress has been slow due to the lack of appropriate data. In this paper, we present the C2D2 dataset, the first expert-supervised **C**hinese **C**ognitive **D**istortion **D**ataset, which contains 7,500 cognitive distortion thoughts in everyday life scenes. Additionally, we examine the presence of cognitive distortions in social media texts shared by individuals diagnosed with mental disorders, providing insights into the association between cognitive distortions and mental health conditions. We propose that incorporating information about users' cognitive distortions can enhance the performance of existing models mental disorder detection. We contribute to a better understanding of how cognitive distortions appear in individuals' language and their impact on mental health.

09:00-10:30 (East Foyer / Virtual)
**DeltaScore: Fine-Grained Story Evaluation with Perturbations**
*Zhuohan Xie, Miao Li, Trevor Cohn and Jey Han Lau*
Numerous evaluation metrics have been developed for natural language generation tasks, but their effectiveness in evaluating stories is limited as they are not specifically tailored to assess intricate aspects of storytelling, such as fluency and interestingness. In this paper, we introduce DeltaScore, a novel methodology that uses perturbation techniques for the evaluation of nuanced story aspects. We posit that the extent to which a story excels in a specific aspect (e.g., fluency) correlates with the magnitude of its susceptibility to particular perturbations (e.g., the introduction of typos). Given this, we measure the quality of an aspect by calculating the likelihood difference between pre- and

post-perturbation states using pre-trained language models. We compare DeltaScore with existing metrics on storytelling datasets from two domains in five fine-grained story aspects: fluency, coherence, relatedness, logicality, and interestingness. DeltaScore demonstrates strong performance, revealing a surprising finding that one specific perturbation proves highly effective in capturing multiple aspects. Source code is available on our GitHub repository.

09:00-10:30 (East Foyer / Virtual)
**Multi-Modal Knowledge Graph Transformer Framework for Multi-Modal Entity Alignment**
*Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang and Jianxin Li*
Multi-Modal Entity Alignment (MMEA) is a critical task that aims to identify equivalent entity pairs across multi-modal knowledge graphs (MMKGs). However, this task faces challenges due to the presence of different types of information, including neighboring entities, multi-modal attributes, and entity types. Directly incorporating the above information (e.g., concatenation or attention) can lead to an unaligned information space. To address these challenges, we propose a novel MMEA transformer, called Meaformer, that hierarchically introduces neighbor features, multi-modal attributes, and entity types to enhance the alignment task. Taking advantage of the transformer's ability to better integrate multiple information, we design a hierarchical modifiable self-attention block in a transformer encoder to preserve the unique semantics of different information. Furthermore, we design two entity-type prefix injection methods to redintegrate entity-type information using type prefixes, which help to restrict the global information of entities not present in the MMKGs.

09:00-10:30 (East Foyer / Virtual)
**Non-Autoregressive Document-Level Machine Translation**
*Guangsheng Bao, Zhiyang Teng, Hao Zhou, Jianhao Yan and Yue Zhang*
Non-autoregressive translation (NAT) models achieve comparable performance and superior speed compared to auto-regressive translation (AT) models in the context of sentence-level machine translation (MT). However, their abilities are unexplored in document-level MT, hindering their usage in real scenarios. In this paper, we conduct a comprehensive examination of typical NAT models in the context of document-level MT and further propose a simple but effective design of sentence alignment between source and target. Experiments show that NAT models achieve high acceleration on documents, and sentence alignment significantly enhances their performance. However, current NAT models still have a significant performance gap compared to their AT counterparts. Further investigation reveals that NAT models suffer more from the multi-modality and misalignment issues in the context of document-level MT, and current NAT models struggle with exploiting document context and handling discourse phenomena. We delve into these challenges and provide our code at https://github.com/baoguangsheng/nat-on-doc.

09:00-10:30 (East Foyer / Virtual)
**Defining a New NLP Playground**
*Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel R. Tetreault, Eduard Hovy and Heng Ji*
The recent explosion of performance of large language models (LLMs) has changed the field of Natural Language Processing (NLP) more abruptly and seismically than any other shift in the field's 80 year history. This has resulted in concerns that the field will become homogenized and resource-intensive. This new status quo has put many academic researchers, especially PhD students, at a disadvantage. This paper aims to define a new NLP playground by proposing 20+ PhD-dissertation-worthy research directions, covering theoretical analysis, new and challenging problems, learning paradigms and interdisciplinary applications.

09:00-10:30 (East Foyer / Virtual)
**Multi-Defendant Legal Judgment Prediction via Hierarchical Reasoning**
*Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang and Zhaochun Ren*
Multiple defendants in a criminal fact description generally exhibit complex interactions, and cannot be well handled by existing Legal Judgment Prediction (LJP) methods which focus on predicting judgment results (e.g., law articles, charges, and terms of penalty) for single-defendant cases. To address this problem, we propose the task of multi-defendant LJP, which aims to automatically predict the judgment results for each defendant of multi-defendant cases. Two challenges arise with the task of multi-defendant LJP: (1) indistinguishable judgment results among various defendants; and (2) the lack of a real-world dataset for training and evaluation. To tackle the first challenge, we formalize the multi-defendant judgment process as hierarchical reasoning chains and introduce a multi-defendant LJP method, named Hierarchical Reasoning Network (HRN), which follows the hierarchical reasoning chains to determine criminal relationships, sentencing circumstances, law articles, charges, and terms of penalty for each defendant. To tackle the second challenge, we collect a real-world multi-defendant LJP dataset, namely MultiLJP, to accelerate the relevant research in the future. Extensive experiments on MultiLJP verify the effectiveness of our proposed HRN.

09:00-10:30 (East Foyer / Virtual)
**On Uncertainty Calibration and Selective Generation in Probabilistic Neural Summarization: A Benchmark Study**
*Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren and Jeremiah Zhe Liu*
Modern deep models for summarization attains impressive benchmark performance, but they are prone to generating miscalibrated predictive uncertainty. This means that they assign high confidence to low-quality predictions, leading to compromised reliability and trustworthiness in real-world applications. Probabilistic deep learning methods are common solutions to the miscalibration problem. However, their relative effectiveness in complex autoregressive summarization tasks are not well-understood. In this work, we thoroughly investigate different state-of-the-art probabilistic methods' effectiveness in improving the uncertainty quality of the neural summarization models, across three large-scale benchmarks with varying difficulty using our newly introduced evaluation protocol. We show that the probabilistic methods consistently improve the model's generation and uncertainty quality, leading to improved selective generation performance (i.e., abstaining from low-quality summaries) in practice. We also reveal notable failure patterns of probabilistic methods widely-adopted in NLP community (e.g., Deep Ensemble and Monte Carlo Dropout), cautioning the importance of choosing appropriate method for the data setting.

09:00-10:30 (East Foyer / Virtual)
**Epsilon Sampling Rocks: Investigating Sampling Strategies for Minimum Bayes Risk Decoding for Machine Translation**
*Markus Freitag, Behrooz Ghorbani and Patrick Fernandes*
Recent advances in machine translation (MT) have shown that Minimum Bayes Risk (MBR) decoding can be a powerful alternative to beam search decoding, especially when combined with neural-based utility functions. However, the performance of MBR decoding depends heavily on how and how many candidates are sampled from the model. In this paper, we explore how different sampling approaches for generating candidate lists for MBR decoding affect performance. We evaluate popular sampling approaches, such as ancestral, nucleus, and top-k sampling. Based on our insights into their limitations, we experiment with the recently proposed epsilon-sampling approach, which prunes away all tokens with a probability smaller than epsilon, ensuring that each token in a sample receives a fair probability mass. Through extensive human evaluations, we demonstrate that MBR decoding based on epsilon-sampling significantly outperforms not only beam search decoding, but also MBR decoding with all other tested sampling methods across four language pairs.

09:00-10:30 (East Foyer / Virtual)

**Simultaneous Machine Translation with Tailored Reference**
*Shoutao Guo, Shaolei Zhang and Yang Feng*
Simultaneous machine translation (SiMT) generates translation while reading the whole source sentence. However, existing SiMT models are typically trained using the same reference disregarding the varying amounts of available source information at different latency. Training the model with ground-truth at low latency may introduce forced anticipations, whereas utilizing reference consistent with the source word order at high latency results in performance degradation. Consequently, it is crucial to train the SiMT model with appropriate reference that avoids forced anticipations during training while maintaining high quality. In this paper, we propose a novel method that provides tailored reference for the SiMT models trained at different latency by rephrasing the ground-truth. Specifically, we introduce the tailor, induced by reinforcement learning, to modify ground-truth to the tailored reference. The SiMT model is trained with the tailored reference and jointly optimized with the tailor to enhance performance. Importantly, our method is applicable to a wide range of current SiMT approaches. Experiments on three translation tasks demonstrate that our method achieves state-of-the-art performance in both fixed and adaptive policies.

09:00-10:30 (East Foyer / Virtual)
**On Event Individuation for Document-Level Information Extraction**
*William Gantt, Reno Kriz, Yunmo Chen, Siddharth Vashishtha and Aaron Steven White*
As information extraction (IE) systems have grown more adept at processing whole documents, the classic task of *template filling* has seen renewed interest as a benchmark for document-level IE. In this position paper, we call into question the suitability of template filling for this purpose. We argue that the task demands definitive answers to thorny questions of *event individuation* — the problem of distinguishing distinct events — about which even human experts disagree. Through an annotation study and error analysis, we show that this raises concerns about the usefulness of template filling metrics, the quality of datasets for the task, and the ability of models to learn it. Finally, we consider possible solutions.

09:00-10:30 (East Foyer / Virtual)
**Evaluating the Knowledge Base Completion Potential of GPT**
*Blerta Veseli, Simon Razniewski, Jan-Christoph Kalo and Gerhard Weikum*
Structured knowledge bases (KBs) are an asset for search engines and other applications but are inevitably incomplete. Language models (LMs) have been proposed for unsupervised knowledge base completion (KBC), yet, their ability to do this at scale and with high accuracy remains an open question. Prior experimental studies mostly fall short because they only evaluate on popular subjects, or sample already existing facts from KBs. In this work, we perform a careful evaluation of GPT's potential to complete the largest public KB: Wikidata. We find that, despite their size and capabilities, models like GPT-3, ChatGPT and GPT-4 do not achieve fully convincing results on this task. Nonetheless, it provides solid improvements over earlier approaches with smaller LMs. In particular, we show that it is feasible to extend Wikidata by 27M facts at 90% precision.

09:00-10:30 (East Foyer / Virtual)
**Guiding LLM to Fool Itself: Automatically Manipulating Machine Reading Comprehension Shortcut Triggers**
*Mosh Levy, Shauli Ravfogel and Yoav Goldberg*
Recent applications of LLMs in Machine Reading Comprehension (MRC) systems have shown impressive results, but the use of shortcuts, mechanisms triggered by features spuriously correlated to the true label, has emerged as a potential threat to their reliability. We analyze the problem from two angles: LLMs as editors, guided to edit text to mislead LLMs; and LLMs as readers, who answer questions based on the edited text. We introduce a framework that guides an editor to add potential shortcuts-triggers to samples. Using GPT4 as the editor, we find it can successfully edit trigger shortcut in samples that fool LLMs. Analysing LLMs as readers, we observe that even capable LLMs can be deceived using shortcut knowledge. Strikingly, we discover that GPT4 can be deceived by its own edits (15% drop in F1). Our findings highlight inherent vulnerabilities of LLMs to shortcut manipulations. We publish ShortcutQA, a curated dataset generated by our framework for future research.

09:00-10:30 (East Foyer / Virtual)
**Unsupervised Lexical Simplification with Context Augmentation**
*Takashi Wada, Timothy Baldwin and Jey Han Lau*
We propose a new unsupervised lexical simplification method that uses only monolingual data and pre-trained language models. Given a target word and its context, our method generates substitutes based on the target context and also additional contexts sampled from monolingual data. We conduct experiments in English, Portuguese, and Spanish on the TSAR-2022 shared task, and show that our model substantially outperforms other unsupervised systems across all languages. We also establish a new state-of-the-art by ensembling our model with GPT-3.5. Lastly, we evaluate our model on the SWORDS lexical substitution data set, achieving a state-of-the-art result.

09:00-10:30 (East Foyer / Virtual)
**EconBERTa: Towards Robust Extraction of Named Entities in Economics**
*Karim Lasri, Pedro Vitor Quinta de Castro, Mona Schirmer, Luis Eduardo San Martin, Linxi Wang, Tomáš Dulka, Haaya Naushan, John Pougué-Biyong, Arianna Legovini and Samuel Fraiberger*
Adapting general-purpose language models has proven to be effective in tackling downstream tasks within specific domains. In this paper, we address the task of extracting entities from the economics literature on impact evaluation. To this end, we release EconBERTa, a large language model pretrained on scientific publications in economics, and ECON-IE, a new expert-annotated dataset of economics abstracts for Named Entity Recognition (NER). We find that EconBERTa reaches state-of-the-art performance on our downstream NER task. Additionally, we extensively analyze the model's generalization capacities, finding that most errors correspond to detecting only a subspan of an entity or failure to extrapolate to longer sequences. This limitation is primarily due to an inability to detect part-of-speech sequences unseen during training, and this effect diminishes when the number of unique instances in the training set increases. Examining the generalization abilities of domain-specific language models paves the way towards improving the robustness of NER models for causal knowledge extraction.

09:00-10:30 (East Foyer / Virtual)
**Cross-lingual Open-Retrieval Question Answering for African Languages**
*Odunayo Ogundepo, Tajuddeen Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Chinenye Emezue, Albert Kahira, Shamsuddeen Hassan Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Ijeoma Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Oluwaseyi Ajayi, Verrah Akinyi Otiende, Andre Niyongabo Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Felix Onwuegbuzia, Falalu Ibrahim Lawan, Ibrahim Said Ahmad, Jesujoba Oluwadara Alabi, Chinedu Emmanuel Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Nasir Iro and Sonia Adhiambo*
African languages have far less in-language content available digitally, making it challenging for question answering systems to satisfy the information needs of users. Cross-lingual open-retrieval question answering (XOR QA) systems – those that retrieve answer content from other languages while serving people in their native language—offer a means of filling this gap. To this end, we create Our Dataset, the

first cross-lingual QA dataset with a focus on African languages. Our Dataset includes 12,000+ XOR QA examples across 10 African languages. While previous datasets have focused primarily on languages where cross-lingual QA augments coverage from the target language, Our Dataset focuses on languages where cross-lingual answer content is the only high-coverage source of answer content. Because of this, we argue that African languages are one of the most important and realistic use cases for XOR QA. Our experiments demonstrate the poor performance of automatic translation and multilingual retrieval methods. Overall, Our Dataset proves challenging for state-of-the-art QA models. We hope that the dataset enables the development of more equitable QA technology.

09:00-10:30 (East Foyer / Virtual)
**Improving Sequential Model Editing with Fact Retrieval**
*Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong, Qinghua Chai and Jeff Z. Pan*
The task of sequential model editing is to fix erroneous knowledge in Pre-trained Language Models (PLMs) efficiently, precisely and continuously. Although existing methods can deal with a small number of modifications, these methods experience a performance decline or require additional annotated data, when the number of edits increases. In this paper, we propose a **R**etrieval **A**ugmented **S**equential Model **E**diting framework (**RASE**) that leverages factual information to enhance editing generalization and to guide the identification of edits by retrieving related facts from the fact-patch memory we constructed. Our main findings are: (i) State-of-the-art models can hardly correct massive mistakes stably and efficiently; (ii) Even if we scale up to thousands of edits, RASE can significantly enhance editing generalization and maintain consistent performance and efficiency; (iii) RASE can edit large-scale PLMs and increase the performance of different editors. Moreover, it can integrate with ChatGPT and further improve performance. Our code and data are available at: https://github.com/sev777/RASE.

09:00-10:30 (East Foyer / Virtual)
**Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection and Preparation**
*A. Seza Doğruöz, Sunayana Sitaram and Zheng Xin Yong*
Multilingualism is widespread around the world and code-switching (CSW) is a common practice among different language pairs/tuples across locations and regions. However, there is still not much progress in building successful CSW systems, despite the recent advances in Massive Multilingual Language Models (MMLMs). We investigate the reasons behind this setback through a critical study about the existing CSW data sets (68) across language pairs in terms of the collection and preparation (e.g. transcription and annotation) stages. This in-depth analysis reveals that **a)** most CSW data involves English ignoring other language pairs/tuples **b)** there are flaws in terms of representativeness in data collection and preparation stages due to ignoring the location based, socio-demographic and register variation in CSW. In addition, lack of clarity on the data selection and filtering stages shadow the representativeness of CSW data sets. We conclude by providing a short check-list to improve the representativeness for forthcoming studies involving CSW data collection and preparation.

09:00-10:30 (East Foyer / Virtual)
**CTQScorer: Combining Multiple Features for In-context Example Selection for Machine Translation**
*Aswanth Kumar M, Ratish Puduppully, Raj Dabre and Anoop Kunchukuttan*
Large language models have demonstrated the capability to perform on machine translation when the input is prompted with a few examples (in-context learning). Translation quality depends on various features of the selected examples, such as their quality and relevance, but previous work has predominantly focused on individual features in isolation. In this paper, we propose a general framework for combining different features influencing example selection. We learn a regression model, CTQ Scorer (Contextual Translation Quality), that selects examples based on multiple features in order to maximize the translation quality. On multiple language pairs and language models, we show that CTQ Scorer helps significantly outperform random selection as well as strong single-factor baselines reported in the literature. We also see an improvement of over 2.5 COMET points on average with respect to a strong BM25 retrieval-based baseline.

09:00-10:30 (East Foyer / Virtual)
**Mind the Gap Between Conversations for Improved Long-Term Dialogue Generation**
*Qiang Zhang, Jason Naradowsky and Yusuke Miyao*
Knowing how to end and resume conversations over time is a natural part of communication, allowing for discussions to span weeks, months, or years. The duration of gaps between conversations dictates which topics are relevant and which questions to ask, and dialogue systems which do not explicitly model time may generate responses that are unnatural. In this work we explore the idea of making dialogue models aware of time, and present GapChat, a multi-session dialogue dataset in which the time between each session varies. While the dataset is constructed in real-time, progress on events in speakers' lives is simulated in order to create realistic dialogues occurring across a long timespan. We expose time information to the model and compare different representations of time and event progress. In human evaluation we show that time-aware models perform better in metrics that judge the relevance of the chosen topics and the information gained from the conversation.

09:00-10:30 (East Foyer / Virtual)
**Retrieval-Augmented Few-shot Text Classification**
*Guoxin Yu, Lemao Liu, Haiyun Jiang, Shuming Shi and Xiang Ao*
Retrieval-augmented methods are successful in the standard scenario where the retrieval space is sufficient; whereas in the few-shot scenario with limited retrieval space, this paper shows it is non-trivial to put them into practice. First, it is impossible to retrieve semantically similar examples by using an off-the-shelf metric and it is crucial to learn a task-specific retrieval metric; Second, our preliminary experiments demonstrate that it is difficult to optimize a plausible metric by minimizing the standard cross-entropy loss. The in-depth analyses quantitatively show minimizing cross-entropy loss suffers from the weak supervision signals and the severe gradient vanishing issue during the optimization. To address these issues, we introduce two novel training objectives, namely EM-L and R-L, which provide more task-specific guidance to the retrieval metric by the EM algorithm and a ranking-based loss, respectively. Extensive experiments on 10 datasets prove the superiority of the proposed retrieval augmented methods on the performance.

09:00-10:30 (East Foyer / Virtual)
**InterroLang: Exploring NLP Models and Datasets through Dialogue-based Explanations**
*Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz and Sebastian Möller*
While recently developed NLP explainability methods let us open the black box in various ways (Madsen et al., 2022), a missing ingredient in this endeavor is an interactive tool offering a conversational interface. Such a dialogue system can help users explore datasets and models with explanations in a contextualized manner, e.g. via clarification or follow-up questions, and through a natural language interface. We adapt the conversational explanation framework TalkToModel (Slack et al., 2022) to the NLP domain, add new NLP-specific operations such as free-text rationalization, and illustrate its generalizability on three NLP tasks (dialogue act classification, question answering, hate speech detection). To recognize user queries for explanations, we evaluate fine-tuned and few-shot prompting models and implement a novel adapter-based approach. We then conduct two user studies on (1) the perceived correctness and helpfulness of the dialogues, and (2) the simulatability, i.e. how objectively helpful dialogical explanations are for humans in figuring out the model's predicted label when it's not shown. We found rationalization and feature attribution were helpful in explaining the model behavior. Moreover, users could more reliably predict the model outcome based on an explanation dialogue rather than one-off explanations.

09:00-10:30 (East Foyer / Virtual)
**Controllable Chest X-Ray Report Generation from Longitudinal Representations**
*Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton and Alison Q O'Neil*
Radiology reports are detailed text descriptions of the content of medical scans. Each report describes the presence/absence and location of relevant clinical findings, commonly including comparison with prior exams of the same patient to describe how they evolved. Radiology reporting is a time-consuming process, and scan results are often subject to delays. One strategy to speed up reporting is to integrate automated reporting systems, however clinical deployment requires high accuracy and interpretability. Previous approaches to automated radiology reporting generally do not provide the prior study as input, precluding comparison which is required for clinical accuracy in some types of scans, and offer only unreliable methods of interpretability. Therefore, leveraging an existing visual input format of anatomical tokens, we introduce two novel aspects: (1) longitudinal representation learning – we input the prior scan as an additional input, proposing a method to align, concatenate and fuse the current and prior visual information into a joint longitudinal representation which can be provided to the multimodal report generation model; (2) sentence-anatomy dropout – a training strategy for controllability in which the report generator model is trained to predict only sentences from the original report which correspond to the subset of anatomical regions given as input. We show through in-depth experiments on the MIMIC-CXR dataset how the proposed approach achieves state-of-the-art results while enabling anatomy-wise controllable report generation.

09:00-10:30 (East Foyer / Virtual)
**Breaking the Language Barrier: Improving Cross-Lingual Reasoning with Structured Self-Attention**
*Negar Foroutan, Mohammadreza Banaei, Karl Aberer and Antoine Bosselut*
In this work, we study whether multilingual language models (MultiLMs) can transfer logical reasoning abilities to other languages when they are fine-tuned for reasoning in a different language. We evaluate the cross-lingual reasoning abilities of MultiLMs in two schemes: (1) where the language of the context and the question remain the same in the new languages that are tested (i.e., the reasoning is still monolingual, but the model must transfer the learned reasoning ability across languages), and (2) where the language of the context and the question is different (which we term code-switched reasoning). On two logical reasoning datasets, RuleTaker and LeapOfThought, we demonstrate that although MultiLMs can transfer reasoning ability across languages in a monolingual setting, they struggle to transfer reasoning abilities in a code-switched setting. Following this observation, we propose a novel attention mechanism that uses a dedicated set of parameters to encourage cross-lingual attention in code-switched sequences, which improves the reasoning performance by up to 14% and 4% on the RuleTaker and LeapOfThought datasets, respectively.

09:00-10:30 (East Foyer / Virtual)
**Tokenization Consistency Matters for Generative Models on Extractive NLP Tasks**
*Kaiser Sun, Peng Qi, Yuhao Zhang, Lan Liu, William Yang Wang and Zhiheng Huang*
Generative models have been widely applied to solve extractive tasks, where parts of the input is extracted to form the desired output, and achieved significant success. For example, in extractive question answering (QA), generative models have constantly yielded state-of-the-art results. In this work, we study the issue of tokenization inconsistency that is commonly neglected in training these models. This issue damages the extractive nature of these tasks after the input and output are tokenized inconsistently by the tokenizer, and thus leads to performance drop as well as hallucination. We propose a simple yet effective fix to this issue and conduct a case study on extractive QA. We show that, with consistent tokenization, the model performs better in both in-domain and out-of-domain datasets, with a notable average of +1.7 F1 gain when a BART model is trained on SQuAD and evaluated on 8 QA datasets. Further, the model converges faster, and becomes less likely to generate out-of-context answers. Our results demonstrate the need for increased scrutiny regarding how tokenization is done in extractive tasks and the benefits of consistent tokenization during training.

09:00-10:30 (East Foyer / Virtual)
**Improving Cross-lingual Transfer through Subtree-aware Word Reordering**
*Ofir Arviv, Dmitry Nikolaev, Taelin Karidi and Omri Abend*
Despite the impressive growth of the abilities of multilingual language models, such as XLM-R and mT5, it has been shown that they still face difficulties when tackling typologically-distant languages, particularly in the low-resource setting. One obstacle for effective cross-lingual transfer is variability in word-order patterns. It can be potentially mitigated via source- or target-side word reordering, and numerous approaches to reordering have been proposed. However, they rely on language-specific rules, work on the level of POS tags, or only target the main clause, leaving subordinate clauses intact. To address these limitations, we present a new powerful reordering method, defined in terms of Universal Dependencies, that is able to learn fine-grained word-order patterns conditioned on the syntactic context from a small amount of annotated data and can be applied at all levels of the syntactic tree. We conduct experiments on a diverse set of tasks and show that our method consistently outperforms strong baselines over different language pairs and model architectures. This performance advantage holds true in both zero-shot and few-shot scenarios.

09:00-10:30 (East Foyer / Virtual)
**QUADRo: Dataset and Models for QUestion-Answer Database Retrieval**
*Stefano Campese, Ivano Lauriola and Alessandro Moschitti*
An effective approach to design automated Question Answering (QA) systems is to efficiently retrieve answers from pre-computed databases containing question/answer pairs. One of the main challenges to this design is the lack of training/testing data. Existing resources are limited in size and topics and either do not consider answers (question-question similarity only) or their quality in the annotation process. To fill this gap, we introduce a novel open-domain annotated resource to train and evaluate models for this task. The resource consists of 15,211 input questions. Each question is paired with 30 similar question/answer pairs, resulting in a total of 443,000 annotated examples. The binary label associated with each pair indicates the relevance with respect to the input question. Furthermore, we report extensive experimentation to test the quality and properties of our resource with respect to various key aspects of QA systems, including answer relevance, training strategies, and models input configuration.

09:00-10:30 (East Foyer / Virtual)
**Toxicity in chatgpt: Analyzing persona-assigned language models**
*Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan and Karthik R Narasimhan*
Large language models (LLMs) have shown incredible capabilities and transcended the natural language processing (NLP) community, with adoption throughout many services like healthcare, therapy, education, and customer service. Since users include people with critical information needs like students or patients engaging with chatbots, the safety of these systems is of prime importance. Legislation has recognized its significance and recently drafted a "Blueprint For An AI Bill Of Rights" which calls for domain experts to identify risks and potential impact of AI systems. To this end, we systematically evaluate toxicity in over half a million generations of ChatGPT, a popular dialogue-based LLM. We find that setting the system parameter of ChatGPT by assigning it a persona, say that of the boxer Muhammad Ali, significantly increases the toxicity of generations. Depending on the persona assigned to ChatGPT, its toxicity can increase up to 6×, with outputs engaging in incorrect stereotypes, harmful dialogue, and hurtful opinions. Furthermore, we find concerning patterns where specific entities (e.g., certain races) are targeted more than others (3× more) irrespective of the assigned persona, reflecting discriminatory biases in the model. Our

findings show that multiple provisions in the legislative blueprint are being violated, and we hope that the broader AI community rethinks the efficacy of current safety guardrails and develops better techniques that lead to robust, safe, and trustworthy AI.

09:00-10:30 (East Foyer / Virtual)
**What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations**
*Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman and Yejin Choi*
Moral or ethical judgments rely heavily on the specific contexts in which they occur. Understanding varying shades of defeasible contextualizations (i.e., additional information that strengthens or attenuates the moral acceptability of an action) is critical to accurately represent the subtlety and intricacy of grounded human moral judgment in real-life scenarios. We introduce defeasible moral reasoning: a task to provide grounded contexts that make an action more or less morally acceptable, along with commonsense rationales that justify the reasoning. To elicit high-quality task data, we take an iterative self-distillation approach that starts from a small amount of unstructured seed knowledge from GPT-3 and then alternates between (1) self-distillation from student models; (2) targeted filtering with a critic model trained by human judgment (to boost validity) and NLI (to boost diversity); (3) self-imitation learning (to amplify the desired data quality). This process yields a student model that produces defeasible contexts with improved validity, diversity, and defeasibility. From this model we distill a high-quality dataset, $\delta$-Rules-of-Thumb, of 1.2M entries of contextualizations and rationales for 115K defeasible moral actions rated highly by human annotators 85.9% to 99.8% of the time. Using $\delta$-RoT we obtain a final student model that wins over all intermediate student models by a notable margin.

09:00-10:30 (East Foyer / Virtual)
**Argument mining as a multi-hop generative machine reading comprehension task**
*Boyang Liu, Viktor Schlegel, Riza Batista-Navarro and Sophia Ananiadou*
Argument mining (AM) is a natural language processing task that aims to generate an argumentative graph given an unstructured argumentative text. An argumentative graph that consists of argumentative components and argumentative relations contains completed information of an argument and exhibits the logic of an argument. As the argument structure of an argumentative text can be regarded as an answer to a "why" question, the whole argument structure is therefore similar to the "chain of thought" concept, i.e., the sequence of ideas that lead to a specific conclusion for a given argument (Wei et al., 2022). For argumentative texts in the same specific genre, the "chain of thought" of such texts is usually similar, i.e., in a student essay, there is usually a major claim supported by several claims, and then a number of premises which are related to the claims are included (Eger et al., 2017). In this paper, we propose a new perspective which transfers the argument mining task into a multi-hop reading comprehension task, allowing the model to learn the argument structure as a "chain of thought". We perform a comprehensive evaluation of our approach on two AM benchmarks and find that we surpass SOTA results. A detailed analysis shows that specifically the "chain of thought" information is helpful for the argument mining task.

09:00-10:30 (East Foyer / Virtual)
**Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model**
*Parishad BehnamGhader, Santiago Miret and Siva Reddy*
Augmenting pretrained language models with retrievers has shown promise in effectively solving common NLP problems, such as language modeling and question answering. In this paper, we evaluate the strengths and weaknesses of popular retriever-augmented language models, namely kNN-LM, REALM, DPR + FiD, Contriever + ATLAS, and Contriever + Flan-T5, in reasoning over retrieved statements across different tasks. Our findings indicate that the simple similarity metric employed by retrievers is insufficient for retrieving all the necessary statements for reasoning. Additionally, the language models do not exhibit strong reasoning even when provided with only the required statements. Furthermore, when combined with imperfect retrievers, the performance of the language models becomes even worse, e.g., Flan-T5's performance drops by 28.6% when retrieving 5 statements using Contriever. While larger language models improve performance, there is still a substantial room for enhancement. Our further analysis indicates that multihop retrieve-and-read is promising for large language models like GPT-3.5, but does not generalize to other language models like Flan-T5-xxl. The code is available at https://github.com/McGill-NLP/retriever-lm-reasoning.

09:00-10:30 (East Foyer / Virtual)
**Demystifying Prompts in Language Models via Perplexity Estimation**
*Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith and Luke Zettlemoyer*
Language models can be prompted to perform a wide variety of tasks with zero- and few-shot in-context learning. However, performance varies significantly with the choice of prompt, and we do not yet understand why this happens. In this paper, we analyze the factors that contribute to this variance and establish a new empirical hypothesis: the performance of a prompt is predicted by the extent to which the model is familiar with the language it contains. Over a wide range of tasks, we show that the lower the perplexity of the prompt, the better it is able to perform the task, when considering reasonable prompts that are related to it. As part of our analysis, we also devise a method to automatically extend a small seed set of manually written prompts by paraphrasing with GPT3 and backtranslation. This larger set allows us to verify that perplexity is a strong predictor of the success of a prompt and we show that the lowest perplexity prompts are consistently effective.

09:00-10:30 (East Foyer / Virtual)
**Towards large language model-based personal agents in the enterprise: Current trends and open problems**
*Vinod Muthusamy, Yara Rizk, Kiran Kate, Praveen Venkateswaran, Vatche Isahagian, Ashu Gulati and Parijat Dube*
There is an emerging trend to use large language models (LLMs) to reason about complex goals and orchestrate a set of pluggable tools or APIs to accomplish a goal. This functionality could, among other use cases, be used to build personal assistants for knowledge workers. While there are impressive demos of LLMs being used as autonomous agents or for tool composition, these solutions are not ready mission-critical enterprise settings. For example, they are brittle to input changes, and can produce inconsistent results for the same inputs. These use cases have many open problems in an exciting area of NLP research, such as trust and explainability, consistency and reproducibility, adherence to guardrails and policies, best practices for composable tool design, and the need for new metrics and benchmarks. This vision paper illustrates some examples of LLM-based autonomous agents that reason and compose tools, highlights cases where they fail, surveys some of the recent efforts in this space, and lays out the research challenges to make these solutions viable for enterprises.

09:00-10:30 (East Foyer / Virtual)
**Machine Reading Comprehension using Case-based Reasoning**
*Dung Ngoc Thai, Dhruv Agarwal, Mudit Chaudhary, Wenlong Zhao, Rajarshi Das, Jay-Yoon Lee, Hannaneh Hajishirzi, Manzil Zaheer and Andrew McCallum*
We present an accurate and interpretable method for answer extraction in machine reading comprehension that is reminiscent of case-based reasoning (CBR) from classical AI. Our method (CBR-MRC) builds upon the hypothesis that contextualized answers to similar questions share semantic similarities with each other. Given a test question, CBR-MRC first retrieves a set of similar cases from a nonparametric memory and then predicts an answer by selecting the span in the test context that is most similar to the contextualized representations of answers in the retrieved cases. The semi-parametric nature of our approach allows it to attribute a prediction to the specific set of evidence cases,

making it a desirable choice for building reliable and debuggable QA systems. We show that CBR-MRC provides high accuracy comparable with large reader models and outperforms baselines by 11.5 and 8.4 EM on NaturalQuestions and NewsQA, respectively. Further, we demonstrate the ability of CBR-MRC in identifying not just the correct answer tokens but also the span with the most relevant supporting evidence. Lastly, we observe that contexts for certain question types show higher lexical diversity than others and find that CBR-MRC is robust to these variations while performance using fully-parametric methods drops.

09:00-10:30 (East Foyer / Virtual)
**A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing**
*William Timkey and Tal Linzen*
Two of the central factors believed to underpin human sentence processing difficulty are expectations and retrieval from working memory. A recent attempt to create a unified cognitive model integrating these two factors have relied on the parallels between the self-attention mechanism of transformer language models and cue-based retrieval theories of working memory in human sentence processing (Ryu and Lewis 2021). While the authors show that attention patterns in specialized attention heads of GPT-2 are consistent with a key prediction of cue-based retrieval models, similarity-based interference effects, their method requires the identification of syntactically specialized attention heads, and makes an cognitively implausible implicit assumption that hundreds of memory retrieval operations take place in parallel. In the present work, we develop a recurrent neural language model with a single self-attention head, which more closely parallels the memory system assumed by cognitive theories. We show that our model's single attention head can capture semantic and syntactic interference effects observed in human experiments.

09:00-10:30 (East Foyer / Virtual)
**Non-Compositionality in Sentiment: New Data and Analyses**
*Verna Dankers and Christopher G. Lucas*
When natural language phrases are combined, their meaning is often more than the sum of their parts. In the context of NLP tasks such as sentiment analysis, where the meaning of a phrase is its sentiment, that still applies. Many NLP studies on sentiment analysis, however, focus on the fact that sentiment computations are largely compositional. We, instead, set out to obtain non-compositionality ratings for phrases with respect to their sentiment. Our contributions are as follows: a) a methodology for obtaining those non-compositionality ratings, b) a resource of ratings for 259 phrases – NonCompSST – along with an analysis of that resource, and c) an evaluation of computational models for sentiment analysis using this new resource.

09:00-10:30 (East Foyer / Virtual)
**Efficient Multilingual Language Model Compression through Vocabulary Trimming**
*Asahi Ushio, Yi Zhou and Jose Camacho-Collados*
Multilingual language models (LMs) have become a powerful tool in NLP, especially for non-English languages. Nevertheless, model parameters of multilingual LMs remain large due to the larger embedding matrix of the vocabulary covering tokens in different languages. Instead, monolingual LMs can be trained in a target language with the language-specific vocabulary only. In this paper, we propose vocabulary-trimming (VT), a method to reduce a multilingual LM vocabulary to a target language by deleting potentially irrelevant tokens from its vocabulary. In theory, VT can compress any existing multilingual LM to any language covered by the original model. In our experiments, we show that VT can retain the original performance of the multilingual LM, while being considerably smaller in size than the original multilingual LM. The evaluation is performed over four NLP tasks (two generative and two classification tasks) among four widely used multilingual LMs in seven languages. The results show that this methodology can keep the best of both monolingual and multilingual worlds by keeping a small size as monolingual models without the need for specifically retraining them, and can even help limit potentially harmful social biases.

09:00-10:30 (East Foyer / Virtual)
**CASSI: Contextual and Semantic Structure-based Interpolation Augmentation for Low-Resource NER**
*Tanmay Surana, Thi-Nga Ho, Kyaw Zin Tun and Eng Siong Chng*
While text augmentation methods have been successful in improving performance in the low-resource setting, they suffer from annotation corruption for a token-level task like NER. Moreover, existing methods cannot reliably add context diversity to the dataset, which has been shown to be crucial for low-resource NER. In this work, we propose Contextual and Semantic Structure-based Interpolation (CASSI), a novel augmentation scheme that generates high-quality contextually diverse augmentations while avoiding annotation corruption by structurally combining a pair of semantically similar sentences to generate a new sentence while maintaining semantic correctness and fluency. To accomplish this, we generate candidate augmentations by performing multiple dependency parsing-based exchanges in a pair of semantically similar sentences that are filtered via scoring with a pretrained Masked Language Model and a metric to promote specificity. Experiments show that CASSI consistently outperforms existing methods at multiple low resource levels, in multiple languages, and for noisy and clean text.

09:00-10:30 (East Foyer / Virtual)
**IRFL: Image Recognition of Figurative Language**
*Ron Yosef, Yonatan Bitton and Dafna Shahaf*
Figures of speech such as metaphors, similes, and idioms are integral parts of human communication. They are ubiquitous in many forms of discourse, allowing people to convey complex, abstract ideas and evoke emotion. As figurative forms are often conveyed through multiple modalities (e.g., both text and images), understanding multimodal figurative language is an important AI challenge, weaving together profound vision, language, commonsense and cultural knowledge. In this work, we develop the Image Recognition of Figurative Language (IRFL) dataset. We leverage human annotation and an automatic pipeline we created to generate a multimodal dataset, and introduce two novel tasks as a benchmark for multimodal figurative language understanding. We experimented with state-of-the-art vision and language models and found that the best (22%) performed substantially worse than humans (97%). We release our dataset, benchmark, and code in hopes of driving the development of models that can better understand figurative language.

09:00-10:30 (East Foyer / Virtual)
**Cross-modality Data Augmentation for End-to-End Sign Language Translation**
*Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu and Hui Xiong*
End-to-end sign language translation (SLT) aims to directly convert sign language videos into spoken language texts without intermediate representations. It has been challenging due to the data scarcity of labeled data and the modality gap between sign videos and texts. To tackle these challenges, we propose a novel Cross-modality Data Augmentation (XmDA) framework to transfer the powerful gloss-to-text translation capabilities to end-to-end sign language translation (i.e., video-to-text). Specifically, XmDA consists of two key components: cross-modality mix-up and cross-modality knowledge distillation. The former one explicitly encourages the alignment between sign video features and gloss embeddings to bridge the modality gap. The latter one utilizes the generation knowledge from gloss-to-text teacher models to guide the spoken language text generation. Experimental results on two widely used SLT datasets, i.e., PHOENIX-2014T and CSL-Daily, demonstrate that the proposed XmDA framework significantly and consistently outperforms the baseline models. Extensive analyses confirm our claim that XmDA enhances end-to-end sign language translation by reducing the representation distance between sign videos and glosses,

as well as improving the translation of low-frequency words and long sentences.

09:00-10:30 (East Foyer / Virtual)
**SmartSpanNER: Making SpanNER Robust in Low Resource Scenarios**
*Min Zhang, Xiaosong Qiao, Yanqing Zhao, Shimin Tao and Hao Yang*
Named Entity Recognition (NER) is one of the most fundamental tasks in natural language processing. Span-level prediction (SpanNER) is more naturally suitable for nested NER than sequence labeling (SeqLab). However, according to our experiments, the SpanNER method is more sensitive to the amount of training data, i.e., the F1 score of SpanNER drops much more than that of SeqLab when the amount of training data drops. In order to improve the robustness of SpanNER in low resource scenarios, we propose a simple and effective method SmartSpanNER, which introduces a Named Entity Head (NEH) prediction task to SpanNER and performs multi-task learning together with the task of span classification. Experimental results demonstrate that the robustness of SpanNER could be greatly improved by SmartSpanNER in low resource scenarios constructed on the CoNLL03, Few-NERD, GENIA and ACE05 standard benchmark datasets.

09:00-10:30 (East Foyer / Virtual)
**Knowledge is a Region in Weight Space for Fine-tuned Language Models**
*Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz and Leshem Choshen*
Research on neural networks has focused on understanding a single model trained on a single dataset. However, relatively little is known about the relationships between different models, particularly those trained or tested on different datasets. We address this by studying how the weight space and the underlying loss landscape of different models are interconnected. Specifically, we demonstrate that finetuned models that were optimized for high performance, reside in well-defined regions in weight space, and vice versa – that any model that resides anywhere in those regions also exhibits high performance. Notably, we show that language models that have been finetuned on the same dataset form a tight cluster in the weight space, while models finetuned on different datasets from the same underlying task form a looser cluster. Moreover, traversing around the region between the models leads to new models that perform comparably or even better than models obtained via finetuning, even on tasks that the original models were not finetuned on. Our findings provide insight into the relationships between models, demonstrating that a model positioned between two similar models can acquire the knowledge of both. We leverage this and design a method for selecting a better model for efficient finetuning. Specifically, we show that starting from the center of the region is as effective, if not more, than using the pretrained model in 11 out of 12 datasets, resulting in an average accuracy improvement of 3.06.

09:00-10:30 (East Foyer / Virtual)
**CAR: Conceptualization-Augmented Reasoner for Zero-Shot Commonsense Question Answering**
*Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song and Antoine Bosselut*
The task of zero-shot commonsense question answering evaluates models on their capacity to reason about general scenarios beyond those presented in specific datasets. Existing approaches for tackling this task leverage external knowledge from CommonSense Knowledge Bases (CSKBs) by pre-training the model on synthetic QA pairs constructed from CSKBs. In these approaches, negative examples (distractors) are formulated by randomly sampling from CSKBs using fairly primitive keyword constraints. However, two bottlenecks limit these approaches: the inherent incompleteness of CSKBs limits the semantic coverage of synthetic QA pairs, and the lack of human annotations makes the sampled negative examples potentially uninformative and contradictory. To tackle these limitations above, we propose Conceptualization-Augmented Reasoner (CAR), a zero-shot commonsense question-answering framework that fully leverages the power of conceptualization. Specifically, CAR abstracts a commonsense knowledge triple to many higher-level instances, which increases the coverage of the CSKB and expands the ground-truth answer space, reducing the likelihood of selecting false negative distractors. Extensive experiments demonstrate that CAR more robustly generalizes to answering questions about zero-shot commonsense scenarios than existing methods, including large language models, such as GPT3.5 and ChatGPT. Our code, data, and model checkpoints are available at https://github.com/HKUST-KnowComp/CAR.

09:00-10:30 (East Foyer / Virtual)
**Automatic Analysis of Substantiation in Scientific Peer Reviews**
*Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis and Chloé Clavel*
With the increasing amount of problematic peer reviews in top AI conferences, the community is urgently in need of automatic quality control measures. In this paper, we restrict our attention to substantiation — one popular quality aspect indicating whether the claims in a review are sufficiently supported by evidence — and provide a solution automatizing this evaluation process. To achieve this goal, we first formulate the problem as claim-evidence pair extraction in scientific peer reviews, and collect SubstanReview, the first annotated dataset for this task. SubstanReview consists of 550 reviews from NLP conferences annotated by domain experts. On the basis of this dataset, we train an argument mining system to automatically analyze the level of substantiation in peer reviews. We also perform data analysis on the SubstanReview dataset to obtain meaningful insights on peer reviewing quality in NLP conferences over recent years. The dataset is available at https://github.com/YanzhuGuo/SubstanReview.

09:00-10:30 (East Foyer / Virtual)
**ECHo: A Visio-Linguistic Dataset for Event Causality Inference via Human-Centric Reasoning**
*Yuxi Xie, Guanzhen Li and Min-Yen Kan*
We introduce ECHo (Event Causality Inference via Human-Centric Reasoning), a diagnostic dataset of event causality inference grounded in visio-linguistic social scenarios. ECHo employs real-world human-centric deductive information building on a television crime drama. ECHo requires the Theory-of-Mind (ToM) ability to understand and reason about social interactions based on multimodal information. Using ECHo, we propose a unified Chain-of-Thought (CoT) framework to assess the reasoning capability of current AI systems. Our ToM-enhanced CoT pipeline accommodates various large foundation models in both zero-shot and few-shot visio-linguistic reasoning. We use this framework to scrutinize recent large foundation models such as InstructGPT and MiniGPT-4 on three diagnostic human-centric tasks. Further analysis demonstrates ECHo as a challenging dataset to expose imperfections and inconsistencies in reasoning. Our data and code are publicly available at https://github.com/YuxiXie/ECHo.

09:00-10:30 (East Foyer / Virtual)
**Adversarial Text Generation by Search and Learning**
*Guoyi Li, Bingkang Shi, Zongzhen Liu, Dehan Kong, Yulei Wu, Xiaodan Zhang, Longtao Huang and Honglei Lyu*
Recent research has shown that evaluating the robustness of natural language processing models using textual attack methods is significant. However, most existing text attack methods only use heuristic replacement strategies or language models to generate replacement words at the word level. The blind pursuit of high attack success rates makes it difficult to ensure the quality of the generated adversarial text. As a result, adversarial text is often difficult for humans to understand. In fact, many methods that perform well in terms of text attacks often generate adversarial text with poor quality. To address this important gap, our work treats black-box text attack as an unsupervised text generation problem and proposes a search and learning framework for Adversarial Text Generation by Search and Learning (ATGSL) and develops three adversarial attack methods (ATGSL-SA, ATGSL-BM, ATGSL-FUSION) for black box text attacks. We first apply a heuristic search attack algorithm (ATGSL-SA) and a linguistic thesaurus to generate adversarial samples with high semantic similarity. After this process,

we train a conditional generative model to learn from the search results while smoothing out search noise. Moreover, we design an efficient ATGSL-BM attack algorithm based on the text generator. Furthermore, we propose a hybrid attack method (ATGSL-FUSION) that integrates the advantages of ATGSL-SA and ATGSL-BM to enhance attack effectiveness. Our proposed attack algorithms are significantly superior to the most advanced methods in terms of attack efficiency and adversarial text quality.

09:00-10:30 (East Foyer / Virtual)
**Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!**
*Yubo Ma, Yixin Cao, Yong Ching Hong and Aixin Sun*
Large Language Models (LLMs) have made remarkable strides in various tasks. Whether LLMs are competitive few-shot solvers for information extraction (IE) tasks, however, remains an open problem. In this work, we aim to provide a thorough answer to this question. Through extensive experiments on nine datasets across four IE tasks, we demonstrate that current advanced LLMs consistently exhibit inferior performance, higher latency, and increased budget requirements compared to fine-tuned SLMs under most settings. Therefore, we conclude that LLMs are not effective few-shot information extractors in general. Nonetheless, we illustrate that with appropriate prompting strategies, LLMs can effectively complement SLMs and tackle challenging samples that SLMs struggle with. And moreover, we propose an adaptive filter-then-rerank paradigm to combine the strengths of LLMs and SLMs. In this paradigm, SLMs serve as filters and LLMs serve as rerankers. By prompting LLMs to rerank a small portion of difficult samples identified by SLMs, our preliminary system consistently achieves promising improvements ($2.4\%$ F1-gain on average) on various IE tasks, with an acceptable time and cost investment.

09:00-10:30 (East Foyer / Virtual)
**Self-supervised Post-processing Method to Enrich Pretrained Word Vectors**
*Hwiyeol Jo*
Retrofitting techniques, which inject external resources into word representations, have compensated for the weakness of distributed representations in semantic and relational knowledge between words. However, the previous methods require additional external resources and strongly depend on the lexicon. To address the issues, we propose a simple extension of extrofitting, self-supervised extrofitting: extrofitting by its own word vector distribution. Our methods improve the vanilla embeddings on all of word similarity tasks without any external resources. Moreover, the method is also effective in various languages, which implies that our method will be useful in lexicon-scarce languages. As downstream tasks, we show its benefits in dialogue state tracking and text classification tasks, reporting better and generalized results compared to other word vector specialization methods.

09:00-10:30 (East Foyer / Virtual)
**Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs**
*Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu and Kam-Fai Wong*
Large Language Models (LLMs), such as ChatGPT, greatly empower dialogue systems with strong language understanding and generation capabilities. However, most of the previous works prompt the LLMs to directly generate a response based on the dialogue context, overlooking the underlying linguistic cues about the user status exhibited in the context. Such in-depth dialogue scenarios are challenging for existing LLMs to figure out the user's hidden needs and respond satisfactorily through a single-step inference. To this end, we propose a novel linguistic cue-based chain-of-thoughts (Cue-CoT), which enhances the LLMs inference with an intermediate reasoning step to find cues exhibited in the dialogue, aiming to provide a more personalized and engaging response. To evaluate the approach, we build a benchmark with in-depth dialogue questions, consisting of 6 datasets in both Chinese and English, targeting 3 major linguistic cues during the conversation: personality, emotion, and psychology. We conducted experiments on the proposed benchmark with 5 LLMs under both zero-shot and one-shot settings. Empirical results demonstrate our proposed Cue-CoT method outperforms standard prompting methods in terms of both helpfulness and acceptability on all datasets.

09:00-10:30 (East Foyer / Virtual)
**SIR-ABSC: Incorporating Syntax into RoBERTa-based Sentiment Analysis Models with a Special Aggregator Token**
*Ikhyun Cho, Yoonhwa Jung and Julia Hockenmaier*
We present a simple, but effective method to incorporate syntactic dependency information directly into transformer-based language models (e.g. RoBERTa) for tasks such as Aspect-Based Sentiment Classification (ABSC), where the desired output depends on specific input tokens. In contrast to prior approaches to ABSC that capture syntax by combining language models with graph neural networks over dependency trees, our model, Syntax-Integrated RoBERTa for ABSC (SIR-ABSC) incorporates syntax directly into the language model by using a novel aggregator token. Yet, SIR-ABSC outperforms these more complex models, yielding new state-of-the-art results on ABSC.

09:00-10:30 (East Foyer / Virtual)
**Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation**
*Olli Kuparinen, Aleksandra Miletić and Yves Scherrer*
Text normalization methods have been commonly applied to historical language or user-generated content, but less often to dialectal transcriptions. In this paper, we introduce dialect-to-standard normalization – i.e., mapping phonetic transcriptions from different dialects to the orthographic norm of the standard variety – as a distinct sentence-level character transduction task and provide a large-scale analysis of dialect-to-standard normalization methods. To this end, we compile a multilingual dataset covering four languages: Finnish, Norwegian, Swiss German and Slovene. For the two biggest corpora, we provide three different data splits corresponding to different use cases for automatic normalization. We evaluate the most successful sequence-to-sequence model architectures proposed for text normalization tasks using different tokenization approaches and context sizes. We find that a character-level Transformer trained on sliding windows of three words works best for Finnish, Swiss German and Slovene, whereas the pre-trained byT5 model using full sentences obtains the best results for Norwegian. Finally, we perform an error analysis to evaluate the effect of different data splits on model performance.

09:00-10:30 (East Foyer / Virtual)
**VISTA: Visual-Textual Knowledge Graph Representation Learning**
*Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo and Joyce Jiyoung Whang*
Knowledge graphs represent human knowledge using triplets composed of entities and relations. While most existing knowledge graph embedding methods only consider the structure of a knowledge graph, a few recently proposed multimodal methods utilize images or text descriptions of entities in a knowledge graph. In this paper, we propose visual-textual knowledge graphs (VTKGs), where not only entities but also triplets can be explained using images, and both entities and relations can accompany text descriptions. By compiling visually expressible commonsense knowledge, we construct new benchmark datasets where triplets themselves are explained by images, and the meanings of entities and relations are described using text. We propose VISTA, a knowledge graph representation learning method for VTKGs, which incorporates the visual and textual representations of entities and relations using entity encoding, relation encoding, and triplet decoding transformers. Experiments show that VISTA outperforms state-of-the-art knowledge graph completion methods in real-world VTKGs.

09:00-10:30 (East Foyer / Virtual)
**impact of sample selection on in-context learning for entity extraction from scientific writing**

*Necva Bölücü, Maciej Rybinski and Stephen Wan*
Prompt-based usage of Large Language Models (LLMs) is an increasingly popular way to tackle many well-known natural language problems. This trend is due, in part, to the appeal of the In-Context Learning (ICL) prompt set-up, in which a few selected training examples are provided along with the inference request. ICL, a type of few-shot learning, is especially attractive for natural language processing (NLP) tasks defined for specialised domains, such as entity extraction from scientific documents, where the annotation is very costly due to expertise requirements for the annotators. In this paper, we present a comprehensive analysis of in-context sample selection methods for entity extraction from scientific documents using GPT-3.5 and compare these results against a fully supervised transformer-based baseline. Our results indicate that the effectiveness of the in-context sample selection methods is heavily domain-dependent, but the improvements are more notable for problems with a larger number of entity types. More in-depth analysis shows that ICL is more effective for low-resource set-ups of scientific information extraction

09:00-10:30 (East Foyer / Virtual)
**Self-Supervised Rule Learning to Link Text Segments to Relational Elements of Structured Knowledge**
*Shajith Ikbal, Udit Sharma, Hima Karanam, Sumit Neelam, Ronny Luss, Dheeraj Sreedhar, Pavan Kapanipathi, Naweed Khan, Kyle Erwin, Ndivhuwo Makondo, Ibrahim Abdelaziz, Achille Fokoue, Alexander G. Gray, Maxwell Crouse, Subhajit Chaudhury and Chitra K Subramanian*
We present a neuro-symbolic approach to self-learn rules that serve as interpretable knowledge to perform relation linking in knowledge base question answering systems. These rules define natural language text predicates as a weighted mixture of knowledge base paths. The weights learned during training effectively serve the mapping needed to perform relation linking. We use popular masked training strategy to self-learn the rules. A key distinguishing aspect of our work is that the masked training operate over logical forms of the sentence instead of their natural language text form. This offers opportunity to extract extended context information from the structured knowledge source and use that to build robust and human readable rules. We evaluate accuracy and usefulness of such learned rules by utilizing them for prediction of missing kinship relation in CLUTRR dataset and relation linking in a KBQA system using SWQ-WD dataset. Results demonstrate the effectiveness of our approach - its generalizability, interpretability and ability to achieve an average performance gain of 17% on CLUTRR dataset.

09:00-10:30 (East Foyer / Virtual)
**ACT-SQL: In-Context Learning for Text-to-SQL with Automatically-Generated Chain-of-Thought**
*Hanchong Zhang, Ruisheng Cao, Lu Chen, Hongshen Xu and Kai Yu*
Recently Large Language Models (LLMs) have been proven to have strong abilities in various domains and tasks. We study the problem of prompt designing in the text-to-SQL task and attempt to improve the LLMs' reasoning ability when generating SQL queries. Besides the trivial few-shot in-context learning setting, we design our chain-of-thought (CoT) prompt with a similar method to schema linking. We provide a method named ACT-SQL to automatically generate auto-CoT exemplars and thus the whole process doesn't need manual labeling. Our approach is cost-saving since we only use the LLMs' API call once when generating one SQL query. Furthermore, we extend our in-context learning method to the multi-turn text-to-SQL task. The experiment results show that the LLMs' performance can benefit from our ACT-SQL approach. Our approach achieves SOTA performance on the Spider dev set among existing in-context learning approaches.

09:00-10:30 (East Foyer / Virtual)
**Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues**
*Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King and Kam-Fai Wong*
Open-domain dialogue system usually requires different sources of knowledge to generate more informative and evidential responses. However, existing knowledge-grounded dialogue systems either focus on a single knowledge source or overlook the dependency between multiple sources of knowledge, which may result in generating inconsistent or even paradoxical responses. To incorporate multiple knowledge sources and dependencies between them, we propose SAFARI, a novel framework that leverages the exceptional capabilities of large language models (LLMs) in planning, understanding, and incorporating under both supervised and unsupervised settings. Specifically, SAFARI decouples the knowledge grounding into multiple sources and response generation, which allows easy extension to various knowledge sources including the possibility of not using any sources. To study the problem, we construct a personalized knowledge-grounded dialogue dataset Knowledge Behind Persona (KBP), which is the first to consider the dependency between persona and implicit knowledge. Experimental results on the KBP dataset demonstrate that the SAFARI framework can effectively produce persona-consistent and knowledge-enhanced responses.

09:00-10:30 (East Foyer / Virtual)
**Boosting Inference Efficiency: Unleashing the Power of Parameter-Shared Pre-trained Language Models**
*Weize Chen, Xiaoyue Xu, Xu Han, Yankai Lin, Ruobing Xie, Zhiyuan Liu, Maosong Sun and Jie Zhou*
Parameter-shared pre-trained language models (PLMs) have emerged as a successful approach in resource-constrained environments, enabling substantial reductions in model storage and memory costs without significant performance compromise. However, it is important to note that parameter sharing does not alleviate computational burdens associated with inference, thus impeding its practicality in situations characterized by limited stringent latency requirements or computational resources. Building upon neural ordinary differential equations (ODEs), we introduce a straightforward technique to enhance the inference efficiency of parameter-shared PLMs. Additionally, we propose a simple pre-training technique that leads to fully or partially shared models capable of achieving even greater inference acceleration. The experimental results demonstrate the effectiveness of our methods on both autoregressive and autoencoding PLMs, providing novel insights into more efficient utilization of parameter-shared models in resource-constrained settings.

09:00-10:30 (East Foyer / Virtual)
**Statistically Profiling Biases in Natural Language Reasoning Datasets and Models**
*Shanshan Huang and Kenny Q. Zhu*
Recent studies have shown that many natural language understanding and reasoning datasets contain statistical cues that can be exploited by NLP models, resulting in an overestimation of their capabilities. Existing methods, such as "hypothesis-only" tests and CheckList, are limited in identifying these cues and evaluating model weaknesses. We introduce ICQ (I-See-Cue), a lightweight, general statistical profiling framework that automatically identifies potential biases in multiple-choice NLU datasets without requiring additional test cases. ICQ assesses the extent to which models exploit these biases through black-box testing, addressing the limitations of current methods. In this work, we conduct a comprehensive evaluation of statistical biases in 10 popular NLU datasets and 4 models, confirming prior findings, revealing new insights, and offering an online demonstration system to encourage users to assess their own datasets and models. Furthermore, we present a case study on investigating ChatGPT's bias, providing valuable recommendations for practical applications.

09:00-10:30 (East Foyer / Virtual)
**LogiCoT: Logical Chain-of-Thought Instruction Tuning**
*Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou and Yue Zhang*
Generative Pre-trained Transformer 4 (GPT-4) demonstrates impressive chain-of-thought reasoning ability. Recent work on self-instruction tuning, such as Alpaca, has focused on enhancing the general proficiency of models. These instructions enable the model to achieve performance comparable to GPT-3.5 on general tasks like open-domain text generation and paraphrasing. However, they fall short of helping the

model handle complex reasoning tasks. To bridge the gap, this paper presents LogiCoT, a new instruction-tuning dataset for Logical Chain-of-Thought reasoning with GPT-4. We elaborate on the process of harvesting instructions for prompting GPT-4 to generate chain-of-thought rationales. LogiCoT serves as an instruction set for teaching models of logical reasoning and elicits general reasoning skills.

09:00-10:30 (East Foyer / Virtual)
**Multi-Task Learning of Query Generation and Classification for Generative Conversational Question Rewriting**
*Sarawoot Kongyoung, Craig MacDonald and Iadh Ounis*
In conversational search settings, users ask questions and receive answers as part of a conversation. The ambiguity in the questions is a common challenge, which can be effectively addressed by leveraging contextual information from the conversation history. In this context, determining topic continuity and reformulating questions into well-defined queries are crucial tasks. Previous approaches have typically addressed these tasks either as a classification task in the case of topic continuity or as a text generation task for question reformulation. However, no prior work has combined both tasks to effectively identify ambiguous questions as part of a conversation. In this paper, we propose a Multi-Task Learning (MTL) approach that uses a text generation model for both question rewriting and classification. Our models, based on BART and T5, are trained to rewrite conversational questions and identify follow-up questions simultaneously. We evaluate our approach on multiple test sets and demonstrate that it outperforms single-task learning baselines on the three LIF test sets, with statistically significant improvements ranging from +3.5% to +10.5% in terms of F1 and Micro-F1 scores. We also show that our approach outperforms single-task question rewriting models in passage retrieval on a large OR-QuAC test set.

09:00-10:30 (East Foyer / Virtual)
**Improving Factual Consistency for Knowledge-Grounded Dialogue Systems via Knowledge Enhancement and Alignment**
*Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu and Kam-Fai Wong*
Pretrained language models (PLMs) based knowledge-grounded dialogue systems are prone to generate responses that are factually inconsistent with the provided knowledge source. In such inconsistent responses, the dialogue models fail to accurately express the external factual knowledge they rely upon. Inspired by previous work which identified that feedforward networks (FFNs) within Transformers are responsible for factual knowledge expressions, we investigate two methods to efficiently improve the factual expression capability of FFNs by knowledge enhancement and alignment respectively. We first propose K-Dial, which explicitly introduces extended FFNs in Transformers to enhance factual knowledge expressions given the specific patterns of knowledge-grounded dialogue inputs. Additionally, we apply the reinforcement learning for factual consistency (RLFC) method to implicitly adjust FFNs' expressions in responses by aligning with gold knowledge for the factual consistency preference. To comprehensively assess the factual consistency and dialogue quality of responses, we employ extensive automatic measures and human evaluations including sophisticated fine-grained NLI-based metrics. Experimental results on WoW and CMU_DoG datasets demonstrate that our methods efficiently enhance the ability of the FFN module to convey factual knowledge, validating the efficacy of improving factual consistency for knowledge-grounded dialogue systems.

09:00-10:30 (East Foyer / Virtual)
**GDA: Grammar-based Data Augmentation for Text Classification using Slot Information**
*Joonghyuk Hahn, Hyunjoon Cheon, Elizabeth Grace Orwig, Su-Hyeon Kim, Sang-Ki Ko and Yo-Sub Han*
Recent studies propose various data augmentation approaches to resolve the low-resource problem in natural language processing tasks. Data augmentation is a successful solution to this problem and recent strategies give variation on sentence structures to boost performance. However, these approaches can potentially lead to semantic errors and produce semantically noisy data due to the unregulated variation of sentence structures. In an effort to combat these semantic errors, we leverage slot information, the representation of the context of keywords from a sentence, and form a data augmentation strategy which we propose, called GDA. Our strategy employs algorithms that construct and manipulate rules of context-aware grammar, utilizing this slot information. The algorithms extract recurrent patterns by distinguishing words with slots and form the "rules of grammar"—a set of injective relations between a sentence's semantics and its syntactical structure—to augment the dataset. The augmentation is done in an automated manner with the constructed rules and thus, GDA is explainable and reliable without any human intervention. We evaluate GDA with state-of-the-art data augmentation techniques, including those using pre-trained language models, and the result illustrates that GDA outperforms all other data augmentation methods by 19.38%. Extensive experiments show that GDA is an effective data augmentation strategy that incorporates word semantics for more accurate and diverse data.

09:00-10:30 (East Foyer / Virtual)
**Multi-label and Multi-target Sampling of Machine Annotation for Computational Stance Detection**
*Zhengyuan Liu, Hai Leong Chieu and Nancy F. Chen*
Data collection from manual labeling provides domain-specific and task-aligned supervision for data-driven approaches, and a critical mass of well-annotated resources is required to achieve reasonable performance in natural language processing tasks. However, manual annotations are often challenging to scale up in terms of time and budget, especially when domain knowledge, capturing subtle semantic features, and reasoning steps are needed. In this paper, we investigate the efficacy of leveraging large language models on automated labeling for computational stance detection. We empirically observe that while large language models show strong potential as an alternative to human annotators, their sensitivity to task-specific instructions and their intrinsic biases pose intriguing yet unique challenges in machine annotation. We introduce a multi-label and multi-target sampling strategy to optimize the annotation quality. Experimental results on the benchmark stance detection corpora show that our method can significantly improve performance and learning efficacy.

09:00-10:30 (East Foyer / Virtual)
**Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting**
*Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia and Furu Wei*
Large language models (LLMs) demonstrate impressive multilingual capability, but their performance varies substantially across different languages. In this work, we introduce a simple yet effective method, called cross-lingual-thought prompting (XLT), to systematically improve the multilingual capability of LLMs. Specifically, XLT is a generic template prompt that stimulates cross-lingual and logical reasoning skills to enhance task performance across languages. We conduct comprehensive evaluations on 7 typical benchmarks related to reasoning, understanding, and generation tasks, covering both high-resource and low-resource languages. Experimental results show that XLT not only remarkably enhances the performance of various multilingual tasks but also significantly reduces the gap between the average performance and the best performance of each task in different languages. Notably, XLT brings over 10 points of average improvement in arithmetic reasoning and open-domain question-answering tasks.

09:00-10:30 (East Foyer / Virtual)
**NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark**
*Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle and Eneko Agirre*
In this position paper we argue that the classical evaluation on Natural Language Processing (NLP) tasks using annotated benchmarks is in trouble. The worst kind of data contamination happens when a Large Language Model (LLM) is trained on the test split of a benchmark, and then evaluated in the same benchmark. The extent of the problem is unknown, as it is not straightforward to measure. Contamination causes an overestimation of the performance of a contaminated model in a target benchmark and associated task with respect to their non-contaminated

counterparts. The consequences can be very harmful, with wrong scientific conclusions being published while other correct ones are discarded. This position paper defines different levels of data contamination and argues for a community effort, including the development of automatic and semi-automatic measures to detect when data from a benchmark was exposed to a model, and suggestions for flagging papers with conclusions that are compromised by data contamination.

09:00-10:30 (East Foyer / Virtual)
**Visual Storytelling with Question-Answer Plans**
*Danyang Liu, Mirella Lapata and Frank Keller*
Visual storytelling aims to generate compelling narratives from image sequences. Existing models often focus on enhancing the representation of the image sequence, e.g., with external knowledge sources or advanced graph structures. Despite recent progress, the stories are often repetitive, illogical, and lacking in detail. To mitigate these issues, we present a novel framework which integrates visual representations with pretrained language models and planning. Our model translates the image sequence into a visual prefix, a sequence of continuous embeddings which language models can interpret. It also leverages a sequence of question-answer pairs as a blueprint plan for selecting salient visual concepts and determining how they should be assembled into a narrative. Automatic and human evaluation on the VIST benchmark demonstrates that blueprint-based models generate stories that are more coherent, interesting, and natural compared to competitive baselines and state-of-the-art systems.

09:00-10:30 (East Foyer / Virtual)
**tagE: Enabling an Embodied Agent to Understand Human Instructions**
*Chayan Sarkar, Avik Mitra, Pradip Pramanick and Tapas Nayak*
Natural language serves as the primary mode of communication when an intelligent agent with a physical presence engages with human beings. While a plethora of research focuses on natural language understanding (NLU), encompassing endeavors such as sentiment analysis, intent prediction, question answering, and summarization, the scope of NLU directed at situations necessitating tangible actions by an embodied agent remains limited. The inherent ambiguity and incompleteness inherent in natural language present challenges for intelligent agents striving to decipher human intention. To tackle this predicament head-on, we introduce a novel system known as task and argument grounding for Embodied agents (tagE). At its core, our system employs an inventive neural network model designed to extract a series of tasks from complex task instructions expressed in natural language. Our proposed model adopts an encoder-decoder framework enriched with nested decoding to effectively extract tasks and their corresponding arguments from these intricate instructions. These extracted tasks are then mapped (or grounded) to the robot's established collection of skills, while the arguments find grounding in objects present within the environment. To facilitate the training and evaluation of our system, we have curated a dataset featuring complex instructions. The results of our experiments underscore the prowess of our approach, as it outperforms robust baseline models.

09:00-10:30 (East Foyer / Virtual)
**DiffusionRet: Diffusion-Enhanced Generative Retriever using Constrained Decoding**
*Shanbao Qiao, Xuebing Liu and Seung-Hoon Na*
Generative retrieval, which maps from a query to its relevant document identifiers (docids), has recently emerged as a new information retrieval (IR) paradigm, however, having suffered from 1) the *lack of the intermediate reasoning step*, caused by the manner of merely using a query to perform the hierarchical classification, and 2) the *pretrain-finetune discrepancy*, which comes from the use of the artificial symbols of docids. To address these limitations, we propose the novel approach of using the document generation from a query as an intermediate step before the retrieval, thus presenting $\underline{diffusion}$-enhanced generative $\underline{ret}$rieval (**DiffusionRet**), which consists of two processing steps: 1) the *diffusion-based document generation*, which employs the sequence-to-sequence diffusion model to produce a pseudo document sample from a query, being expected to semantically close to a relevant document; 2) *N-gram-based generative retrieval*, which use another sequence-to-sequence model to generate n-grams that appear in the collection index for linking a generated sample to an original document. Experiment results on MS MARCO and Natural Questions dataset show that the proposed DiffusionRet significantly outperforms all the existing generative retrieval methods and leads to the state-of-the-art performances, even with much smaller number of parameters.

09:00-10:30 (East Foyer / Virtual)
**DocAsRef: An Empirical Study on Repurposing Reference-based Summary Quality Metrics as Reference-free Metrics**
*Forrest Sheng Bao, Ruixuan Tu, Ge Luo, Yinfei Yang, Hebi Li, Minghui Qiu, Youbiao He and Cen Chen*
Automated summary quality assessment falls into two categories: reference-based and reference-free. Reference-based metrics, historically deemed more accurate due to the additional information provided by human-written references, are limited by their reliance on human input. In this paper, we hypothesize that the comparison methodologies used by some reference-based metrics to evaluate a system summary against its corresponding reference can be effectively adapted to assess it against its source document, thereby transforming these metrics into reference-free ones. Experimental results support this hypothesis. After being repurposed reference-freely, the zero-shot BERTScore using the pretrained DeBERTa-large-MNLI model of $<$0.5B parameters consistently outperforms its original reference-based version across various aspects on the SummEval and Newsroom datasets. It also excels in comparison to most existing reference-free metrics and closely competes with zero-shot summary evaluators based on GPT-3.5.

09:00-10:30 (East Foyer / Virtual)
**Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?**
*Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton and Genevieve Grant*
Large Language Models (LLMs), such as ChatGPT, have drawn a lot of attentions recently in the legal domain due to its emergent ability to tackle a variety of legal tasks. However, it is still unknown if LLMs are able to analyze a legal case and perform reasoning in the same manner as lawyers. Therefore, we constructed a novel corpus consisting of scenarios pertain to Contract Acts Malaysia and Australian Social Act for Dependent Child. ChatGPT is applied to perform analysis on the corpus using the IRAC method, which is a framework widely used by legal professionals for organizing legal analysis. Each scenario in the corpus is annotated with a complete IRAC analysis in a semi-structured format so that both machines and legal professionals are able to interpret and understand the annotations. In addition, we conducted the first empirical assessment of ChatGPT for IRAC analysis in order to understand how well it aligns with the analysis of legal professionals. Our experimental results shed lights on possible future research directions to improve alignments between LLMs and legal experts in terms of legal reasoning.

09:00-10:30 (East Foyer / Virtual)
**IndiSocialFT: Multilingual Word Representation for Indian languages in code-mixed environment**
*Saurabh Kumar, Ranbir Singh Sanasam and Sukumar Nandi*
The increasing number of Indian language users on the internet necessitates the development of Indian language technologies. In response to this demand, our paper presents a generalized representation vector for diverse text characteristics, including native scripts, transliterated text, multilingual, code-mixed, and social media-related attributes. We gather text from both social media and well-formed sources and utilize the FastText model to create the "IndiSocialFT" embedding. Through intrinsic and extrinsic evaluation methods, we compare IndiSocialFT with three popular pretrained embeddings trained over Indian languages. Our findings show that the proposed embedding surpasses the baselines

in most cases and languages, demonstrating its suitability for various NLP applications.

09:00-10:30 (East Foyer / Virtual)
**Learning Dynamic Representations for Discourse Dependency Parsing**
*Tianyi Liu, Yansong Feng and Dongyan Zhao*
Transition systems have been widely used for the discourse dependency parsing task. Existing works often characterize transition states by examining a certain number of elementary discourse units (EDUs), while neglecting the arcs obtained from the transition history. In this paper, we propose to employ GAT-based encoder to learn dynamic representations for sub-trees constructed in previous transition steps. By incorporating these representations, our model is able to retain accessibility to all parsed EDUs through the obtained arcs, thus better utilizing the structural information of the document, particularly when handling lengthy text spans with complex structures. For the discourse relation recognition task, we employ edge-featured GATs to derive better representations for EDU pairs. Experimental results show that our model can achieve state-of-the-art performance on widely adopted datasets including RST-DT, SciDTB and CDTB. Our code is available at $https://github.com/lty-lty/Discourse-Dependency-Parsing$.

09:00-10:30 (East Foyer / Virtual)
**LDM$^2$: A Large Decision Model Imitating Human Cognition with Dynamic Memory Enhancement**
*Xingjin Wang, Linjing Li and Daniel Dajun Zeng*
With the rapid development of large language models (LLMs), it is highly demanded that LLMs can be adopted to make decisions to enable the artificial general intelligence. Most approaches leverage manually crafted examples to prompt the LLMs to imitate the decision process of human. However, designing optimal prompts is difficult and the patterned prompts can hardly be generalized to more complex environments. In this paper, we propose a novel model named Large Decision Model with Memory (LDM$^2$), which leverages a dynamic memory mechanism to construct dynamic prompts, guiding the LLMs in making proper decisions according to the faced state. LDM$^2$ consists of two stages: memory formation and memory refinement. In the former stage, human behaviors are decomposed into state-action tuples utilizing the powerful summarizing ability of LLMs. Then, these tuples are stored in the memory, whose indices are generated by the LLMs, to facilitate the retrieval of the most relevant subset of memorized tuples based on the current state. In the latter stage, our LDM$^2$ employs tree exploration to discover more suitable decision processes and enrich the memory by adding valuable state-action tuples. The dynamic circle of exploration and memory enhancement provides LDM$^2$ a better understanding of the global environment. Extensive experiments conducted in two interactive environments have shown that our LDM$^2$ outperforms the baselines in terms of both score and success rate, which demonstrates its effectiveness.

09:00-10:30 (East Foyer / Virtual)
**Causal Inference from Text: Unveiling Interactions between Variables**
*Yuxiang Zhou and Yulan He*
Adjusting for latent covariates is crucial for estimating causal effects from observational textual data. Most existing methods only account for confounding covariates that affect both treatment and outcome, potentially leading to biased causal effects. This bias arises from insufficient consideration of non-confounding covariates, which are relevant only to either the treatment or the outcome. In this work, we aim to mitigate the bias by unveiling interactions between different variables to disentangle the non-confounding covariates when estimating causal effects from text. The disentangling process ensures covariates only contribute to their respective objectives, enabling independence between variables. Additionally, we impose a constraint to balance representations from the treated group and control group to alleviate selection bias. We conduct experiments on two different treatment factors under various scenarios, and the proposed model significantly outperforms recent strong baselines. Furthermore, our thorough analysis on earnings call transcripts demonstrates that our model can effectively disentangle the variables, and further investigations into real-world scenarios provide guidance for investors to make informed decisions.

09:00-10:30 (East Foyer / Virtual)
**DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text**
*Jinyan Su, Terry Yue Zhuo, Di Wang and Preslav Nakov*
With the rapid progress of Large language models (LLMs) and the huge amount of text they generate, it becomes impractical to manually distinguish whether a text is machine-generated. The growing use of LLMs in social media and education, prompts us to develop methods to detect machine-generated text, preventing malicious use such as plagiarism, misinformation, and propaganda. In this paper, we introduce two novel zero-shot methods for detecting machine-generated text by leveraging the Log-Rank information. One is called DetectLLM-LRR, which is fast and efficient, and the other is called DetectLLM-NPR, which is more accurate, but slower due to the need for perturbations. Our experiments on three datasets and seven language models show that our proposed methods improve over the state of the art by 3.9 and 1.75 AUROC points absolute. Moreover, DetectLLM-NPR needs fewer perturbations than previous work to achieve the same level of performance, which makes it more practical for real-world use. We also investigate the efficiency-performance trade-off based on users' preference for these two measures and provide intuition for using them in practice effectively. We release the data and the code of both methods in https://github.com/mbzuai-nlp/DetectLLM.

09:00-10:30 (East Foyer / Virtual)
**Robustness Tests for Automatic Machine Translation Metrics with Adversarial Attacks**
*Yichen Huang and Timothy Baldwin*
We investigate MT evaluation metric performance on adversarially-synthesized texts, to shed light on metric robustness. We experiment with word- and character-level attacks on three popular machine translation metrics: BERTScore, BLEURT, and COMET. Our human experiments validate that automatic metrics tend to overpenalize adversarially-degraded translations. We also identify inconsistencies in BERTScore ratings, where it judges the original sentence and the adversarially-degraded one as similar, while judging the degraded translation as notably worse than the original with respect to the reference. We identify patterns of brittleness that motivate more robust metric development.

09:00-10:30 (East Foyer / Virtual)
**The Locality and Symmetry of Positional Encodings**
*Lihu Chen, Gael Varoquaux and Fabian M. Suchanek*
Positional Encodings (PEs) are used to inject word-order information into transformer-based language models. While they can significantly enhance the quality of sentence representations, their specific contribution to language models is not fully understood, especially given recent findings that various positional encodings are insensitive to word order. In this work, we conduct a systematic study of positional encodings in **Bidirectional Masked Language Models** (BERT-style) , which complements existing work in three aspects: (1) We uncover the core function of PEs by identifying two common properties, Locality and Symmetry; (2) We show that the two properties are closely correlated with the performances of downstream tasks; (3) We quantify the weakness of current PEs by introducing two new probing tasks, on which current PEs perform poorly. We believe that these results are the basis for developing better PEs for transformer-based language models.

09:00-10:30 (East Foyer / Virtual)
**Image and Text: Fighting the same Battle? Super Resolution Learning for Imbalanced Text Classification**
*Romain Meunier, Benamara Farah, Véronique Moriceau and Patricia Stolf*
In this paper, we propose SRL4NLP, a new approach for data augmentation by drawing an analogy between image and text processing: Super-resolution learning. This method is based on using high-resolution images to overcome the problem of low resolution images. While this technique is a common usage in image processing when images have a low resolution or are too noisy, it has never been used in NLP. We therefore propose the first adaptation of this method for text classification and evaluate its effectiveness on urgency detection from tweets posted in crisis situations, a very challenging task where messages are scarce and highly imbalanced. We show that this strategy is efficient when compared to competitive state-of-the-art data augmentation techniques on several benchmarks datasets in two languages.

09:00-10:30 (East Foyer / Virtual)
**Learning to Abstract with Nonparametric Variational Information Bottleneck**
*Melika Behjati, Fabio James Fehr and James Henderson*
Learned representations at the level of characters, sub-words, words, and sentences, have each contributed to advances in understanding different NLP tasks and linguistic phenomena. However, learning textual embeddings is costly as they are tokenization specific and require different models to be trained for each level of abstraction. We introduce a novel language representation model which can learn to compress to different levels of abstraction at different layers of the same model. We apply Nonparametric Variational Information Bottleneck (NVIB) to stacked Transformer self-attention layers in the encoder, which encourages an information-theoretic compression of the representations through the model. We find that the layers within the model correspond to increasing levels of abstraction and that their representations are more linguistically informed. Finally, we show that NVIB compression results in a model which is more robust to adversarial perturbations.

09:00-10:30 (East Foyer / Virtual)
**Disentangling Structure and Style: Political Bias Detection in News by Inducing Document Hierarchy**
*Jiwoo Hong, Yejin Cho, Jiyoung Han, Jaemin Jung and James Thorne*
We address an important gap in detecting political bias in news articles. Previous works that perform document classification can be influenced by the writing style of each news outlet, leading to overfitting and limited generalizability. Our approach overcomes this limitation by considering both the sentence-level semantics and the document-level rhetorical structure, resulting in a more robust and style-agnostic approach to detecting political bias in news articles. We introduce a novel multi-head hierarchical attention model that effectively encodes the structure of long documents through a diverse ensemble of attention heads. While journalism follows a formalized rhetorical structure, the writing style may vary by news outlet. We demonstrate that our method overcomes this domain dependency and outperforms previous approaches for robustness and accuracy. Further analysis and human evaluation demonstrate the ability of our model to capture common discourse structures in journalism.

09:00-10:30 (East Foyer / Virtual)
**Aspect-Category Enhanced Learning with a Neural Coherence Model for Implicit Sentiment Analysis**
*Jin Cui, Fumiyo Fukumoto, Xinfeng Wang, Yoshimi Suzuki, Jiyi Li and Wanzeng Kong*
Aspect-based sentiment analysis (ABSA) has been widely studied since the explosive growth of social networking services. However, the recognition of implicit sentiments that do not contain obvious opinion words remains less explored. In this paper, we propose aspect-category enhanced learning with a neural coherence model (ELCoM). It captures document-level coherence by using contrastive learning, and sentence-level by a hypergraph to mine opinions from explicit sentences to aid implicit sentiment classification. To address the issue of sentences with different sentiment polarities in the same category, we perform cross-category enhancement to offset the impact of anomalous nodes in the hypergraph and obtain sentence representations with enhanced aspect-category. Extensive experiments on benchmark datasets show that the ELCoM achieves state-of-the-art performance. Our source codes and data are released at https://github.com/cuijin-23/ELCoM.

09:00-10:30 (East Foyer / Virtual)
**Dynamic Stance: Modeling Discussions by Labeling the Interactions**
*Blanca Calvo Figueras, Irene Baucells and Tommaso Caselli*
Stance detection is an increasingly popular task that has been mainly modeled as a static task, by assigning the expressed attitude of a text toward a given topic. Such a framing presents limitations, with trained systems showing poor generalization capabilities and being strongly topic-dependent. In this work, we propose modeling stance as a dynamic task, by focusing on the interactions between a message and their replies. For this purpose, we present a new annotation scheme that enables the categorization of all kinds of textual interactions. As a result, we have created a new corpus, the Dynamic Stance Corpus (DySC), consisting of three datasets in two middle-resourced languages: Catalan and Dutch. Our data analysis further supports our modeling decisions, empirically showing differences between the annotation of stance in static and dynamic contexts. We fine-tuned a series of monolingual and multilingual models on DySC, showing portability across topics and languages.

09:00-10:30 (East Foyer / Virtual)
**Responsible AI Considerations in Text Summarization Research: A Review of Current Practices**
*Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu and Adam Trischler*
AI and NLP publication venues have increasingly encouraged researchers to reflect on possible ethical considerations, adverse impacts, and other responsible AI issues their work might engender. However, for specific NLP tasks our understanding of how prevalent such issues are, or when and why these issues are likely to arise, remains limited. Focusing on text summarization—a common NLP task largely overlooked by the responsible AI community—we examine research and reporting practices in the current literature. We conduct a multi-round qualitative analysis of 333 summarization papers from the ACL Anthology published between 2020–2022. We focus on how, which, and when responsible AI issues are covered, which relevant stakeholders are considered, and mismatches between stated and realized research goals. We also discuss current evaluation practices and consider how authors discuss the limitations of both prior work and their own work. Overall, we find that relatively few papers engage with possible stakeholders or contexts of use, which limits their consideration of potential downstream adverse impacts or other responsible AI issues. Based on our findings, we make recommendations on concrete practices and research directions.

09:00-10:30 (East Foyer / Virtual)
**Task-Aware Self-Supervised Framework for Dialogue Discourse Parsing**
*Wei Li, Luyao Zhu, Wei Shao, Zonglin Yang and Erik Cambria*
Dialogue discourse parsing is a fundamental natural language processing task. It can benefit a series of conversation-related downstream tasks including dialogue summarization and emotion recognition in conversations. However, existing parsing approaches are constrained by predefined relation types, which can impede the adaptability of the parser for downstream tasks. To this end, we propose to introduce a task-aware paradigm to improve the versatility of the parser in this paper. Moreover, to alleviate error propagation and learning bias, we design a graph-based discourse parsing model termed DialogDP. Building upon the symmetrical property of matrix-embedded parsing graphs, we have developed an innovative self-supervised mechanism that leverages both bottom-up and top-down parsing strategies. This approach allows the

parsing graphs to mutually regularize and enhance each other. Empirical studies on dialogue discourse parsing datasets and a downstream task demonstrate the effectiveness and flexibility of our framework.

09:00-10:30 (East Foyer / Virtual)
**StyleBART: Decorate Pretrained Model with Style Adapters for Unsupervised Stylistic Headline Generation**
*Hanqing Wang, Yajing Luo, Boya Xiong, Guanhua Chen and Yun Chen*
Stylistic headline generation is the task to generate a headline that not only summarizes the content of an article, but also reflects a desired style that attracts users. As style-specific article-headline pairs are scarce, previous researches focus on unsupervised approaches with a standard headline generation dataset and mono-style corpora. In this work, we follow this line and propose StyleBART, an unsupervised approach for stylistic headline generation. Our method decorates the pretrained BART model with adapters that are responsible for different styles and allows the generation of headlines with diverse styles by simply switching the adapters. Different from previous works, StyleBART separates the task of style learning and headline generation, making it possible to freely combine the base model and the style adapters during inference. We further propose an inverse paraphrasing task to enhance the style adapters. Extensive automatic and human evaluations show that StyleBART achieves new state-of-the-art performance in the unsupervised stylistic headline generation task, producing high-quality headlines with the desired style.

09:00-10:30 (East Foyer / Virtual)
**Self-Polish: Enhance Reasoning in Large Language Models via Problem Refinement**
*Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang and Xuanjing Huang*
To enhance the multi-step reasoning capabilities of large language models, researchers have extensively explored prompting methods, notably the Chain-of-Thought (CoT) method which explicitly elicits human-like rationales. However, they have inadvertently overlooked the potential of enhancing model reasoning performance by formulating higher-quality problems. In this work, we start from the problem side and propose Self-Polish (SP), a novel method that facilitates the model's reasoning by guiding it to progressively refine the given problems to be more comprehensible and solvable. We also explore several automatic prompting varients and propose the Self-Polish prompt bank for the community. SP is orthogonal to all other prompting methods of answer/reasoning side like CoT, allowing for seamless integration with state-of-the-art techniques for further improvement. Thorough experiments show that the proposed method attains notable and consistent effectiveness on five reasoning benchmarks across different models. Furthermore, our method also showcases impressive performance on robustness evaluation. Codes and prompts are available at https://github.com/WooooDyy/Self-Polish.

09:00-10:30 (East Foyer / Virtual)
**Exploring the Potential of Large Language Models in Generating Code-Tracing Questions for Introductory Programming Courses**
*Aysa Xuemo Fan, Haoran Ranran Zhang, Luc Paquette and Rui Zhang*
In this paper, we explore the application of large language models (LLMs) for generating code-tracing questions in introductory programming courses. We designed targeted prompts for GPT4, guiding it to generate code-tracing questions based on code snippets and descriptions. We established a set of human evaluation metrics to assess the quality of questions produced by the model compared to those created by human experts. Our analysis provides insights into the capabilities and potential of LLMs in generating diverse code-tracing questions. Additionally, we present a unique dataset of human and LLM-generated tracing questions, serving as a valuable resource for both the education and NLP research communities. This work contributes to the ongoing dialogue on the potential uses of LLMs in educational settings.

09:00-10:30 (East Foyer / Virtual)
**InheritSumm: A General, Versatile and Compact Summarizer by Distilling from GPT**
*Yichong Xu, Ruochen Xu, Dan Iter, Yang Liu, Shuohang Wang, Chenguang Zhu and Michael Zeng*
While large models such as GPT-3 demonstrate exceptional performance in zeroshot and fewshot summarization tasks, their extensive serving and fine-tuning costs hinder their utilization in various applications. Conversely, previous studies have found that although automatic metrics tend to favor smaller fine-tuned models, the quality of the summaries they generate is inferior to that of larger models like GPT-3 when assessed by human evaluators. To address this issue, we propose InheritSumm, a versatile and compact summarization model derived from GPT-3.5 through distillation. InheritSumm not only exhibits comparable zeroshot and fewshot summarization capabilities to GPT-3.5 but is also sufficiently compact for fine-tuning purposes. Experimental results demonstrate that InheritSumm achieves similar or superior performance to GPT-3.5 in zeroshot and fewshot settings. Furthermore, it outperforms the previously established best small models in both prefix-tuning and full-data fine-tuning scenarios.

09:00-10:30 (East Foyer / Virtual)
**A Rewriting Approach for Gender Inclusivity in Portuguese**
*Leonor Veloso, Luisa Coheur and Rui Ribeiro*
In recent years, there has been a notable rise in research interest regarding the integration of gender-inclusive and gender-neutral language in natural language processing models. A specific area of focus that has gained practical and academic significant interest is gender-neutral rewriting, which involves converting binary-gendered text to its gender-neutral counterpart. However, current approaches to gender-neutral rewriting for gendered languages tend to rely on large datasets, which may not be an option for languages with fewer resources, such as Portuguese. In this paper, we present a rule-based and a neural-based tool for gender-neutral rewriting for Portuguese, a heavily gendered Romance language whose morphology creates different challenges from the ones tackled by other gender-neutral rewriters. Our neural approach relies on fine-tuning large multilingual machine translation models on examples generated by the rule-based model. We evaluate both models on texts from different sources and contexts. We provide the first Portuguese dataset explicitly containing gender-neutral language and neopronouns, as well as a manually annotated golden collection of 500 sentences that allows for evaluation of future work.

09:00-10:30 (East Foyer / Virtual)
**A Read-and-Select Framework for Zero-shot Entity Linking**
*Zhenran Xu, Yulin Chen, Baotian Hu and Min Zhang*
Zero-shot entity linking (EL) aims at aligning entity mentions to unseen entities to challenge the generalization ability. Previous methods largely focus on the candidate retrieval stage and ignore the essential candidate ranking stage, which disambiguates among entities and makes the final linking prediction. In this paper, we propose a read-and-select (ReS) framework by modeling the main components of entity disambiguation, i.e., mention-entity matching and cross-entity comparison. First, for each candidate, the reading module leverages mention context to output mention-aware entity representations, enabling mention-entity matching. Then, in the selecting module, we frame the choice of candidates as a sequence labeling problem, and all candidate representations are fused together to enable cross-entity comparison. Our method achieves the state-of-the-art performance on the established zero-shot EL dataset ZESHEL with a 2.55% micro-average accuracy gain, with no need for laborious multi-phase pre-training used in most of the previous work, showing the effectiveness of both mention-entity and cross-entity interaction.

09:00-10:30 (East Foyer / Virtual)
**Analysis of Style-Shifting on Social Media: Using Neural Language Model Conditioned by Social Meanings**

*Seiya Kawano, Shota Kanezaki, Angel Fernando Garcia Contreras, Akishige Yuguchi, Marie Katsurai and Koichiro Yoshino*
In this paper, we propose a novel framework for evaluating style-shifting in social media conversations. Our proposed framework captures changes in an individual's conversational style based on surprisals predicted by a personalized neural language model for individuals. Our personalized language model integrates not only the linguistic contents of conversations but also non-linguistic factors, such as social meanings, including group membership, personal attributes, and individual beliefs. We incorporate these factors directly or implicitly into our model, leveraging large, pre-trained language models and feature vectors derived from a relationship graph on social media. Compared to existing models, our personalized language model demonstrated superior performance in predicting an individual's language in a test set. Furthermore, an analysis of style-shifting utilizing our proposed metric based on our personalized neural language model reveals a correlation between our metric and various conversation factors as well as human evaluation of style-shifting.

09:00-10:30 (East Foyer / Virtual)
**RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning**
*Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li and Jiang Liu*
Automating radiology report generation can significantly alleviate radiologists' workloads. Previous research has primarily focused on realizing highly concise observations while neglecting the precise attributes that determine the severity of diseases (e.g., small pleural effusion). Since incorrect attributes will lead to imprecise radiology reports, strengthening the generation process with precise attribute modeling becomes necessary. Additionally, the temporal information contained in the historical records, which is crucial in evaluating a patient's current condition (e.g., heart size is unchanged), has also been largely disregarded. To address these issues, we propose RECAP, which generates precise and accurate radiology reports via dynamic disease progression reasoning. Specifically, RECAP first predicts the observations and progressions (i.e., spatiotemporal information) given two consecutive radiographs. It then combines the historical records, spatiotemporal information, and radiographs for report generation, where a disease progression graph and dynamic progression reasoning mechanism are devised to accurately select the attributes of each observation and progression. Extensive experiments on two publicly available datasets demonstrate the effectiveness of our model.

09:00-10:30 (East Foyer / Virtual)
**The Past, Present, and Future of Typological Databases in NLP**
*Emi Baylor, Esther Ploeger and Johannes Bjerva*
Typological information has the potential to be beneficial in the development of NLP models, particularly for low-resource languages. Unfortunately, current large-scale typological databases, notably WALS and Grambank, are inconsistent both with each other and with other sources of typological information, such as linguistic grammars. Some of these inconsistencies stem from coding errors or linguistic variation, but many of the disagreements are due to the discrete categorical nature of these databases. We shed light on this issue by systematically exploring disagreements across typological databases and resources, and their uses in NLP, covering the past and present. We next investigate the future of such work, offering an argument that a continuous view of typological features is clearly beneficial, echoing recommendations from linguistics. We propose that such a view of typology has significant potential in the future, including in language modeling in low-resource scenarios.

09:00-10:30 (East Foyer / Virtual)
**PROTEGE: Prompt-based Diverse Question Generation from Web Articles**
*Vinayak S Puranik, Anirban Majumder and Vineet Chaoji*
Rich and diverse knowledge bases (KB) are foundational building blocks for online knowledge sharing communities such as StackOverflow and Quora, and applications such as conversational assistants (aka chatbots). A popular format for knowledge bases is question-answer pairs (or FAQs), where questions are designed to accurately match a multitude of queries. In this paper, we address the problem of automatic creation of such Q&A-based knowledge bases from domain-specific, long-form textual content (e.g., web articles). Specifically, we consider the problem of question generation, which is the task of generating questions given a paragraph of text as input, with a goal to achieve both diversity and fidelity of the generated questions. Towards this goal we propose PROTEGE, a diverse question generation framework which consists of (1) a novel encoder-decoder based Large Language Model (LLM) architecture which can take a variety of prompts and generate a diverse set of candidate questions, and (2) a hill-climbing algorithm that maximizes a sub-modular objective function to balance diversity with fidelity. Through our experiments on three popular public Q&A datasets, we demonstrate that PROTEGE improves diversity by +16% and fidelity by +8% over diverse beam search and prompt-based baselines.

09:00-10:30 (East Foyer / Virtual)
**Toxicity, Morality, and Speech Act Guided Stance Detection**
*Apoorva Upadhyaya, Marco Fisichella and Wolfgang Nejdl*
In this work, we focus on the task of determining the public attitude toward various social issues discussed on social media platforms. Platforms such as Twitter, however, are often used to spread misinformation, fake news through polarizing views. Existing literature suggests that higher levels of toxicity prevalent in Twitter conversations often spread negativity and delay addressing issues. Further, the embedded moral values and speech acts specifying the intention of the tweet correlate with public opinions expressed on various topics. However, previous works, which mainly focus on stance detection, either ignore the speech act, toxic, and moral features of these tweets that can collectively help capture public opinion or lack an efficient architecture that can detect the attitudes across targets. Therefore, in our work, we focus on the main task of stance detection by exploiting the toxicity, morality, and speech act as auxiliary tasks. We propose a multitasking model TWISTED that initially extracts the valence, arousal, and dominance aspects hidden in the tweets and injects the emotional sense into the embedded text followed by an efficient attention framework to correctly detect the tweet's stance by using the shared features of toxicity, morality, and speech acts present in the tweet. Extensive experiments conducted on 4 benchmark stance detection datasets (SemEval-2016, P-Stance, COVID19-Stance, and ClimateChange) comprising different domains demonstrate the effectiveness and generalizability of our approach.

09:00-10:30 (East Foyer / Virtual)
**Strong and Efficient Baselines for Open Domain Conversational Question Answering**
*Andrei Catalin Coman, Gianni Barlacchi and Adrià de Gispert*
Unlike the Open Domain Question Answering (ODQA) setting, the conversational (ODConvQA) domain has received limited attention when it comes to reevaluating baselines for both efficiency and effectiveness. In this paper, we study the State-of-the-Art (SotA) Dense Passage Retrieval (DPR) retriever and Fusion-in-Decoder (FiD) reader pipeline, and show that it significantly underperforms when applied to ODConvQA tasks due to various limitations. We then propose and evaluate strong yet simple and efficient baselines, by introducing a fast reranking component between the retriever and the reader, and by performing targeted finetuning steps. Experiments on two ODConvQA tasks, namely TopiOCQA and OR-QuAC, show that our method improves the SotA results, while reducing reader's latency by 60%. Finally, we provide new and valuable insights into the development of challenging baselines that serve as a reference for future, more intricate approaches, including those that leverage Large Language Models (LLMs).

09:00-10:30 (East Foyer / Virtual)
**BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs**

*Giuseppe Samo, Vivi Nastase, Chunyang Jiang and Paola Merlo*
Current NLP models appear to be achieving performance comparable to human capabilities on well-established benchmarks. New benchmarks are now necessary to test deeper layers of understanding of natural languages by these models. Blackbird's Language Matrices are a recently developed framework that draws inspiration from tests of human analytic intelligence. The BLM task has revealed that successful performances in previously studied linguistic problems do not yet stem from a deep understanding of the generative factors that define these problems. In this study, we define a new BLM task for predicate-argument structure, and develop a structured dataset for its investigation, concentrating on the spray-load verb alternations in English, as a case study. The context sentences include one alternant from the spray-load alternation and the target sentence is the other alternant, to be chosen among a minimally contrastive and adversarial set of answers. We describe the generation process of the dataset and the reasoning behind the generating rules. The dataset aims to facilitate investigations into how verb information is encoded in sentence embeddings and how models generalize to the complex properties of argument structures. Benchmarking experiments conducted on the dataset and qualitative error analysis on the answer set reveal the inherent challenges associated with the problem even for current high-performing representations.

09:00-10:30 (East Foyer / Virtual)
**Search Augmented Instruction Learning**
*Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen M. Meng and James R. Glass*
Large language models (LLMs) have been significantly improved by instruction fine-tuning, but still lack transparency and the ability to utilize up-to-date knowledge and information. In this work, we propose search-augmented instruction learning (SAIL), which grounds the language generation and instruction following abilities on complex search results generated by in-house and external search engines. With an instruction tuning corpus, we collect search results for each training case from different search APIs and domains, and construct a new search-grounded training set containing (instruction, grounding information, response) triplets. We then fine-tune the LLaMA-7B model on the constructed training set. Since the collected results contain unrelated and disputing languages, the model needs to learn to ground on trustworthy search results, filter out distracting passages, and generate the target response. The search result-denoising process entails explicit trustworthy information selection and multi-hop reasoning, since the retrieved passages might be informative but not contain the instruction-following answer. Experiments show that the fine-tuned SAIL-7B model has a strong instruction-following ability, and it performs significantly better on transparency-sensitive tasks, including open-ended question answering and fact checking.

09:00-10:30 (East Foyer / Virtual)
**Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture**
*Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler and Dakuo Wang*
Real-world domain experts (e.g., doctors) rarely annotate only a decision label in their day-to-day workflow without providing explanations. Yet, existing low-resource learning techniques, such as Active Learning (AL), that aim to support human annotators mostly focus on the label while neglecting the natural language explanation of a data point. This work proposes a novel AL architecture to support experts' real-world need for label and explanation annotations in low-resource scenarios. Our AL architecture leverages an explanation-generation model to produce explanations guided by human explanations, a prediction model that utilizes generated explanations toward prediction faithfully, and a novel data diversity-based AL sampling strategy that benefits from the explanation annotations. Automated and human evaluations demonstrate the effectiveness of incorporating explanations into AL sampling and the improved human annotation efficiency and trustworthiness with our AL architecture. Additional ablation studies illustrate the potential of our AL architecture for transfer learning, generalizability, and integration with large language models (LLMs). While LLMs exhibit exceptional explanation-generation capabilities for relatively simple tasks, their effectiveness in complex real-world tasks warrants further in-depth study.

09:00-10:30 (East Foyer / Virtual)
**Is a Prestigious Job the same as a Prestigious Country? A Case Study on Multilingual Sentence Embeddings and European Countries**
*Jindřich Libovický*
We study how multilingual sentence representations capture European countries and occupations and how this differs across European languages. We prompt the models with templated sentences that we machine-translate into 12 European languages and analyze the most prominent dimensions in the embeddings. Our analysis reveals that the most prominent feature in the embedding is the political distinction between Eastern and Western Europe and the country's economic strength in terms of GDP. When prompted specifically for job prestige, the embedding space clearly distinguishes high and low-prestige jobs. The occupational dimension is uncorrelated with the most dominant country dimensions in three out of four studied models. The exception is a small distilled model that exhibits a connection between occupational prestige and country of origin, which is a potential source of nationality-based discrimination. Our findings are consistent across languages.

09:00-10:30 (East Foyer / Virtual)
**Domain Private Transformers for Multi-Domain Dialog Systems**
*Anmol Kabra and Ethan R. Elenberg*
Large, general purpose language models have demonstrated impressive performance across many different conversational domains. While multi-domain language models achieve low overall perplexity, their outputs are not guaranteed to stay within the domain of a given input prompt. This paper proposes *domain privacy* as a novel way to quantify how likely a conditional language model will leak across domains. We also develop policy functions based on token-level domain classification, and propose an efficient fine-tuning method to improve the trained model's domain privacy. Experiments on membership inference attacks show that our proposed method has comparable resiliency to methods adapted from recent literature on differentially private language models.

09:00-10:30 (East Foyer / Virtual)
**CoVariance-based Causal Debiasing for Entity and Relation Extraction**
*Lin Ren, Yongbin Liu, Yixin Cao and Chunping Ouyang*
Joint entity and relation extraction tasks aim to recognize named entities and extract relations simultaneously. Suffering from a variety of data biases, such as data selection bias, and distribution bias (out of distribution, long-tail distribution), serious concerns can be witnessed to threaten the model's transferability, robustness, and generalization. In this work, we address the above problems from a causality perspective. We propose a novel causal framework called **co**variance and **y**variance **op**timization framework (OVO) to optimize feature representations and conduct general debiasing. In particular, the proposed **c**ovariance **op**timizing (COP) minimizes characterizing features' covariance for alleviating the selection and distribution bias and enhances feature representation in the feature space. Furthermore, based on the causal backdoor adjustment, we propose $underline{\textbf{o}}$variance **op**timizing (VOP) separates samples in terms of label information and minimizes the variance of each dimension in the feature vectors of the same class label for mitigating the distribution bias further. By applying it to three strong baselines in two widely used datasets, the results demonstrate the effectiveness and generalization of OVO for joint entity and relation extraction tasks. Furthermore, a fine-grained analysis reveals that OVO possesses the capability to mitigate the impact of long-tail distribution.

09:00-10:30 (East Foyer / Virtual)
**Misery Loves Complexity: Exploring Linguistic Complexity in the Context of Emotion Detection**
*Pranaydeep Singh, Luna De Bruyne, Orphée De Clercq and Els Lefever*
Given the omnipresence of social media in our society, thoughts and opinions are being shared online in an unprecedented manner. This means that both positive and negative emotions can be equally and freely expressed. However, the negativity bias posits that human beings are inherently drawn to and more moved by negativity and, as a consequence, negative emotions get more traffic. Correspondingly, when writing about emotions this negativity bias could lead to expressions of negative emotions that are linguistically more complex. In this paper, we attempt to use readability and linguistic complexity metrics to better understand the manifestation of emotions on social media platforms like Reddit based on the widely-used GoEmotions dataset. We demonstrate that according to most metrics, negative emotions indeed tend to generate more complex text than positive emotions. In addition, we examine whether a higher complexity hampers the automatic identification of emotions. To answer this question, we fine-tuned three state-of-the-art transformers (BERT, RoBERTa, and SpanBERT) on the same emotion detection dataset. We demonstrate that these models often fail to predict emotions for the more complex texts. More advanced LLMs like RoBERTa and SpanBERT also fail to improve by significant margins on complex samples. This calls for a more nuanced interpretation of the emotion detection performance of transformer models. We make the automatically annotated data available for further research at: https://huggingface.co/datasets/pranaydeeps/CAMEO

09:00-10:30 (East Foyer / Virtual)
**Ensemble-Instruct: Instruction Tuning Data Generation with a Heterogeneous Mixture of LMs**
*Young-Suk Lee, Md Arafat Sultan, Yousef El-Kurdi, Tahira Naseem, Asim Munawar, Radu Florian, Salim Roukos and Ramón Fernandez Astudillo*
Using in-context learning (ICL) for data generation, techniques such as Self-Instruct (Wang et al., 2023) or the follow-up Alpaca (Taori et al., 2023) can train strong conversational agents with only a small amount of human supervision. One limitation of these approaches is that they resort to very large language models (around 175B parameters) that are also proprietary and non-public. Here we explore the application of such techniques to language models that are much smaller (around 10B–40B parameters) and have permissive licenses. We find the Self-Instruct approach to be less effective at these sizes and propose new ICL methods that draw on two main ideas: (a) categorization and simplification of the ICL templates to make prompt learning easier for the LM, and (b) ensembling over multiple LM outputs to help select high-quality synthetic examples. Our algorithm leverages the 175 Self-Instruct seed tasks and employs separate pipelines for instructions that require an input and instructions that do not. Empirical investigations with different LMs show that: (1) Our proposed method yields higher-quality instruction tuning data than Self-Instruct, (2) It improves performances of both vanilla and instruction-tuned LMs by significant margins, and (3) Smaller instruction-tuned LMs generate more useful examples than their larger un-tuned counterparts.

09:00-10:30 (East Foyer / Virtual)
**SDOH-NLI: a Dataset for Inferring Social Determinants of Health from Clinical Notes**
*Adam D Lelkes, Eric Loreaux, Tal Schuster, Ming-Jun Chen and Alvin Rajkomar*
Social and behavioral determinants of health (SDOH) play a significant role in shaping health outcomes, and extracting these determinants from clinical notes is a first step to help healthcare providers systematically identify opportunities to provide appropriate care and address disparities. Progress on using NLP methods for this task has been hindered by the lack of high-quality publicly available labeled data, largely due to the privacy and regulatory constraints on the use of real patients' information. This paper introduces a new dataset, SDOH-NLI, that is based on publicly available notes and which we release publicly. We formulate SDOH extraction as a natural language inference task, and provide binary textual entailment labels obtained from human raters for a cross product of a set of social history snippets as premises and SDOH factors as hypotheses. Our dataset differs from standard NLI benchmarks in that our premises and hypotheses are obtained independently. We evaluate both "off-the-shelf" entailment models as well as models fine-tuned on our data, and highlight the ways in which our dataset appears more challenging than commonly used NLI datasets.

### Industry 4

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**Automatic Marketing Theme and Commodity Construction System for E-commerce**
*Zhiping Wang, Peng Lin, Hainan Zhang, Hongshen Chen, Tianhao Li, Zhuoye Ding, Sulong Xu and Jinghe Hu*
When consumers' shopping needs are concentrated, they are more interested in the collection of commodities under the specific marketing theme. Therefore, mining marketing themes and their commodities collections can help customers save shopping costs and improve user clicks and purchases for recommendation system. However, the current system invites experts to write marketing themes and select the relevant commodities, which suffer from difficulty in mass production, poor timeliness and low online indicators. Therefore, we propose a automatic marketing theme and commodity construction system, which can not only generate popular marketing themes and select the relevant commodities automatically, but also improve the theme online effectiveness in the recommendation system. Specifically, we firstly utilize the pretrained language model to generate the marketing themes. And then, we utilize the theme-commodity consistency module to select the relevant commodities for the above generative theme. What's more, we also build the indicator simulator to evaluate the effectiveness of the above generative theme. When the indicator is lower, the above selective commodities will be input into the theme-rewriter module to generate more efficient marketing themes. Finally, we utilize the human screening to control the system quality. Both the offline experiments and online A/B test demonstrate the superior performance of our proposed system compared with state-of-the-art methods.

# Coffee Break

10:30-11:00 - Location: West Foyer

# Session 6: Oral & Poster - 11:00-12:30

### Interpretability, Interactivity, and Analysis of Models for NLP 2

11:00-12:30 (East Ballroom)

---

11:00-11:15 (East Ballroom)

**Absolute Position Embedding Learns Sinusoid-like Waves for Attention Based on Relative Position**
*Yuji Yamamoto and Takuya Matsuzaki*
Attention weight is a clue to interpret how a Transformer-based model makes an inference. In some attention heads, the attention focuses on the neighbors of each token. This allows the output vector of each token to depend on the surrounding tokens and contributes to make the inference context-dependent. We analyze the mechanism behind the concentration of attention on nearby tokens. We show that the phenomenon emerges as follows: (1) learned position embedding has sinusoid-like components, (2) such components are transmitted to the query and the key in the self-attention, (3) the attention head shifts the phases of the sinusoid-like components so that the attention concentrates on nearby tokens at specific relative positions. In other words, a certain type of Transformer-based model acquires the sinusoidal positional encoding to some extent on its own through Masked Language Modeling.

11:15-11:30 (East Ballroom)
**Statistical Depth for Ranking and Characterizing Transformer-Based Text Embeddings**
*Parker Seegmiller and Sarah Masud Preum*
The popularity of transformer-based text embeddings calls for better statistical tools for measuring distributions of such embeddings. One such tool would be a method for ranking texts within a corpus by centrality, i.e. assigning each text a number signifying how representative that text is of the corpus as a whole. However, an intrinsic center-outward ordering of high-dimensional text representations is not trivial. A *statistical depth* is a function for ranking $k$-dimensional objects by measuring centrality with respect to some observed $k$-dimensional distribution. We adopt a statistical depth to measure distributions of transformer-based text embeddings, *transformer-based text embedding (TTE) depth*, and introduce the practical use of this depth for both modeling and distributional inference in NLP pipelines. We first define TTE depth and an associated rank sum test for determining whether two corpora differ significantly in embedding space. We then use TTE depth for the task of in-context learning prompt selection, showing that this approach reliably improves performance over statistical baseline approaches across six text classification tasks. Finally, we use TTE depth and the associated rank sum test to characterize the distributions of synthesized and human-generated corpora, showing that five recent synthetic data augmentation processes cause a measurable distributional shift away from associated human-generated text.

11:30-11:45 (East Ballroom)
**Explaining Interactions Between Text Spans**
*Sagnik Ray Choudhury, Pepa Atanasova and Isabelle Augenstein*
Reasoning over spans of tokens from different parts of the input is essential for natural language understanding (NLU) tasks such as fact-checking (FC), machine reading comprehension (MRC) or natural language inference (NLI). However, existing highlight-based explanations primarily focus on identifying individual important features or interactions only between adjacent tokens or tuples of tokens. Most notably, there is a lack of annotations capturing the human decision-making process with respect to the necessary interactions for informed decision-making in such tasks. To bridge this gap, we introduce SpanEx, a multi-annotator dataset of human span interaction explanations for two NLU tasks: NLI and FC. We then investigate the decision-making processes of multiple fine-tuned large language models in terms of the employed connections between spans in separate parts of the input and compare them to the human reasoning processes. Finally, we present a novel community detection based unsupervised method to extract such interaction explanations. We make the code and the dataset available on [Github](https://github.com/copenlu/spanex). The dataset is also available on [Huggingface datasets](https://huggingface.co/datasets/copenlu/spanex).

11:45-12:00 (East Ballroom)
**Bridging Information-Theoretic and Geometric Compression in Language Models**
*Emily Cheng, Corentin Kervadec and Marco Baroni*
For a language model (LM) to faithfully model human language, it must compress vast, potentially infinite information into relatively few dimensions. We propose analyzing compression in (pre-trained) LMs from two points of view: geometric and information-theoretic. We demonstrate that the two views are highly correlated, such that the intrinsic geometric dimension of linguistic data predicts their coding length under the LM. We then show that, in turn, high compression of a linguistic dataset predicts rapid adaptation to that dataset, confirming that being able to compress linguistic information is an important part of successful LM performance. As a practical byproduct of our analysis, we evaluate a battery of intrinsic dimension estimators for the first time on linguistic data, showing that only some encapsulate the relationship between information-theoretic compression, geometric compression, and ease-of-adaptation.

12:00-12:15 (East Ballroom)
**What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability**
*Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández and Barbara Plank*
In Natural Language Generation (NLG) tasks, for any input, multiple communicative goals are plausible, and any goal can be put into words, or produced, in multiple ways. We characterise the extent to which human production varies lexically, syntactically, and semantically across four NLG tasks, connecting human production variability to aleatoric or data uncertainty. We then inspect the space of output strings shaped by a generation system's predicted probability distribution and decoding algorithm to probe its uncertainty. For each test input, we measure the generator's calibration to human production variability. Following this instance-level approach, we analyse NLG models and decoding strategies, demonstrating that probing a generator with multiple samples and, when possible, multiple references, provides the level of detail necessary to gain understanding of a model's representation of uncertainty.

12:15-12:30 (East Ballroom)
**Data Factors for Better Compositional Generalization**
*Xiang Zhou, Yichen Jiang and Mohit Bansal*
Recent diagnostic datasets on compositional generalization, such as SCAN (Lake and Baroni, 2018) and COGS (Kim and Linzen, 2020), expose severe problems in models trained from scratch on these datasets. However, in contrast to this poor performance, state-of-the-art models trained on larger and more general datasets show better generalization ability. In this work, to reconcile this inconsistency, we conduct an empirical analysis by training Transformer models on a variety of training sets with different data factors, including dataset scale, pattern complexity, example difficulty, etc. First, we show that increased dataset complexity can lead to better generalization behavior on multiple different generalization challenges. To further understand this improvement, we show two axes of the benefit from more complex datasets: they provide more diverse examples so compositional understanding becomes more effective, and they also prevent ungeneralizable memorization of the examples due to reduced example repetition frequency. Finally, we explore how training examples of different difficulty levels influence generalization differently. On synthetic datasets, simple examples invoke stronger compositionality than hard examples do. On larger-scale real language datasets, while hard examples become more important potentially to ensure decent data coverage, a balanced mixture of simple and hard examples manages to induce the strongest generalizability.

## Language Modeling and Analysis of Language Models 2

11:00-12:30 (Central 1 Ballroom)

11:00-11:15 (Central 1 Ballroom)
**Inverse Scaling Can Become U-Shaped**
*Jason Wei, Najoung Kim, Yi Tay and Quoc V Le*
Scaling up language models has been empirically shown to improve performance on a wide range of downstream tasks. However, if we were to observe worse performance as a function of scale (inverse scaling) on certain tasks, this would indicate that scaling can also encourage behaviors that are misaligned with human preferences. The Inverse Scaling Prize (McKenzie et al. 2023) identified eleven such inverse scaling tasks, evaluated on models of up to 280B parameters and up to 500 zettaFLOPs of training compute. This paper takes a closer look at these inverse scaling tasks. In this paper, we evaluate models of up to 540B parameters, trained on five times more compute than those evaluated in the Inverse Scaling Prize. With this increased range of model sizes and compute, only four out of the eleven tasks remain inverse scaling. Six tasks exhibit U-shaped scaling, where performance decreases up to a certain size, and then increases again up to the largest model evaluated (the one remaining task displays positive scaling). In addition, 1-shot examples and chain-of-thought can help mitigate undesirable scaling patterns even further. U-shaped scaling suggests that the inverse scaling trend observed in McKenzie et al. (2023) may not continue to hold for larger models, which we attribute to the presence of distractor tasks that only sufficiently large models can avoid.

11:15-11:30 (Central 1 Ballroom)
**Revisiting Instruction Fine-tuned Model Evaluation to Guide Industrial Applications**
*Manuel Faysse, Gautier Viaud, Céline Hudelot and Pierre Colombo*
Instruction Fine-Tuning (IFT) is a powerful paradigm that strengthens the zero-shot capabilities of Large Language Models (LLMs), but in doing so induces new evaluation metric requirements. We show LLM-based metrics to be well adapted to these requirements, and leverage them to conduct an investigation of task-specialization strategies, quantifying the trade-offs that emerge in practical industrial settings. Our findings offer practitioners actionable insights for real-world IFT model deployment.

11:30-11:45 (Central 1 Ballroom)
**FinGPT: Large Generative Models for a Small Language**
*Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kristiina Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao and Sampo Pyysalo*
Large language models (LLMs) excel in many tasks in NLP and beyond, but most open models have very limited coverage of smaller languages and LLM work tends to focus on languages where nearly unlimited data is available for pretraining. In this work, we study the challenges of creating LLMs for Finnish, a language spoken by less than 0.1% of the world population. We compile an extensive dataset of Finnish combining web crawls, news, social media and eBooks. We pursue two approaches to pretrain models: 1) we train seven monolingual models from scratch (186M to 13B parameters) dubbed FinGPT, 2) we continue the pretraining of the multilingual BLOOM model on a mix of its original training data and Finnish, resulting in a 176 billion parameter model we call BLUUMI. For model evaluation, we introduce FIN-bench, a version of BIG-bench with Finnish tasks. We also assess other model qualities such as toxicity and bias. Our models and tools are openly available at https://turkunlp.org/gpt3-finnish.

11:45-12:00 (Central 1 Ballroom)
**Consistency Analysis of ChatGPT**
*Myeongjun Erik Jang and Thomas Lukasiewicz*
ChatGPT has gained a huge popularity since its introduction. Its positive aspects have been reported through many media platforms, and some analyses even showed that ChatGPT achieved a decent grade in professional exams, adding extra support to the claim that AI can now assist and even replace humans in industrial fields. Others, however, doubt its reliability and trustworthiness. This paper investigates the trustworthiness of ChatGPT and GPT-4 regarding logically consistent behaviour, focusing specifically on semantic consistency and the properties of negation, symmetric, and transitive consistency. Our findings suggest that while both models appear to show an enhanced language understanding and reasoning ability, they still frequently fall short of generating logically consistent predictions. We also ascertain via experiments that prompt designing, few-shot learning and employing larger large language models (LLMs) are unlikely to be the ultimate solution to resolve the inconsistency issue of LLMs.

12:00-12:15 (Central 1 Ballroom)
**How Abstract Is Linguistic Generalization in Large Language Models? Experiments with Argument Structure**
*Michael Wilson, Jackson Petty and Robert Frank*
Language models are typically evaluated on their success at predicting the distribution of specific words in specific contexts. Yet linguistic knowledge also encodes relationships between contexts, allowing inferences between word distributions. We investigate the degree to which pre-trained Transformer-based large language models (LLMs) represent such relationships, focusing on the domain of argument structure. We find that LLMs perform well in generalizing the distribution of a novel noun argument between related contexts that were seen during pre-training (e.g., the active object and passive subject of the verb spray), succeeding by making use of the semantically-organized structure of the embedding space for word embeddings. However, LLMs fail at generalizations between related contexts that have not been observed during pre-training, but which instantiate more abstract, but well-attested structural generalizations (e.g., between the active object and passive subject of an arbitrary verb). Instead, in this case, LLMs show a bias to generalize based on linear order. This finding points to a limitation with current models and points to a reason for which their training is data-intensive.

12:15-12:30 (Central 1 Ballroom)
**How is a "Kitchen Chair" like a "Farm Horse"? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models**
*Mark Ormerod, Jesús Martínez del Rincón and Barry Devereux*
Despite the success of Transformer-based language models in a wide variety of natural language processing tasks, our understanding of how these models process a given input in order to represent task-relevant information remains incomplete. In this work, we focus on semantic composition and examine how Transformer-based language models represent semantic information related to the meaning of English noun-noun compounds. We probe Transformer-based language models for their knowledge of the thematic relations that link the head nouns and modifier words of compounds (e.g., KITCHEN CHAIR: a chair located in a kitchen). Firstly, using a dataset featuring groups of compounds with shared lexical or semantic features, we find that token representations of six Transformer-based language models distinguish between pairs of compounds based on whether they use the same thematic relation. Secondly, we utilize fine-grained vector representations of compound semantics derived from human annotations, and find that token vectors from several models elicit a strong signal of the semantic relations used in the compounds. In a novel 'compositional probe' setting, where we compare the semantic relation signal in mean-pooled token vectors of compounds to mean-pooled token vectors when the two constituent words appear in separate sentences, we find that the

Transformer-based language models that best represent the semantics of noun-noun compounds also do so substantially better than in the control condition where the two constituent works are processed separately. Overall, our results shed light on the ability of Transformer-based language models to support compositional semantic processes in representing the meaning of noun-noun compounds.

## Multilinguality and Linguistic Diversity 2

11:00-12:30 (Central 3 Ballroom)

---

11:00-11:15 (Central 3 Ballroom)
**Multilingual Large Language Models Are Not (Yet) Code-Switchers**
*Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata and Alham Fikri Aji*
Multilingual Large Language Models (LLMs) have recently shown great capabilities in a wide range of tasks, exhibiting state-of-the-art performance through zero-shot or few-shot prompting methods. While there have been extensive studies on their abilities in monolingual tasks, the investigation of their potential in the context of code-switching (CSW), the practice of alternating languages within an utterance, remains relatively uncharted. In this paper, we provide a comprehensive empirical analysis of various multilingual LLMs, benchmarking their performance across four tasks: sentiment analysis, machine translation, summarization and word-level language identification. Our results indicate that despite multilingual LLMs exhibiting promising outcomes in certain tasks using zero or few-shot prompting, they still underperform in comparison to fine-tuned models of much smaller scales. We argue that current "multilingualism' in LLMs does not inherently imply proficiency with code-switching texts, calling for future research to bridge this discrepancy.

11:15-11:30 (Central 3 Ballroom)
**Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages**
*Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang and Wanxiang Che*
Chain-of-thought (CoT) is capable of eliciting models to explicitly generate reasoning paths, thus promoting reasoning accuracy and attracting increasing attention. Specifically, zero-shot CoT achieves remarkable improvements in a wide range of reasoning tasks by simply instructing the LLM with the prompt "Let's think step by step!". Despite the success of zero-shot CoT, the existing zero-shot prompting techniques remain limited to a single language, making it challenging to generalize to other languages and hindering global development. In this work, we introduce cross-lingual prompting (CLP), aiming to improve zero-shot CoT reasoning across languages. Specifically, CLP consists of two main components: (1) cross-lingual alignment prompting and (2) task-specific solver prompting. The cross-lingual alignment prompting is responsible for aligning representations across different languages, whereas the task-specific solver prompting is used to generate the final chain of thoughts and results for the reasoning task. In addition, we further introduce cross-lingual self-consistent prompting (CLSP) to ensemble different reasoning paths across languages. Our experimental evaluations on several benchmarks demonstrate that CLP and CLSP significantly outperform the existing prompting methods and achieve state-of-the-art performance. We hope this work will inspire further breakthroughs in cross-lingual CoT.

11:30-11:45 (Central 3 Ballroom)
**Investigating Bias in Multilingual Language Models: Cross-Lingual Transfer of Debiasing Techniques**
*Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt and Bart Baesens*
This paper investigates the transferability of debiasing techniques across different languages within multilingual models. We examine the applicability of these techniques in English, French, German, and Dutch. Using multilingual BERT (mBERT), we demonstrate that cross-lingual transfer of debiasing techniques is not only feasible but also yields promising results. Surprisingly, our findings reveal no performance disadvantages when applying these techniques to non-English languages. Using translations of the CrowS-Pairs dataset, our analysis identifies SentenceDebias as the best technique across different languages, reducing bias in mBERT by an average of 13%. We also find that debiasing techniques with additional pretraining exhibit enhanced cross-lingual effectiveness for the languages included in the analyses, particularly in lower-resource languages. These novel insights contribute to a deeper understanding of bias mitigation in multilingual language models and provide practical guidance for debiasing techniques in different language contexts.

11:45-12:00 (Central 3 Ballroom)
**FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models**
*Konstantin Dobler and Gerard de Melo*
Using model weights pretrained on a high-resource language as a warm start can reduce the need for data and compute to obtain high-quality language models for other, especially low-resource, languages. However, if we want to use a new tokenizer specialized for the target language, we cannot transfer the source model's embedding matrix. In this paper, we propose FOCUS - **F**ast **O**verlapping Token **C**ombinations **U**sing **S**parsemax, a novel embedding initialization method that effectively initializes the embedding matrix for a new tokenizer based on information in the source model's embedding matrix. FOCUS represents newly added tokens as combinations of tokens in the overlap of the source and target vocabularies. The overlapping tokens are selected based on semantic similarity in an auxiliary static token embedding space. We focus our study on using the multilingual XLM-R as a source model and empirically show that FOCUS outperforms random initialization and previous work on language modeling and on a range of downstream tasks (NLI, QA, and NER). We publish our model checkpoints and code on GitHub.

12:00-12:15 (Central 3 Ballroom)
**GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP**
*Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi and Muhammad Abdul-Mageed*
ChatGPT's emergence heralds a transformative phase in NLP, particularly demonstrated through its excellent performance on many English benchmarks. However, the model's efficacy across diverse linguistic contexts remains largely uncharted territory. This work aims to bridge this knowledge gap, with a primary focus on assessing ChatGPT's capabilities on Arabic languages and dialectal varieties. Our comprehensive study conducts a large-scale automated and human evaluation of ChatGPT, encompassing 44 distinct language understanding and generation tasks on over 60 different datasets. To our knowledge, this marks the first extensive performance analysis of ChatGPT's deployment in Arabic NLP. Our findings indicate that, despite its remarkable performance in English, ChatGPT is consistently surpassed by smaller models that have undergone finetuning on Arabic. We further undertake a meticulous comparison of ChatGPT and GPT-4's Modern Standard Arabic (MSA) and Dialectal Arabic (DA), unveiling the relative shortcomings of both models in handling Arabic dialects compared to MSA. Although we further explore and confirm the utility of employing GPT-4 as a potential alternative for human evaluation, our work adds to a growing body of research underscoring the limitations of ChatGPT.

12:15-12:30 (Central 3 Ballroom)
**Shared Lexical Items as Triggers of Code Switching**
*Shuly Wintner, Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak and Yuval Nov*

Why do bilingual speakers code-switch (mix their two languages)? Among the several theories that attempt to explain this natural and ubiquitous phenomenon, the Triggering Hypothesis relates code-switching to the presence of lexical triggers, specifically cognates and proper names, adjacent to the switch point. We provide a fuller, more nuanced and refined exploration of the triggering hypothesis, based on five large datasets in three language pairs, reflecting both spoken and written bilingual interactions. Our results show that words that are assumed to reside in a mental lexicon shared by both languages indeed trigger code-switching; that the tendency to switch depends on the distance of the trigger from the switch point; and on whether the trigger precedes or succeeds the switch; but not on the etymology of the trigger words. We thus provide strong, robust, evidence-based confirmation to several hypotheses on the relationships between lexical triggers and code-switching.

## Natural Language Generation 2

11:00-12:30 (West 1 Ballroom)

---

11:00-11:15 (West 1 Ballroom)
**MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark**
*Dominik Macko, Robert Moro, Adaku Uchendu, Jason S Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko and Maria Bielikova*
There is a lack of research into capabilities of recent LLMs to generate convincing text in languages other than English and into performance of detectors of machine-generated text in multilingual settings. This is also reflected in the available benchmarks which lack authentic texts in languages other than English and predominantly cover older generators. To fill this gap, we introduce MULTITuDE, a novel benchmarking dataset for multilingual machine-generated text detection comprising of 74,081 authentic and machine-generated texts in 11 languages (ar, ca, cs, de, en, es, nl, pt, ru, uk, and zh) generated by 8 multilingual LLMs. Using this benchmark, we compare the performance of zero-shot (statistical and black-box) and fine-tuned detectors. Considering the multilinguality, we evaluate 1) how these detectors generalize to unseen languages (linguistically similar as well as dissimilar) and unseen LLMs and 2) whether the detectors improve their performance when trained on multiple languages.

11:15-11:30 (West 1 Ballroom)
**Active Learning for Natural Language Generation**
*Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim and Liat Ein-Dor*
The field of Natural Language Generation (NLG) suffers from a severe shortage of labeled data due to the extremely expensive and time-consuming process involved in manual annotation. A natural approach for coping with this problem is active learning (AL), a well-known machine learning technique for improving annotation efficiency by selectively choosing the most informative examples to label. However, while AL has been well-researched in the context of text classification, its application to NLG remains largely unexplored. In this paper, we present a first systematic study of active learning for NLG, considering a diverse set of tasks and multiple leading selection strategies, and harnessing a strong instruction-tuned model. Our results indicate that the performance of existing AL strategies is inconsistent, surpassing the baseline of random example selection in some cases but not in others. We highlight some notable differences between the classification and generation scenarios, and analyze the selection behaviors of existing AL strategies. Our findings motivate exploring novel approaches for applying AL to generation tasks.

11:30-11:45 (West 1 Ballroom)
**Interactive Text Generation**
*Felix Faltings, Michel Galley, Kianté Brantley, Baolin Peng, Weixin Cai, Yizhe Zhang, Jianfeng Gao and Bill Dolan*
Users interact with text, image, code, or other editors on a daily basis. However, machine learning models are rarely trained in the settings that reflect the interactivity between users and their editor. This is understandable as training AI models with real users is not only slow and costly, but what these models learn may be specific to user interface design choices. Unfortunately, this means most of the research on text, code, and image generation has focused on non-interactive settings, whereby the model is expected to get everything right without accounting for any input from a user who may be willing to help. We introduce a new Interactive Text Generation task that allows training generation models interactively without the costs of involving real users, by using user simulators that provide edits that guide the model towards a given target text. We train our interactive models using Imitation Learning, and our experiments against competitive non-interactive generation models show that models trained interactively are superior to their non-interactive counterparts, even when all models are given the same budget of user inputs or edits.

11:45-12:00 (West 1 Ballroom)
**INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback**
*Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang and Lei Li*
Automatically evaluating the quality of language generation is critical. Although recent learned metrics show high correlation with human judgement, these metrics do not provide explicit explanation of their verdict, nor associate the scores with defects in the generated text. To address this limitation, we present INSTRUCTSCORE, a fine-grained explainable evaluation metric for text generation. By harnessing both explicit human instruction and the implicit knowledge of GPT-4, we fine-tune a text evaluation metric based on LLaMA, producing both a score for generated text and a human readable diagnostic report. We evaluate INSTRUCTSCORE on a variety of generation tasks, including translation, captioning, data-to-text, and commonsense generation. Experiments show that our 7B model surpasses all other unsupervised metrics, including those based on 175B GPT-3 and GPT-4. Surprisingly, our INSTRUCTSCORE, even without direct supervision from human-rated data, achieves performance levels on par with state-of-the-art metrics like COMET22, which were fine-tuned on human ratings.

12:00-12:15 (West 1 Ballroom)
**Pre-training Language Models for Comparative Reasoning**
*Mengxia Yu, Zhihan Zhang, Wenhao Yu and Meng Jiang*
Comparative reasoning is a process of comparing objects, concepts, or entities to draw conclusions, which constitutes a fundamental cognitive ability. In this paper, we propose a novel framework to pre-train language models for enhancing their abilities of comparative reasoning over texts. While there have been approaches for NLP tasks that require comparative reasoning, they suffer from costly manual data labeling and limited generalizability to different tasks. Our approach introduces a novel method of collecting scalable data for text-based entity comparison, which leverages both structured and unstructured data. Moreover, we present a framework of pre-training language models via three novel objectives on comparative reasoning. Evaluation on downstream tasks including comparative question answering, question generation, and summarization shows that our pre-training framework significantly improves the comparative reasoning abilities of language models, especially under low-resource conditions. This work also releases the first integrated benchmark for comparative reasoning.

12:15-12:30 (West 1 Ballroom)

**Composable Text Controls in Latent Space with ODEs**
*Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li and Zhiting Hu*
Real-world text applications often involve composing a wide range of text control operations, such as editing the text w.r.t. an attribute, manipulating keywords and structure, and generating new text of desired properties. Prior work typically learns/finetunes a language model (LM) to perform individual or specific subsets of operations. Recent research has studied combining operations in a plug-and-play manner, often with costly search or optimization in the complex sequence space. This paper proposes a new efficient approach for composable text operations in the compact latent space of text. The low-dimensionality and differentiability of the text latent vector allow us to develop an efficient sampler based on ordinary differential equations (ODEs) given arbitrary plug-in operators (e.g., attribute classifiers). By connecting pretrained LMs (e.g., GPT2) to the latent space through efficient adaption, we then decode the sampled vectors into desired text sequences. The flexible approach permits diverse control operators (sentiment, tense, formality, keywords, etc.) acquired using any relevant data from different domains. Experiments show that composing those operators within our approach manages to generate or edit high-quality text, substantially improving over previous methods in terms of generation quality and efficiency.

## Question Answering

11:00-12:30 (West 2 Ballroom)

---

11:00-11:15 (West 2 Ballroom)
**Compressing and Debiasing Vision-Language Pre-Trained Models for Visual Question Answering**
*Qingyi Si, Yuanxin Liu, Zheng Lin, Peng Fu, Yanan Cao and Weiping Wang*
Despite the excellent performance of vision-language pre-trained models (VLPs) on conventional VQA task, they still suffer from two problems: First, VLPs tend to rely on language biases in datasets and fail to generalize to out-of-distribution (OOD) data. Second, they are inefficient in terms of memory footprint and computation. Although promising progress has been made in both problems, most existing works tackle them independently. To facilitate the application of VLP to VQA tasks, it is imperative to jointly study VLP compression and OOD robustness, which, however, has not yet been explored. This paper investigates whether a VLP can be compressed and debiased simultaneously by searching sparse and robust subnetworks. To this end, we systematically study the design of a training and compression pipeline to search the subnetworks, as well as the assignment of sparsity to different modality-specific modules. Our experiments involve 2 VLPs, 2 compression methods, 4 training methods, 2 datasets and a range of sparsity levels. Our results show that there indeed exist sparse and robust subnetworks, which are competitive with the debiased full VLP and clearly outperform the debiasing SoTAs with fewer parameters on OOD datasets VQA-CP v2 and VQA-VS. The codes can be found at https://github.com/PhoebusSi/Compress-Robust-VQA.

11:15-11:30 (West 2 Ballroom)
**Merging Generated and Retrieved Knowledge for Open-Domain QA**
*Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee and Lu Wang*
Open-domain question answering (QA) systems are often built with retrieval modules. However, retrieving passages from a given source is known to suffer from insufficient knowledge coverage. Alternatively, prompting large language models (LLMs) to generate contextual passages based on their parametric knowledge has been shown to improve QA performance. Yet, LLMs tend to "hallucinate" content that conflicts with the retrieved knowledge. Based on the intuition that answers supported by both sources are more likely to be correct, we propose COMBO, a Compatibility-Oriented knowledge Merging for Better Open-domain QA framework, to effectively leverage the two sources of information. Concretely, we match LLM-generated passages with retrieved counterparts into compatible pairs, based on discriminators trained with silver compatibility labels. Then a Fusion-in-Decoder-based reader model handles passage pairs to arrive at the final answer. Experiments show that COMBO outperforms competitive baselines on three out of four tested open-domain QA benchmarks. Further analysis reveals that our proposed framework demonstrates greater efficacy in scenarios with a higher degree of knowledge conflicts.

11:30-11:45 (West 2 Ballroom)
**Diversity Enhanced Narrative Question Generation for Storybooks**
*Hokeun Yoon and JinYeong Bak*
Question generation (QG) from a given context can enhance comprehension, engagement, assessment, and overall efficacy in learning or conversational environments. Despite recent advancements in QG, the challenge of enhancing or measuring the diversity of generated questions often remains unaddressed. In this paper, we introduce a multi-question generation model (mQG), which is capable of generating multiple, diverse, and answerable questions by focusing on context and questions. To validate the answerability of the generated questions, we employ a SQuAD 2.0 fine-tuned question answering model, classifying the questions as answerable or not. We train and evaluate mQG on the FairytaleQA dataset, a well-structured QA dataset based on storybooks, with narrative questions. We further apply a zero-shot adaptation on the TellMeWhy and SQuAD1.1 datasets. mQG shows promising results across various evaluation metrics, among strong baselines.

11:45-12:00 (West 2 Ballroom)
**The Art of SOCRATIC QUESTIONING: Recursive Thinking with Large Language Models**
*Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang and Lifu Huang*
Chain-of-Thought (CoT) prompting enables large language models to solve complex reasoning problems by generating intermediate steps. However, confined by its inherent single-pass and sequential generation process, CoT heavily relies on the initial decisions, causing errors in early steps to accumulate and impact the final answers. In contrast, humans adopt recursive thinking when tackling complex reasoning problems, i.e. iteratively breaking the original problem into approachable sub-problems and aggregating their answers to resolve the original one. Inspired by the human cognitive process, we propose SOCRATIC QUESTIONING, a divide-and-conquer style algorithm that mimics the recursive thinking process. Specifically, SOCRATIC QUESTIONING leverages large language models to raise and answer sub-questions until collecting enough information to tackle the original question. Unlike CoT, SOCRATIC QUESTIONING explicitly navigates the thinking space, stimulates effective recursive thinking, and is more robust towards errors in the thinking process. Extensive experiments on several complex reasoning tasks, including MMLU, MATH, LogiQA, and visual question-answering demonstrate significant performance improvements over the state-of-the-art prompting methods, such as CoT, and Tree-of-Thought. The qualitative analysis clearly shows that the intermediate reasoning steps elicited by SOCRATIC QUESTIONING are similar to humans' recursively thinking process of complex reasoning problems.

12:00-12:15 (West 2 Ballroom)
**Once Upon a $Time$ in $Graph$: Relative-Time Pretraining for Complex Temporal Reasoning**
*Sen Yang, Xin Li, Lidong Bing and Wai Lam*
Our physical world is constantly evolving over time, rendering challenges for pre-trained language models to understand and reason over the temporal contexts of texts. Existing work focuses on strengthening the direct association between a piece of text and its time-stamp. However, the knowledge-time association is usually insufficient for the downstream tasks that require reasoning over temporal dependencies

between knowledge. In this work, we make use of the underlying nature of time, all temporally-scoped sentences are strung together through a one-dimensional time axis, and suggest creating a graph structure based on the relative placements of events along the time axis. Inspired by the graph view, we propose REMEMO ($\underline{R}elative\underline{T}ime\underline{M}odeling$), which explicitly connects all temporally-scoped facts by modeling the time relations between any two sentences. Experimental results show that REMEMO outperforms the baseline T5 on multiple temporal question answering datasets under various settings. Further analysis suggests that REMEMO is especially good at modeling long-range complex temporal dependencies.

12:15-12:30 (West 2 Ballroom)
**On the Robustness of Dialogue History Representation in Conversational Question Answering: A Comprehensive Study and a New Prompt-based Method**
*Roi Reichart, Zorik Gekhman, Nadav Oved, Orgad Keller and Idan Szpektor*
Most works on modeling the conversation history in Conversational Question Answering (CQA) report a single main result on a common CQA benchmark. While existing models show impressive results on CQA leaderboards, it remains unclear whether they are robust to shifts in setting (sometimes to more realistic ones), training data size (e.g. from large to small sets) and domain.In this work, we design and conduct the first large-scale robustness study of history modeling approaches for CQA. We find that high benchmark scores do not necessarily translate to strong robustness, and that various methods can perform extremely differently under different settings. Equipped with the insights from our study, we design a novel prompt-based history modeling approach and demonstrate its strong robustness across various settings.Our approach is inspired by existing methods that highlight historic answers in the passage. However, instead of highlighting by modifying the passage token embeddings, we add textual prompts directly in the passage text. Our approach is simple, easy-to-plug into practically any model, and highly effective, thus we recommend it as a starting point for future model developers. We also hope that our study and insights will raise awareness to the importance of robustness-focused evaluation, in addition to obtaining high leaderboard scores, leading to better CQA systems.

## Resources and Evaluation 1

11:00-12:30 (West 3 Ballroom)

---

11:00-11:15 (West 3 Ballroom)
**HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models**
*Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie and Ji-Rong Wen*
Large language models (LLMs), such as ChatGPT, are prone to generate hallucinations, i.e., content that conflicts with the source or cannot be verified by the factual knowledge. To understand what types of content and to which extent LLMs are apt to hallucinate, we introduce the Hallucination Evaluation for Large Language Models (HaluEval) benchmark, a large collection of generated and human-annotated hallucinated samples for evaluating the performance of LLMs in recognizing hallucination. To generate these samples, we propose a ChatGPT-based two-step framework, i.e., sampling-then-filtering. Besides, we also hire some human labelers to annotate the hallucinations in ChatGPT responses. The empirical results suggest that ChatGPT is likely to generate hallucinated content in specific topics by fabricating unverifiable information (i.e., about $19.5\%$ user queries). Moreover, existing LLMs face great challenges in recognizing the hallucinations in texts. While, our experiments also prove that the hallucination recognition can be improved by providing external knowledge or adding reasoning steps.

11:15-11:30 (West 3 Ballroom)
**TRIGO: Benchmarking Formal Mathematical Proof Reduction for Generative Language Models**
*Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, Chuanyang Zheng, Xiaodan Liang, Ming Zhang and Qun Liu*
Automated theorem proving (ATP) has become an appealing domain for exploring the reasoning ability of the recent successful generative language models. However, current ATP benchmarks are mainly focus on symbolic inference, but rarely involve the understanding of complex number combination reasoning. In this work, we propose TRIGO, an ATP benchmark that not only requires a model to reduce a trigonometric expression with step-by-step proof but also evaluates a generative LM's reasoning ability on formulas and capability to manipulate, group, and factor number terms. We gather trigonometric expressions and their reduced forms from web, annotate the simplification process manually, and translate it into the "Lean" formal language system. We then automatically generate additional examples from the annotated samples to expand the dataset. Furthermore, we also create three automatically generated training and testing datasets of varying difficulty and distributions. Our extensive experiments show our proposed TRIGO poses a new challenge for advanced generative LM's including GPT-4 which is pre-trained on a considerable amount of open-source formal theorem-proving language data, and provide a new tool to study the generative LM's ability on both formal and mathematical reasoning.

11:30-11:45 (West 3 Ballroom)
**BanglaAbuseMeme: A Dataset for Bengali Abusive Meme Classification**
*Mithun Das and Animesh Mukherjee*
The dramatic increase in the use of social media platforms for information sharing has also fueled a steep growth in online abuse. A simple yet effective way of abusing individuals or communities is by creating memes, which often integrate an image with a short piece of text layered on top of it. Such harmful elements are in rampant use and are a threat to online safety. Hence it is necessary to develop efficient models to detect and flag abusive memes. The problem becomes more challenging in a low-resource setting (e.g., Bengali memes, i.e., images with Bengali text embedded on it) because of the absence of benchmark datasets on which AI models could be trained. In this paper we bridge this gap by building a Bengali meme dataset. To setup an effective benchmark we implement several baseline models for classifying abusive memes using this dataset. We observe that multimodal models that use both textual and visual information outperform unimodal models. Our best-performing model achieves a macro F1 score of 70.51. Finally, we perform a qualitative error analysis of the misclassified memes of the best-performing text-based, image-based and multimodal models.

11:45-12:00 (West 3 Ballroom)
**IDTraffickers: An Authorship Attribution Dataset to link and connect Potential Human-Trafficking Operations on Text Escort Advertisements**
*Vageesh Kumar Saxena, Benjamin Ashpole, Gijs van Dijck and Gerasimos Spanakis*
Human trafficking (HT) is a pervasive global issue affecting vulnerable individuals, violating their fundamental human rights. Investigations reveal that a significant number of HT cases are associated with online advertisements (ads), particularly in escort markets. Consequently, identifying and connecting HT vendors has become increasingly challenging for Law Enforcement Agencies (LEAs). To address this issue, we introduce IDTraffickers, an extensive dataset consisting of 87,595 text ads and 5,244 vendor labels to enable the verification and identification of potential HT vendors on online escort markets. To establish a benchmark for authorship identification, we train a DeCLUTR-small model, achieving a macro-F1 score of 0.8656 in a closed-set classification environment. Next, we leverage the style representations

extracted from the trained classifier to conduct authorship verification, resulting in a mean r-precision score of 0.8852 in an open-set ranking environment. Finally, to encourage further research and ensure responsible data sharing, we plan to release IDTraffickers for the authorship attribution task to researchers under specific conditions, considering the sensitive nature of the data. We believe that the availability of our dataset and benchmarks will empower future researchers to utilize our findings, thereby facilitating the effective linkage of escort ads and the development of more robust approaches for identifying HT indicators.

12:00-12:15 (West 3 Ballroom)
**This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models**
*Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios and German Rigau*
Although large language models (LLMs) have apparently acquired a certain level of grammatical knowledge and the ability to make generalizations, they fail to interpret negation, a crucial step in Natural Language Processing. We try to clarify the reasons for the sub-optimal performance of LLMs understanding negation. We introduce a large semi-automatically generated dataset of circa 400,000 descriptive sentences about commonsense knowledge that can be true or false in which negation is present in about 2/3 of the corpus in different forms. We have used our dataset with the largest available open LLMs in a zero-shot approach to grasp their generalization and inference capability and we have also fine-tuned some of the models to assess whether the understanding of negation can be trained. Our findings show that, while LLMs are proficient at classifying affirmative sentences, they struggle with negative sentences and lack a deep understanding of negation, often relying on superficial cues. Although fine-tuning the models on negative sentences improves their performance, the lack of generalization in handling negation is persistent, highlighting the ongoing challenges of LLMs regarding negation understanding and generalization. The dataset and code are publicly available.

12:15-12:30 (West 3 Ballroom)
**You Told Me That Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models**
*Alexander Baranov, Vladimir Kniazhevsky and Pavel Braslavski*
In this study, we focus on automatic humor detection, a highly relevant task for conversational AI. To date, there are several English datasets for this task, but little research on how models trained on them generalize and behave in the wild. To fill this gap, we carefully analyze existing datasets, train RoBERTa-based and Naïve Bayes classifiers on each of them, and test on the rest. Training and testing on the same dataset yields good results, but the transferability of the models varies widely. Models trained on datasets with jokes from different sources show better transferability, while the amount of training data has a smaller impact. The behavior of the models on out-of-domain data is unstable, suggesting that some of the models overfit, while others learn non-specific humor characteristics. An adversarial attack shows that models trained on pun datasets are less robust. We also evaluate the sense of humor of the chatGPT and Flan-UL2 models in a zero-shot scenario. The LLMs demonstrate competitive results on humor datasets and a more stable behavior on out-of-domain data. We believe that the obtained results will facilitate the development of new datasets and evaluation methodologies in the field of computational humor. We've made all the data from the study and the trained models publicly available at https://github.com/Humor-Research/Humor-detection.

## Demo session 5

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**TP-Detector: Detecting Turning Points in the Engineering Process of Large-scale Projects**
*Qi Wu, WenHan Chao, Xian Zhou and Zhunchen Luo*
This paper introduces a novel task of detecting turning points in the engineering process of large-scale projects, wherein the turning points signify significant transitions occurring between phases. Given the complexities involving diverse critical events and limited comprehension in individual news reports, we approach the problem by treating the sequence of related news streams as a window with multiple instances. To capture the evolution of changes effectively, we adopt a deep Multiple Instance Learning (MIL) framework and employ the multiple instance ranking loss to discern the transition patterns exhibited in the turning point window. Extensive experiments consistently demonstrate the effectiveness of our proposed approach on the constructed dataset compared to baseline methods. We deployed the proposed mode and provided a demonstration video to illustrate its functionality. The code and dataset are available on GitHub.

11:00-12:30 (East Foyer / Virtual)
**MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models**
*Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang and Jiang Bian*
AI-empowered music processing is a diverse feld that encompasses dozens of tasks, ranging from generation tasks (e.g., timbre synthesis) to comprehension tasks (e.g., music classifcation). For developers and amateurs, it is very diffcult to grasp all of these task to satisfy their requirements in music processing, especially considering the huge differences in the representations of music data and the model applicability across platforms among various tasks. Consequently, it is necessary to build a system to organize and integrate these tasks, and thus help practitioners to automatically analyze their demand and call suitable tools as solutions to fulfill their requirements. Inspired by the recent success of large language models (LLMs) in task automation, we develop a system, named MusicAgent, which integrates numerous music-related tools and an autonomous workflow to address user requirements. More specifically, we build 1) toolset that collects tools from diverse sources, including Hugging Face, GitHub, and Web API, etc. 2) an autonomous workflow empowered by LLMs (e.g., ChatGPT) to organize these tools and automatically decompose user requests into multiple sub-tasks and invoke corresponding music tools. The primary goal of this system is to free users from the intricacies of AI-music tools, enabling them to concentrate on the creative aspect. By granting users the freedom to effortlessly combine tools, the system offers a seamless and enriching music experience. The code is available on GitHub along with a brief instructional video.

11:00-12:30 (East Foyer / Virtual)
**CHAMP: Efficient Annotation and Consolidation of Cluster Hierarchies**
*Arie Cattan, Tom Hope, Doug Downey, Roy Bar-Haim, Lilach Eden, Yoav Kantor and Ido Dagan*
Various NLP tasks require a complex hierarchical structure over nodes, where each node is a cluster of items. Examples include generating entailment graphs, hierarchical cross-document coreference resolution, annotating event and subevent relations, etc. To enable efficient annotation of such hierarchical structures, we release CHAMP, an open source tool allowing to incrementally construct both clusters and hierarchy simultaneously over any type of texts. This incremental approach significantly reduces annotation time compared to the common pairwise annotation approach and also guarantees maintaining transitivity at the cluster and hierarchy levels. Furthermore, CHAMP includes a consolidation mode, where an adjudicator can easily compare multiple cluster hierarchy annotations and resolve disagreements.

11:00-12:30 (East Foyer / Virtual)
**Descriptive Knowledge Graph in Biomedical Domain**

*Kerui Zhu, Jie Huang and Kevin Chen-Chuan Chang*
We present a novel system that automatically extracts and generates informative and descriptive sentences from the biomedical corpus and facilitates the efficient search for relational knowledge. Unlike previous search engines or exploration systems that retrieve unconnected passages, our system organizes descriptive sentences as a relational graph, enabling researchers to explore closely related biomedical entities (e.g., diseases treated by a chemical) or indirectly connected entities (e.g., potential drugs for treating a disease). Our system also uses ChatGPT and a fine-tuned relation synthesis model to generate concise and reliable descriptive sentences from retrieved information, reducing the need for extensive human reading effort. With our system, researchers can easily obtain both high-level knowledge and detailed references and interactively steer to the information of interest. We spotlight the application of our system in COVID-19 research, illustrating its utility in areas such as drug repurposing and literature curation.

11:00-12:30 (East Foyer / Virtual)
**ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models**
*Baoli Zhang, Haining Xie, Pengfan Du, Junhao Chen, Pengfei Cao, Yubo Chen, Shengping Liu, Kang Liu and Jun Zhao*
The unprecedented performance of LLMs requires comprehensive and accurate evaluation. We argue that for LLMs evaluation, benchmarks need to be comprehensive and systematic. To this end, we propose the Zhujiu benchmark, which has the following strengths: (1) Multi-dimensional ability coverage: We comprehensively evaluate LLMs across 7 ability dimensions covering 51 tasks. Especially, we also propose a new benchmark that focus on knowledge ability of LLMs. (2) Multi-faceted evaluation methods collaboration: We use 3 different yet complementary evaluation methods to comprehensively evaluate LLMs, which can ensure the authority and accuracy of the evaluation results. (3) Comprehensive Chinese benchmark: ZhuJiu is the pioneering benchmark that fully assesses LLMs in Chinese, while also providing equally robust evaluation abilities in English. (4) Avoiding potential data leakage: To avoid data leakage, we construct evaluation data specifically for 37 tasks. We evaluate 10 current mainstream LLMs, and conduct an in-depth discussion and analysis of their results. The ZhuJiu benchmark and open-participation leaderboard are publicly released at http://www.zhujiu-benchmark.com and we also provide a demo video at https://youtu.be/qypkJ89L1Ic.

11:00-12:30 (East Foyer / Virtual)
**CoLLiE: Collaborative Training of Large Language Models in an Efficient Way**
*Kai Lv, Shuo Zhang, Tianle Gu, Shuhao Xing, Jiawei Hong, Keyu Chen, Xiaoran Liu, Yuqing Yang, Honglin Guo, Tengxiao Liu, Yu Sun, Qipeng Guo, Hang Yan and Xipeng Qiu*
Large language models (LLMs) are increasingly pivotal in a wide range of natural language processing tasks. Access to pre-trained models, courtesy of the open-source community, has made it possible to adapt these models to specific applications for enhanced performance. However, the substantial resources required for training these models necessitate efficient solutions. This paper introduces CoLLiE, an efficient library that facilitates collaborative training of large language models using 3D parallelism, parameter-efficient fine-tuning (PEFT) methods, and optimizers such as Lion, Adan, Sophia, and LOMO. With its modular design and comprehensive functionality, CoLLiE offers a balanced blend of efficiency, ease of use, and customization. CoLLiE has proven superior training efficiency in comparison with prevalent solutions in pre-training and fine-tuning scenarios. Furthermore, we provide an empirical evaluation of the correlation between model size and GPU memory consumption under different optimization methods, as well as an analysis of the throughput. Lastly, we carry out a comprehensive comparison of various optimizers and PEFT methods within the instruction-tuning context. CoLLiE is available at https://github.com/OpenLMLab/collie.

11:00-12:30 (East Foyer / Virtual)
**Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding**
*Hang Zhang, Xin Li and Lidong Bing*
We present Video-LLaMA, a multi-modal framework that empowers Large Language Models (LLMs) with the capability of understanding both visual and auditory content in the video. Video-LLaMA bootstraps cross-modal training from the frozen pre-trained visual & audio encoders and the frozen LLMs. Unlike previous works that complement LLMs to process the visual or audio signals only, Video-LLaMA enables video comprehension by tackling two challenges: (1) capturing the temporal changes in visual scenes, (2) integrating audio-visual signals. To counter the first challenge, we propose a Video Q-former to assemble a pre-trained image encoder into our video encoder and introduce a video-to-text generation task to learn video-language correspondence. For the second challenge, we leverage ImageBind, a universal embedding model aligning multiple modalities, as the pre-trained audio encoder and introduce an Audio Q-former on top of ImageBind to learn reasonable auditory query embeddings for the LLM module. To align the output of both visual & audio encoders with LLM's embedding space, we first train Video-LLaMA on massive video/image-caption pairs and then tune our model with visual-instruction datasets of moderate amount but higher quality. We found Video-LLaMA shows the ability to perceive and comprehend video content and generate meaningful responses grounded in the visual and auditory information presented in the videos.

11:00-12:30 (East Foyer / Virtual)
**SummHelper: Collaborative Human-Computer Summarization**
*Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira and Ido Dagan*
Current approaches for text summarization are predominantly automatic, with rather limited space for human intervention and control over the process. In this paper, we introduce SummHelper, and screencast demo at https://www.youtube.com/watch?v=nGcknJwGhxk a 2-phase summarization assistant designed to foster human-machine collaboration. The initial phase involves content selection, where the system recommends potential content, allowing users to accept, modify, or introduce additional selections. The subsequent phase, content consolidation, involves SummHelper generating a coherent summary from these selections, which users can then refine using visual mappings between the summary and the source text. Small-scale user studies reveal the effectiveness of our application, with participants being especially appreciative of the balance between automated guidance and opportunities for personal input.

11:00-12:30 (East Foyer / Virtual)
**ModelScope-Agent: Building Your Customizable Agent System with Open-source Large Language Models**
*Chenliang Li, He Chen, Ming Yan, Weizhou Shen, Haiyang Xu, Zhikai Wu, Zhicheng Zhang, Wenmeng Zhou, Yingda Chen, Chen Cheng, Hongzhu Shi, Ji Zhang, Fei Huang and Jingren Zhou*
Large language models (LLMs) have recently demonstrated remarkable capabilities to comprehend human intentions, engage in reasoning, and design planning-like behavior. To further unleash the power of LLMs to accomplish complex tasks, there is a growing trend to build agent frameworks that equips LLMs, such as ChatGPT, with tool-use abilities to connect with massive external APIs. In this work, we introduce ModelScope-Agent, a general and customizable agent framework for real-world applications, based on open-source LLMs as controllers. It provides a user-friendly system library, with a customizable engine design to support model training on multiple open-source LLMs, while also enabling seamless integration with both model APIs and common APIs in a unified way. To equip the LLMs with tool-use abilities, a comprehensive framework has been proposed spanning tool-use data collection, tool retrieval, tool registration, memory control, customized model training, and evaluation for practical real-world applications. Finally, we showcase ModelScopeGPT, a real-world intelligent assistant of ModelScope Community based on the ModelScope-Agent framework, which is able to connect open-source LLMs with more than 1000 public AI models and localized community knowledge in ModelScope. The ModelScope-Agent online demo, library are now publicly avail-

able.

11:00-12:30 (East Foyer / Virtual)
**EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge**
*Tom Bryan, Jacob Carlson, Abhishek Arora and Melissa Dell*
Billions of public domain documents remain trapped in hard copy or lack an accurate digitization. Modern natural language processing methods cannot be used to index, retrieve, and summarize their texts; conduct computational textual analyses; or extract information for statistical analyses, and these texts cannot be incorporated into language model training. Given the diversity and sheer quantity of public domain texts, liberating them at scale requires optical character recognition (OCR) that is accurate, extremely cheap to deploy, and sample-efficient to customize to novel collections, languages, and character sets. Existing OCR engines, largely designed for small-scale commercial applications in high resource languages, often fall short of these requirements. EffOCR (EfficientOCR), a novel open-source OCR package, meets both the computational and sample efficiency requirements for liberating texts at scale by abandoning the sequence-to-sequence architecture typically used for OCR, which takes representations from a learned vision model as inputs to a learned language model. Instead, EffOCR models OCR as a character or word-level image retrieval problem. EffOCR is cheap and sample efficient to train, as the model only needs to learn characters' visual appearance and not how they are used in sequence to form language. Models in the EffOCR model zoo can be deployed off-the-shelf with only a few lines of code and include lightweight models designed for mobile phones that are extremely cheap to deploy. Importantly, EffOCR also allows for easy, sample efficient customization with a simple model training interface and minimal labeling requirements due to its sample efficiency. We illustrate the utility of EffOCR by cheaply and accurately digitizing 20 million historical U.S. newspaper scans, evaluating zero-shot performance on randomly selected documents from the U.S. National Archives, and accurately digitizing a Japanese document collection for which all other OCR solutions failed.

## Poster session 5

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**#1 Mirages. On Anthropomorphism in Dialogue Systems**
*Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser and Zeerak Talat*
Automated dialogue or conversational systems are anthropomorphised by developers and personified by users. While a degree of anthropomorphism is inevitable, conscious and unconscious design choices can guide users to personify them to varying degrees. Encouraging users to relate to automated systems as if they were human can lead to transparency and trust issues, and high risk scenarios caused by over-reliance on their outputs. As a result, natural language processing researchers have investigated the factors that induce personification and develop resources to mitigate such effects. However, these efforts are fragmented, and many aspects of anthropomorphism have yet to be explored. In this paper, we discuss the linguistic factors that contribute to the anthropomorphism of dialogue systems and the harms that can arise thereof, including reinforcing gender stereotypes and conceptions of acceptable language. We recommend that future efforts towards developing dialogue systems take particular care in their design, development, release, and description; and attend to the many linguistic cues that can elicit personification by users.

11:00-12:30 (East Foyer / Virtual)
**#2 DecipherPref: Analyzing Influential Factors in Human Preference Judgments via GPT-4**
*Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh and Fei Liu*
Human preference judgments are pivotal in guiding large language models (LLMs) to produce outputs that align with human values. Human evaluations are also used in summarization tasks to compare outputs from various systems, complementing existing automatic metrics. Despite their significance, however, there has been limited research probing these pairwise or $k$-wise comparisons. The collective impact and relative importance of factors such as output length, informativeness, fluency, and factual consistency are still not well understood. It is also unclear if there are other hidden factors influencing human judgments. In this paper, we conduct an in-depth examination of a collection of pairwise human judgments released by OpenAI. Utilizing the Bradley-Terry-Luce (BTL) model, we reveal the inherent preferences embedded in these human judgments. We find that the most favored factors vary across tasks and genres, whereas the least favored factors tend to be consistent, e.g., outputs are too brief, contain excessive off-focus content or hallucinated facts. Our findings have implications on the construction of balanced datasets in human preference evaluations, which is a crucial step in shaping the behaviors of future LLMs.

11:00-12:30 (East Foyer / Virtual)
**#3 Adaptive Gating in Mixture-of-Experts based Language Models**
*Jiamin Li, Qiang Su, Yitao Yang, Yimin Jiang, Cong Wang and Hong Xu*
Large language models have demonstrated exceptional language understanding capabilities in many NLP tasks. Sparsely activated mixture-of-experts (MoE) has emerged as a promising solution for scaling models while maintaining a constant number of computational operations. Existing MoE models adopt a fixed gating network where each token is computed by the same number of experts. This contradicts our intuition that the tokens in each sequence vary in terms of their linguistic complexity and, consequently, require different computational costs. Little is discussed in prior research on the trade-off between computation per token and model performance. This paper introduces adaptive gating in MoE, a flexible training strategy that allows tokens to be processed by a variable number of experts based on expert probability distribution. Adaptive gating preserves sparsity while improving training efficiency. We further draw upon curriculum learning to better align the order of training samples and maximize the training time savings. Extensive experiments on diverse NLP tasks show that adaptive gating reduces at most 22.5% training time while maintaining inference quality. Moreover, we conduct a comprehensive analysis of the gating decisions and present our insights on which tokens are inherently difficult to process, depending on the specific language task.

11:00-12:30 (East Foyer / Virtual)
**#4 CoLT5: Faster Long-Range Transformers with Conditional Computation**
*Joshua Ainslie, Tao Lei, Michiel de Jong, Santiago Ontanon, Siddhartha Brahma, Yury Zemlyanskiy, David Uthus, Mandy Guo, James Lee-Thorp, Yi Tay, Yun-Hsuan Sung and Sumit Sanghai*
Many natural language processing tasks benefit from long inputs, but processing long documents with Transformers is expensive – not only due to quadratic attention complexity but also from applying feedforward and projection layers to every token. However, not all tokens are equally important, especially for longer documents. We propose CoLT5, a long-input Transformer model that builds on this intuition by employing conditional computation, devoting more resources to important tokens in both feedforward and attention layers. We show that CoLT5 achieves stronger performance than LongT5 with much faster training and inference, achieving SOTA on the long-input SCROLLS benchmark. Moreover, CoLT5 can effectively and tractably make use of extremely long inputs, showing strong gains up to 64k input length.

11:00-12:30 (East Foyer / Virtual)

**#5 GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints**
*Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron and Sumit Sanghai*
Multi-query attention (MQA), which only uses a single key-value head, drastically speeds up decoder inference. However, MQA can lead to quality degradation, and moreover it may not be desirable to train a separate model just for faster inference. We (1) propose a recipe for uptraining existing multi-head language model checkpoints into models with MQA using 5% of original pre-training compute, and (2) introduce grouped-query attention (GQA), a generalization of multi-query attention which uses an intermediate (more than one, less than number of query heads) number of key-value heads. We show that uptrained GQA achieves quality close to multi-head attention with comparable speed to MQA.

11:00-12:30 (East Foyer / Virtual)
**#6 Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication**
*Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang and Xipeng Qiu*
Large Language Models (LLMs) have recently made significant strides in complex reasoning tasks through the Chain-of-Thought technique. Despite this progress, their reasoning is often constrained by their intrinsic understanding, lacking external insights. To address this, we propose Exchange-of-Thought (EoT), a novel framework that enables cross-model communication during problem-solving. Drawing inspiration from network topology, EoT integrates four unique communication paradigms: Memory, Report, Relay, and Debate. This paper delves into the communication dynamics and volume associated with each paradigm. To counterbalance the risks of incorrect reasoning chains, we implement a robust confidence evaluation mechanism within these communications. Our experiments across diverse complex reasoning tasks demonstrate that EoT significantly surpasses established baselines, underscoring the value of external insights in enhancing LLM performance. Furthermore, we show that EoT achieves these superior results in a cost-effective manner, marking a promising advancement for efficient and collaborative AI problem-solving.

11:00-12:30 (East Foyer / Virtual)
**#7 GEMINI: Controlling The Sentence-Level Summary Style in Abstractive Text Summarization**
*Guangsheng Bao, Zebin Ou and Yue Zhang*
Human experts write summaries using different techniques, including extracting a sentence from the document and rewriting it, or fusing various information from the document to abstract it. These techniques are flexible and thus difficult to be imitated by any single method. To address this issue, we propose an adaptive model, GEMINI, that integrates a rewriter and a generator to mimic the sentence rewriting and abstracting techniques, respectively. GEMINI adaptively chooses to rewrite a specific document sentence or generate a summary sentence from scratch. Experiments demonstrate that our adaptive approach outperforms the pure abstractive and rewriting baselines on three benchmark datasets, achieving the best results on WikiHow. Interestingly, empirical results show that the human summary styles of summary sentences are consistently predictable given their context. We release our code and model at https://github.com/baoguangsheng/gemini.

11:00-12:30 (East Foyer / Virtual)
**#8 Boosting Summarization with Normalizing Flows and Aggressive Training**
*Yu Yang and Xiaotong Shen*
This paper presents FlowSUM, a normalizing flows-based variational encoder-decoder framework for Transformer-based summarization. Our approach tackles two primary challenges in variational summarization: insufficient semantic information in latent representations and posterior collapse during training. To address these challenges, we employ normalizing flows to enable flexible latent posterior modeling, and we propose a controlled alternate aggressive training (CAAT) strategy with an improved gate mechanism. Experimental results show that FlowSUM significantly enhances the quality of generated summaries and unleashes the potential for knowledge distillation with minimal impact on inference time. Furthermore, we investigate the issue of posterior collapse in normalizing flows and analyze how the summary quality is affected by the training strategy, gate initialization, and the type and number of normalizing flows used, offering valuable insights for future research.

11:00-12:30 (East Foyer / Virtual)
**#9 Exploring the Impact of Model Scaling on Parameter-Efficient Tuning**
*Yusheng Su, Chi-Min Chan, Jiali Cheng, Yujia Qin, Yankai Lin, Shengding Hu, Zonghan Yang, Ning Ding, Xingzhi Sun, Guotong Xie, Zhiyuan Liu and Maosong Sun*
Parameter-efficient tuning (PET) methods can effectively drive extremely large pre-trained language models (PLMs) by training only minimal parameters. Different PET methods utilize different manually designed tunable modules. In small PLMs, there are usually noticeable performance differences among PET methods. Nevertheless, as the model scale increases, the performance differences become marginal. Hence, we hypothesize that model scaling mitigates the impact of design differences on PET methods. To investigate this hypothesis, we introduce a more flexible PET method called Arbitrary PET (APET) method. The APET method is compatible with a tunable module, which consists of any number of parameters distributed in arbitrary positions. Then, we utilize it and conduct experiments on 11 NLP tasks across 3 representative PLMs. Our investigations reveal that model scaling (1) mitigates the effects of the positions of tunable parameters on performance, and (2) enables tuning methods to achieve performance comparable to full-parameter fine-tuning by optimizing fewer tunable parameters. Intriguingly, we also observe that tuning methods optimize the similar number of tunable parameters to exceed random guess performance on different tasks. We collectively discuss this phenomenon and the two aforementioned findings from an optimization perspective to understand the underlying mechanisms. These conclusions enhance our understanding of the impact of model scaling on PET and assist in designing more effective and efficient PET methods for PLMs of different scales. The source code can be obtained from this GitHub repository: https://github.com/yushengsu-thu/PET_Scaling.

11:00-12:30 (East Foyer / Virtual)
**#10 MeaeQ: Mount Model Extraction Attacks with Efficient Queries**
*Chengwei Dai, Minxuan Lv, Kun Li and Wei Zhou*
We study model extraction attacks in natural language processing (NLP) where attackers aim to steal victim models by repeatedly querying the open Application Programming Interfaces (APIs). Recent works focus on limited-query budget settings and adopt random sampling or active learning-based sampling strategies on publicly available, unannotated data sources. However, these methods often result in selected queries that lack task relevance and data diversity, leading to limited success in achieving satisfactory results with low query costs. In this paper, we propose MeaeQ (Model extraction attack with efficient Queries), a straightforward yet effective method to address these issues. Specifically, we initially utilize a zero-shot sequence inference classifier, combined with API service information, to filter task-relevant data from a public text corpus instead of a problem domain-specific dataset. Furthermore, we employ a clustering-based data reduction technique to obtain representative data as queries for the attack. Extensive experiments conducted on four benchmark datasets demonstrate that MeaeQ achieves higher functional similarity to the victim model than baselines while requiring fewer queries.

11:00-12:30 (East Foyer / Virtual)
**#11 CT-GAT: Cross-Task Generative Adversarial Attack based on Transferability**
*Minxuan Lv, Chengwei Dai, Kun Li, Wei Zhou and Songlin Hu*

Neural network models are vulnerable to adversarial examples, and adversarial transferability further increases the risk of adversarial attacks. Current methods based on transferability often rely on substitute models, which can be impractical and costly in real-world scenarios due to the unavailability of training data and the victim model's structural details. In this paper, we propose a novel approach that directly constructs adversarial examples by extracting transferable features across various tasks. Our key insight is that adversarial transferability can extend across different tasks. Specifically, we train a sequence-to-sequence generative model named CT-GAT (Cross-Task Generative Adversarial Attack) using adversarial sample data collected from multiple tasks to acquire universal adversarial features and generate adversarial examples for different tasks.We conduct experiments on ten distinct datasets, and the results demonstrate that our method achieves superior attack performance with small cost.

11:00-12:30 (East Foyer / Virtual)
**#12 ATFormer: A Learned Performance Model with Transfer Learning Across Devices for Deep Learning Tensor Programs**
*Yang Bai, Wenqian Zhao, Shuo Yin, Zixiao Wang and Bei Yu*
The training and inference efficiency of ever-larger deep neural networks highly rely on the performance of tensor operators on specific hardware platforms. Therefore, a compilation-based optimization flow with automatic tensor generation and parameter tuning is necessary for efficient model deployment. While compilation-based methods with performance models can provide dynamic and suitable code optimization, they suffer from a large design space exploration with rough measurement accuracy and poor transferability among different hardware platforms. This paper presents ATFormer, a simple yet efficient design with attention-inspired modules to accurately predict the performance of optimized operators by capturing global and long-range dependencies within a complete scheduling space. Compared with state-of-the-arts, ATFormer can predict the optimal implementation of tensor operators to reduce inference time with minimal effort on modern DNN benchmarks. Furthermore, ATFormer with pre-trained parameters can quickly adapt to different workloads and hardware via transfer learning.

11:00-12:30 (East Foyer / Virtual)
**#13 Can We Edit Factual Knowledge by In-Context Learning?**
*Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu and Baobao Chang*
Previous studies have shown that large language models (LLMs) like GPTs store massive factual knowledge in their parameters. However, the stored knowledge could be false or outdated. Traditional knowledge editing methods refine LLMs via fine-tuning on texts containing specific knowledge. However, with the increasing scales of LLMs, these gradient-based approaches bring large computation costs. The trend of model-as-a-service also makes it impossible to modify knowledge in black-box LMs. Inspired by in-context learning (ICL), a new paradigm based on demonstration contexts without parameter updating, we explore whether ICL can edit factual knowledge. To answer this question, we give a comprehensive empirical study of ICL strategies. Experiments show that in-context knowledge editing (IKE), without any gradient and parameter updating, achieves a competitive success rate compared to gradient-based methods on GPT-J (6B) but with much fewer side effects, including less over-editing on similar but unrelated facts and less knowledge forgetting on previously stored knowledge. We also apply the method to larger LMs with tens or hundreds of parameters like OPT-175B, which shows the scalability of our method. The code is available at https://github.com/pkunlp-icler/IKE.

11:00-12:30 (East Foyer / Virtual)
**#14 Discovering Universal Geometry in Embeddings with ICA**
*Hiroaki Yamagiwa, Momose Oyama and Hidetoshi Shimodaira*
This study utilizes Independent Component Analysis (ICA) to unveil a consistent semantic structure within embeddings of words or images. Our approach extracts independent semantic components from the embeddings of a pre-trained model by leveraging anisotropic information that remains after the whitening process in Principal Component Analysis (PCA). We demonstrate that each embedding can be expressed as a composition of a few intrinsic interpretable axes and that these semantic axes remain consistent across different languages, algorithms, and modalities. The discovery of a universal semantic structure in the geometric patterns of embeddings enhances our understanding of the representations in embeddings.

11:00-12:30 (East Foyer / Virtual)
**#15 LLM-FP4: 4-Bit Floating-Point Quantized Transformers**
*Shih-yang Liu, Zechun Liu, Xijie Huang, Pingcheng Dong and Kwang-Ting Cheng*
We propose LLM-FP4 for quantizing both weights and activations in large language models (LLMs) down to 4-bit floating-point values, in a post-training manner. Existing post-training quantization (PTQ) solutions are primarily integer-based and struggle with bit widths below 8 bits. Compared to integer quantization, floating-point (FP) quantization is more flexible and can better handle long-tail or bell-shaped distributions, and it has emerged as a default choice in many hardware platforms. One characteristic of FP quantization is that its performance largely depends on the choice of exponent bits and clipping range. In this regard, we construct a strong FP-PTQ baseline by searching for the optimal quantization parameters. Furthermore, we observe a high inter-channel variance and low intra-channel variance pattern in activation distributions, which adds activation quantization difficulty. We recognize this pattern to be consistent across a spectrum of transformer models designed for diverse tasks such as LLMs, BERT, and Vision Transformer models. To tackle this, we propose per-channel activation quantization and show that these additional scaling factors can be reparameterized as exponential biases of weights, incurring a negligible cost. Our method, for the first time, can quantize both weights and activations in the LLaMA-13B to only 4-bit and achieves an average score of 63.1 on the common sense zero-shot reasoning tasks, which is only 5.8 lower than the full-precision model, significantly outperforming the previous state-of-the-art by 12.7 points. Code is available at: https://github.com/nbasyl/LLM-FP4.

11:00-12:30 (East Foyer / Virtual)
**#16 Are Compressed Language Models Less Subgroup Robust?**
*Leonidas Gee, Andrea Zugarini and Novi Quadrianto*
To reduce the inference cost of large language models, model compression is increasingly used to create smaller scalable models. However, little is known about their robustness to minority subgroups defined by the labels and attributes of a dataset. In this paper, we investigate the effects of 18 different compression methods and settings on the subgroup robustness of BERT language models. We show that worst-group performance does not depend on model size alone, but also on the compression method used. Additionally, we find that model compression does not always worsen the performance on minority subgroups. Altogether, our analysis serves to further research into the subgroup robustness of model compression.

11:00-12:30 (East Foyer / Virtual)
**#17 Prompt as Triggers for Backdoor Attack: Examining the Vulnerability in Language Models**
*Shuai Zhao, Jinming Wen, Anh Tuan Luu, Junbo Zhao and Jie Fu*
The prompt-based learning paradigm, which bridges the gap between pre-training and fine-tuning, achieves state-of-the-art performance on several NLP tasks, particularly in few-shot settings. Despite being widely applied, prompt-based learning is vulnerable to backdoor attacks. Textual backdoor attacks are designed to introduce targeted vulnerabilities into models by poisoning a subset of training samples through trigger injection and label modification. However, they suffer from flaws such as abnormal natural language expressions resulting from the trigger and incorrect labeling of poisoned samples. In this study, we propose ProAttack, a novel and efficient method for performing clean-

label backdoor attacks based on the prompt, which uses the prompt itself as a trigger. Our method does not require external triggers and ensures correct labeling of poisoned samples, improving the stealthy nature of the backdoor attack. With extensive experiments on rich-resource and few-shot text classification tasks, we empirically validate ProAttack's competitive performance in textual backdoor attacks. Notably, in the rich-resource setting, ProAttack achieves state-of-the-art attack success rates in the clean-label backdoor attack benchmark without external triggers.

11:00-12:30 (East Foyer / Virtual)
**#18 Vicinal Risk Minimization for Few-Shot Cross-lingual Transfer in Abusive Language Detection**
*Gretel Liz De la Peña Sarracén, Paolo Rosso, Robert Litschko, Goran Glavaš and Simone Paolo Ponzetto*
Cross-lingual transfer learning from high-resource to medium and low-resource languages has shown encouraging results. However, the scarcity of resources in target languages remains a challenge. In this work, we resort to data augmentation and continual pre-training for domain adaptation to improve cross-lingual abusive language detection. For data augmentation, we analyze two existing techniques based on vicinal risk minimization and propose MIXAG, a novel data augmentation method which interpolates pairs of instances based on the angle of their representations. Our experiments involve seven languages typologically distinct from English and three different domains. The results reveal that the data augmentation strategies can enhance few-shot cross-lingual abusive language detection. Specifically, we observe that consistently in all target languages, MIXAG improves significantly in multidomain and multilingual environments. Finally, we show through an error analysis how the domain adaptation can favour the class of abusive texts (reducing false negatives), but at the same time, declines the precision of the abusive language detection model.

11:00-12:30 (East Foyer / Virtual)
**#19 Compressing Context to Enhance Inference Efficiency of Large Language Models**
*Yucheng Li, Bo Dong, Frank Guerin and Chenghua Lin*
Large language models (LLMs) achieved remarkable performance across various tasks. However, they face challenges in managing long documents and extended conversations, due to significantly increased computational requirements, both in memory and inference time, and potential context truncation when the input exceeds the LLM's fixed context length. This paper proposes a method called *Selective Context* that enhances the inference efficiency of LLMs by identifying and pruning redundancy in the input context to make the input more compact. We test our approach using common data sources requiring long context processing: arXiv papers, news articles, and long conversations, on tasks of summarisation, question answering, and response generation. Experimental results show that Selective Context significantly reduces memory cost and decreases generation latency while maintaining comparable performance compared to that achieved when full context is used. Specifically, we achieve a 50% reduction in context cost, resulting in a 36% reduction in inference memory usage and a 32% reduction in inference time, while observing only a minor drop of .023 in BERTscore and .038 in faithfulness on four downstream applications, indicating that our method strikes a good balance between efficiency and performance.

11:00-12:30 (East Foyer / Virtual)
**#20 QTSumm: Query-Focused Summarization over Tabular Data**
*Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev and Arman Cohan*
People primarily consult tables to conduct data analysis or answer specific questions. Text generation systems that can provide accurate table summaries tailored to users' information needs can facilitate more efficient access to relevant data insights. Motivated by this, we define a new query-focused table summarization task, where text generation models have to perform human-like reasoning and analysis over the given table to generate a tailored summary. We introduce a new benchmark named QTSumm for this task, which contains 7,111 human-annotated query-summary pairs over 2,934 tables covering diverse topics. We investigate a set of strong baselines on QTSumm, including text generation, table-to-text generation, and large language models. Experimental results and manual analysis reveal that the new task presents significant challenges in table-to-text generation for future research. Moreover, we propose a new approach named ReFactor, to retrieve and reason over query-relevant information from tabular data to generate several natural language facts. Experimental results demonstrate that ReFactor can bring effective improvements to baselines by concatenating the generated facts to the model input. Our data and code are publicly available at https://github.com/yale-nlp/QTSumm.

11:00-12:30 (East Foyer / Virtual)
**#21 Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis**
*Seraphina Goldfarb-Tarrant, Björn Ross and Adam Lopez*
Sentiment analysis (SA) systems are widely deployed in many of the world's languages, and there is well-documented evidence of demographic bias in these systems. In languages beyond English, scarcer training data is often supplemented with transfer learning using pre-trained models, including multilingual models trained on other languages. In some cases, even supervision data comes from other languages. Does cross-lingual transfer also import new biases? To answer this question, we use counterfactual evaluation to test whether gender or racial biases are imported when using cross-lingual transfer, compared to a monolingual transfer setting. Across five languages, we find that systems using cross-lingual transfer usually become more biased than their monolingual counterparts. We also find racial biases to be much more prevalent than gender biases. To spur further research on this topic, we release the sentiment models we used for this study, and the intermediate checkpoints throughout training, yielding 1,525 distinct models; we also release our evaluation code.

11:00-12:30 (East Foyer / Virtual)
**#22 Generating Summaries with Controllable Readability Levels**
*Leonardo F. R. Ribeiro, Mohit Bansal and Markus Dreyer*
Readability refers to how easily a reader can understand a written text. Several factors affect the readability level, such as the complexity of the text, its subject matter, and the reader's background knowledge. Generating summaries based on different readability levels is critical for enabling knowledge consumption by diverse audiences. However, current text generation approaches lack refined control, resulting in texts that are not customized to readers' proficiency levels. In this work, we bridge this gap and study techniques to generate summaries at specified readability levels. Unlike previous methods that focus on a specific readability level (e.g., lay summarization), we generate summaries with fine-grained control over their readability. We develop three text generation techniques for controlling readability: (1) instruction-based readability control, (2) reinforcement learning to minimize the gap between requested and observed readability and (3) a decoding approach that uses lookahead to estimate the readability of upcoming decoding steps. We show that our generation methods significantly improve readability control on news summarization (CNN/DM dataset), as measured by various readability metrics and human judgement, establishing strong baselines for controllable readability in summarization.

11:00-12:30 (East Foyer / Virtual)
**#23 Disentangling Transformer Language Models as Superposed Topic Models**
*Jia Peng Lim and Hady W. Lauw*
Topic Modelling is an established research area where the quality of a given topic is measured using coherence metrics. Often, we infer topics from Neural Topic Models (NTM) by interpreting their decoder weights, consisting of top-activated words projected from individual neurons.

Transformer-based Language Models (TLM) similarly consist of decoder weights. However, due to its hypothesised superposition properties, the final logits originating from the residual path are considered uninterpretable. Therefore, we posit that we can interpret TLM as superposed NTM by proposing a novel weight-based, model-agnostic and corpus-agnostic approach to search and disentangle decoder-only TLM, potentially mapping individual neurons to multiple coherent topics. Our results show that it is empirically feasible to disentangle coherent topics from GPT-2 models using the Wikipedia corpus. We validate this approach for GPT-2 models using Zero-Shot Topic Modelling. Finally, we extend the proposed approach to disentangle and analyse LLaMA models.

11:00-12:30 (East Foyer / Virtual)
**#24 Context Compression for Auto-regressive Transformers with Sentinel Tokens**
*Siyu Ren, Qi Jia and Kenny Q. Zhu*
The quadratic complexity of the attention module makes it gradually become the bulk of compute in Transformer-based LLMs during generation. Moreover, the excessive key-value cache that arises when dealing with long inputs also brings severe issues on memory footprint and inference latency. In this work, we propose a plug-and-play approach that is able to incrementally compress the intermediate activation of a specified span of tokens into compact ones, thereby reducing both memory and computational cost when processing subsequent context. Experiments on both in-domain language modeling and zero-shot open-ended document generation demonstrate the advantage of our approach over sparse attention baselines in terms of fluency, n-gram matching, and semantic similarity. At last, we comprehensively profile the benefit of context compression on improving the system throughout. Code is available at `https://github.com/DRSY/KV_Compression`.

11:00-12:30 (East Foyer / Virtual)
**#25 Transformer-based Live Update Generation for Soccer Matches from Microblog Posts**
*Masashi Oshika, Kosuke Yamada, Ryohei Sasano and Koichi Takeda*
It has been known to be difficult to generate adequate sports updates from a sequence of vast amounts of diverse live tweets, although the live sports viewing experience with tweets is gaining the popularity. In this paper, we focus on soccer matches and work on building a system to generate live updates for soccer matches from tweets so that users can instantly grasp a match's progress and enjoy the excitement of the match from raw tweets. Our proposed system is based on a large pre-trained language model and incorporates a mechanism to control the number of updates and a mechanism to reduce the redundancy of duplicate and similar updates.

11:00-12:30 (East Foyer / Virtual)
**#26 LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models**
*Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria and Roy Ka-Wei Lee*
The success of large language models (LLMs), like GPT-4 and ChatGPT, has led to the development of numerous cost-effective and accessible alternatives that are created by finetuning open-access LLMs with task-specific data (e.g., ChatDoctor) or instruction data (e.g., Alpaca). Among the various fine-tuning methods, adapter-based parameter-efficient fine-tuning (PEFT) is undoubtedly one of the most attractive topics, as it only requires fine-tuning a few external parameters instead of the entire LLMs while achieving comparable or even better performance. To enable further research on PEFT methods of LLMs, this paper presents LLM-Adapters, an easy-to-use framework that integrates various adapters into LLMs and can execute these adapter-based PEFT methods of LLMs for different tasks. The framework includes state-of-the-art open-access LLMs such as LLaMA, BLOOM, and GPT-J, as well as widely used adapters such as Series adapters, Parallel adapter, Prompt-based learning and Reparametrization-based methods. Moreover, we conduct extensive empirical studies on the impact of adapter types, placement locations, and hyper-parameters to the best design for each adapter-based methods. We evaluate the effectiveness of the adapters on fourteen datasets from two different reasoning tasks, Arithmetic Reasoning and Commonsense Reasoning. The results demonstrate that using adapter-based PEFT in smaller-scale LLMs (7B) with few extra trainable parameters yields comparable, and in some cases superior, performance to powerful LLMs (175B) in zero-shot inference on simple math reasoning datasets.

11:00-12:30 (East Foyer / Virtual)
**#27 A Multi-Task Dataset for Assessing Discourse Coherence in Chinese Essays: Structure, Theme, and Logic Analysis**
*Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai and Yuanbin Wu*
This paper introduces the **C**hinese **E**ssay **D**iscourse **C**oherence **C**orpus (**CEDCC**), a multi-task dataset for assessing discourse coherence. Existing research tends to focus on isolated dimensions of discourse coherence, a gap which the CEDCC addresses by integrating coherence grading, topical continuity, and discourse relations. This approach, alongside detailed annotations, captures the subtleties of real-world texts and stimulates progress in Chinese discourse coherence analysis. Our contributions include the development of the CEDCC, the establishment of baselines for further research, and the demonstration of the impact of coherence on discourse relation recognition and automated essay scoring. The dataset and related codes is available at `https://github.com/cubenlp/CEDCC_corpus`.

11:00-12:30 (East Foyer / Virtual)
**#28 Unraveling Feature Extraction Mechanisms in Neural Networks**
*Xiaobing Sun, Jiaxi Li and Wei Lu*
The underlying mechanism of neural networks in capturing precise knowledge has been the subject of consistent research efforts. In this work, we propose a theoretical approach based on Neural Tangent Kernels (NTKs) to investigate such mechanisms. Specifically, considering the infinite network width, we hypothesize the learning dynamics of target models may intuitively unravel the features they acquire from training data, deepening our insights into their internal mechanisms. We apply our approach to several fundamental models and reveal how these models leverage statistical features during gradient descent and how they are integrated into final decisions. We also discovered that the choice of activation function can affect feature extraction. For instance, the use of the ReLU activation function could potentially introduce a bias in features, providing a plausible explanation for its replacement with alternative functions in recent pre-trained language models. Additionally, we find that while self-attention and CNN models may exhibit limitations in learning n-grams, multiplication-based models seem to excel in this area. We verify these theoretical findings through experiments and find that they can be applied to analyze language modeling tasks, which can be regarded as a special variant of classification. Our work may offer insights into the roles and capacities of fundamental modules within deep neural networks including large language models.

11:00-12:30 (East Foyer / Virtual)
**#29 HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts**
*Truong Giang Do, Le Huy Khiem, Quang Pham, TrungTin Nguyen, Thanh-Nam Doan, Binh T. Nguyen, Chenghao Liu, Savitha Ramasamy, Xiaoli Li and Steven Hoi*
By routing input tokens to only a few split experts, Sparse Mixture-of-Experts has enabled efficient training of large language models. Recent findings suggest that fixing the routers can achieve competitive performance by alleviating the collapsing problem, where all experts eventually learn similar representations. However, this strategy has two key limitations: (i) the policy derived from random routers might be sub-optimal, and (ii) it requires extensive resources during training and evaluation, leading to limited efficiency gains. This work introduces `HyperRouter`, which dynamically generates the router's parameters through a fixed hypernetwork and trainable embeddings to achieve a balance between training the routers and freezing them to learn an improved routing policy. Extensive experiments across a wide range of tasks demonstrate the superior performance and efficiency gains of `HyperRouter` compared to existing routing methods. Our implementa-

tion is publicly available at https://github.com/giangdip2410/HyperRouter.

11:00-12:30 (East Foyer / Virtual)
**#30 TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models**
*Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind and Idan Szpektor*
Factual consistency evaluation is often conducted using Natural Language Inference (NLI) models, yet these models exhibit limited success in evaluating summaries. Previous work improved such models with synthetic training data. However, the data is typically based on perturbed human-written summaries, which often differ in their characteristics from real model-generated summaries and have limited coverage of possible factual errors. Alternatively, large language models (LLMs) have recently shown promising results in directly evaluating generative tasks, but are too computationally expensive for practical use. Motivated by these limitations, we introduce TrueTeacher, a method for generating synthetic data by annotating diverse model-generated summaries using a LLM. Unlike prior work, TrueTeacher does not rely on human-written summaries, and is multilingual by nature. Experiments on the TRUE benchmark show that a student model trained using our data, substantially outperforms both the state-of-the-art model with similar capacity, and the LLM teacher. In a systematic study, we compare TrueTeacher to existing synthetic data generation methods and demonstrate its superiority and robustness to domain-shift. We also show that our method generalizes to multilingual scenarios. Lastly, we release our large scale synthetic dataset (1.4M examples), generated using TrueTeacher, and a checkpoint trained on this data.

11:00-12:30 (East Foyer / Virtual)
**#31 Anchoring Fine-tuning of Sentence Transformer with Semantic Label Information for Efficient Truly Few-shot Classification**
*Amalie Brogaard Pauli, Leon Derczynski and Ira Assent*
Few-shot classification is a powerful technique, but training requires substantial computing power and data. We propose an efficient method with small model sizes and less training data with only 2-8 training instances per class. Our proposed method, AncSetFit, targets low data scenarios by anchoring the task and label information through sentence embeddings in fine-tuning a Sentence Transformer model. It uses contrastive learning and a triplet loss to enforce training instances of a class to be closest to its own textual semantic label information in the embedding space - and thereby learning to embed different class instances more distinct. AncSetFit obtains strong performance in data-sparse scenarios compared to existing methods across SST-5, Emotion detection, and AG News data, even with just two examples per class.

11:00-12:30 (East Foyer / Virtual)
**#32 This Reads Like That: Deep Learning for Interpretable Natural Language Processing**
*Claudio Fanconi, Moritz Vandenhirtz, Severin Husmann and Julia E Vogt*
Prototype learning, a popular machine learning method designed for inherently interpretable decisions, leverages similarities to learned prototypes for classifying new data. While it is mainly applied in computer vision, in this work, we build upon prior research and further explore the extension of prototypical networks to natural language processing. We introduce a learned weighted similarity measure that enhances the similarity computation by focusing on informative dimensions of pre-trained sentence embeddings. Additionally, we propose a post-hoc explainability mechanism that extracts prediction-relevant words from both the prototype and input sentences. Finally, we empirically demonstrate that our proposed method not only improves predictive performance on the AG News and RT Polarity datasets over a previous prototype-based approach, but also improves the faithfulness of explanations compared to rationale-based recurrent convolutions.

11:00-12:30 (East Foyer / Virtual)
**#33 Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning**
*Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou and Xu Sun*
In-context learning (ICL) emerges as a promising capability of large language models (LLMs) by providing them with demonstration examples to perform diverse tasks. However, the underlying mechanism of how LLMs learn from the provided context remains under-explored. In this paper, we investigate the working mechanism of ICL through an information flow lens. Our findings reveal that label words in the demonstration examples function as anchors: (1) semantic information aggregates into label word representations during the shallow computation layers' processing; (2) the consolidated information in label words serves as a reference for LLMs' final predictions. Based on these insights, we introduce an anchor re-weighting method to improve ICL performance, a demonstration compression technique to expedite inference, and an analysis framework for diagnosing ICL errors in GPT2-XL. The promising applications of our findings again validate the uncovered ICL working mechanism and pave the way for future studies.

11:00-12:30 (East Foyer / Virtual)
**#34 Memorisation Cartography: Mapping out the Memorisation-Generalisation Continuum in Neural Machine Translation**
*Verna Dankers, Ivan Titov and Dieuwke Hupkes*
When training a neural network, it will quickly memorise some source-target mappings from your dataset but never learn some others. Yet, memorisation is not easily expressed as a binary feature that is good or bad: individual datapoints lie on a memorisation-generalisation continuum. What determines a datapoint's position on that spectrum, and how does that spectrum influence neural models' performance? We address these two questions for neural machine translation (NMT) models. We use the counterfactual memorisation metric to (1) build a resource that places 5M NMT datapoints on a memorisation-generalisation map, (2) illustrate how the datapoints' surface-level characteristics and a models' per-datum training signals are predictive of memorisation in NMT, (3) and describe the influence that subsets of that map have on NMT systems' performance.

11:00-12:30 (East Foyer / Virtual)
**#35 A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis**
*Alessandro Stolfo, Yonatan Belinkov and Mrinmaya Sachan*
Mathematical reasoning in large language models (LMs) has garnered significant attention in recent work, but there is a limited understanding of how these models process and store information related to arithmetic tasks within their architecture. In order to improve our understanding of this aspect of language models, we present a mechanistic interpretation of Transformer-based LMs on arithmetic questions using a causal mediation analysis framework. By intervening on the activations of specific model components and measuring the resulting changes in predicted probabilities, we identify the subset of parameters responsible for specific predictions. This provides insights into how information related to arithmetic is processed by LMs. Our experimental results indicate that LMs process the input by transmitting the information relevant to the query from mid-sequence early layers to the final token using the attention mechanism. Then, this information is processed by a set of MLP modules, which generate result-related information that is incorporated into the residual stream. To assess the specificity of the observed activation dynamics, we compare the effects of different model components on arithmetic queries with other tasks, including number retrieval from prompts and factual knowledge questions.

11:00-12:30 (East Foyer / Virtual)
**#36 A Frustratingly Easy Post-Training Quantization Scheme for LLMs**
*Yongkweon Jeon, Chungman Lee, Kyungphil Park and Ho-young Kim*
Efficient inference has become crucial for hyper-scale AI models, including large language models, as their parameter count continues to

increase for enhanced performance. This necessity holds true regardless of the computing environment, whether it be mobile devices or cloud servers. Quantization emerges as a solution to alleviate the computational burden during inference. By representing models with a reduced bit-width, quantization minimizes the frequency of DRAM access while fully exploiting the parallelism of operations through a dense matrix format. Consequently, quantized models achieve low end-to-end latency and optimize resource utilization by addressing both memory and computing bottlenecks. In this paper, we propose a straightforward post-training quantization scheme, called Z-FOLD, that fully utilizes the feature of the Transformer structure widely employed in large language models.

11:00-12:30 (East Foyer / Virtual)
**#37 Conceptor-Aided Debiasing of Large Language Models**
*Li S. Yifei, Lyle Ungar and João Sedoc*
Pre-trained large language models (LLMs) reflect the inherent social biases of their training corpus. Many methods have been proposed to mitigate this issue, but they often fail to debias or they sacrifice model accuracy. We use *conceptors*–a soft projection method–to identify and remove the bias subspace in LLMs such as BERT and GPT. We propose two methods of applying conceptors (1) bias subspace projection by post-processing by the conceptor NOT operation; and (2) a new architecture, conceptor-intervened BERT (CI-BERT), which explicitly incorporates the conceptor projection into all layers during training. We find that conceptor post-processing achieves state-of-the-art (SoTA) debiasing results while maintaining LLMs' performance on the GLUE benchmark. Further, it is robust in various scenarios and can mitigate intersectional bias efficiently by its AND operation on the existing bias subspaces. Although CI-BERT's training takes all layers' bias into account and can beat its post-processing counterpart in bias mitigation, CI-BERT reduces the language model accuracy. We also show the importance of carefully constructing the bias subspace. The best results are obtained by removing outliers from the list of biased words, combining them (via the OR operation), and computing their embeddings using the sentences from a cleaner corpus.

11:00-12:30 (East Foyer / Virtual)
**#38 A Rose by Any Other Name would not Smell as Sweet: Social Bias in Names Mistranslation**
*Sandra Camille Sandoval, Jieyu Zhao, Marine Carpuat and Hal Daumé III*
We ask the question: Are there widespread disparities in machine translations of names across race/ethnicity, and gender? We hypothesize that the translation quality of names and surrounding context will be lower for names associated with US racial and ethnic minorities due to these systems' tendencies to standardize language to predominant language patterns. We develop a dataset of names that are strongly demographically aligned and propose a translation evaluation procedure based on round-trip translation. We analyze the effect of name demographics on translation quality using generalized linear mixed effects models and find that the ability of translation systems to correctly translate female-associated names is significantly lower than male-associated names. This effect is particularly pronounced for female-associated names that are also associated with racial (Black) and ethnic (Hispanic) minorities. This disparity in translation quality between social groups for something as personal as someone's name has significant implications for people's professional, personal, and cultural identities, self-worth and ease of communication. Our findings suggest that more MT research is needed to improve the translation of names and to provide high-quality service for users regardless of gender, race, and ethnicity.

11:00-12:30 (East Foyer / Virtual)
**#39 Outlier Suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling**
*Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo and Xianglong Liu*
Post-training quantization (PTQ) of transformer language models faces significant challenges due to the existence of detrimental outliers in activations. We observe that these outliers are concentrated in specific channels and are asymmetric across channels. To address this issue, we propose the Outlier Suppression+ (OS+) framework, which contains the channel-wise shifting for asymmetry and channel-wise scaling for concentration. We show that these operations can be seamlessly migrated into subsequent modules while maintaining equivalence. Second, we propose a fast and stable scheme to calculate effective shifting and scaling values. The channel-wise shifting aligns the center of each channel for removal of outlier asymmetry. The channel-wise scaling quantitatively evaluates changes brought by migration and quantization for better quantization burden balance. We validate our OS+ under both standard and fine-grained quantization settings with models including BERT, OPT, BLOOM, BLOOMZ, and LLaMA. Comprehensive results across various tasks demonstrate the superiority of our approach. Especially, with standard quantization, OS+ can achieve near-floating-point performance on both small models and large language models on 8-bit and 6-bit. Besides, we establish a new state-of-the-art for 4-bit BERT with 15.5% improvement. Our code is available at https://github.com/ModelTC/Outlier_Suppression_Plus.

11:00-12:30 (East Foyer / Virtual)
**#40 Comparing Biases and the Impact of Multilingual Training across Multiple Languages**
*Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli and Dan Roth*
Studies in bias and fairness in natural language processing have primarily examined social biases within a single language and/or across few attributes (e.g. gender, race). However, biases can manifest differently across various languages for individual attributes. As a result, it is critical to examine biases within each language and attribute. Of equal importance is to study how these biases compare across languages and how the biases are affected when training a model on multilingual data versus monolingual data. We present a bias analysis across Italian, Chinese, English, Hebrew, and Spanish on the downstream sentiment analysis task to observe whether specific demographics are viewed more positively. We study bias similarities and differences across these languages and investigate the impact of multilingual vs. monolingual training data. We adapt existing sentiment bias templates in English to Italian, Chinese, Hebrew, and Spanish for four attributes: race, religion, nationality, and gender. Our results reveal similarities in bias expression such as favoritism of groups that are dominant in each language's culture (e.g. majority religions and nationalities). Additionally, we find an increased variation in predictions across protected groups, indicating bias amplification, after multilingual finetuning in comparison to multilingual pretraining.

11:00-12:30 (East Foyer / Virtual)
**#41 Text encoders bottleneck compositionality in contrastive vision-language models**
*Amita Kamath, Jack Hessel and Kai-Wei Chang*
Performant vision-language (VL) models like CLIP represent captions using a single vector. How much information about language is lost in this bottleneck? We first curate CompPrompts, a set of increasingly compositional image captions that VL models should be able to capture (e.g., single object, to object+property, to multiple interacting objects). Then, we train text-only recovery probes that aim to reconstruct captions from single-vector text representations produced by several VL models. This approach does not require images, allowing us to test on a broader range of scenes compared to prior work. We find that: 1) CLIP's text encoder falls short on more compositional inputs, including object relationships, attribute-object association, counting, and negations; 2) some text encoders work significantly better than others; and 3) text-only recovery performance predicts multimodal matching performance on ControlledImCaps: a new evaluation benchmark we collect and release consisting of fine-grained compositional images and captions. Specifically, our results suggest text-only recoverability is a necessary (but not sufficient) condition for modeling compositional factors in contrastive VL models. We release our datasets and code.

11:00-12:30 (East Foyer / Virtual)
**#42 "Are Your Explanations Reliable?" Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and**

**Adversarial Attack**
*Christopher Burger, Lingwei Chen and Thai Le*
LIME has emerged as one of the most commonly referenced tools in explainable AI (XAI) frameworks that is integrated into critical machine learning applications (e.g., healthcare and finance). However, its stability remains little explored, especially in the context of text data, due to the unique text-space constraints. To address these challenges, in this paper, we first evaluate the inherent instability of LIME on text data to establish a baseline, and then propose a novel algorithm XAIFooler to perturb text inputs and manipulate explanations that casts investigation on the stability of LIME as a text perturbation optimization problem. XAIFooler conforms to the constraints to preserve text semantics and original prediction with small perturbations, and introduces Rank-biased Overlap (RBO) as a key part to guide the optimization of XAIFooler that satisfies all the requirements for explanation similarity measure. Extensive experiments on real-world text datasets demonstrate that XAIFooler significantly outperforms all baselines by large margins in its ability to manipulate LIME's explanations with high semantic preservability.

11:00-12:30 (East Foyer / Virtual)
**#43 Fair Without Leveling Down: A New Intersectional Fairness Definition**
*Gaurav Maheshwari, Aurélien Bellet, Pascal Denis and Mikaela Keller*
In this work, we consider the problem of intersectional group fairness in the classification setting, where the objective is to learn discrimination-free models in the presence of several intersecting sensitive groups. First, we illustrate various shortcomings of existing fairness measures commonly used to capture intersectional fairness. Then, we propose a new definition called the $\alpha$-Intersectional Fairness, which combines the absolute and the relative performance across sensitive groups and can be seen as a generalization of the notion of differential fairness. We highlight several desirable properties of the proposed definition and analyze its relation to other fairness measures. Finally, we benchmark multiple popular in-processing fair machine learning approaches using our new fairness definition and show that they do not achieve any improvement over a simple baseline. Our results reveal that the increase in fairness measured by previous definitions hides a "leveling down" effect, i.e., degrading the best performance over groups rather than improving the worst one.

11:00-12:30 (East Foyer / Virtual)
**#44 Discourse Structures Guided Fine-grained Propaganda Identification**
*Yuanyuan Lei and Ruihong Huang*
Propaganda is a form of deceptive narratives that instigate or mislead the public, usually with a political purpose. In this paper, we aim to identify propaganda in political news at two fine-grained levels: sentence-level and token-level. We observe that propaganda content is more likely to be embedded in sentences that attribute causality or assert contrast to nearby sentences, as well as seen in opinionated evaluation, speculation and discussions of future expectation. Hence, we propose to incorporate both local and global discourse structures for propaganda discovery and construct two teacher models for identifying PDTB-style discourse relations between nearby sentences and common discourse roles of sentences in a news article respectively. We further devise two methods to incorporate the two types of discourse structures for propaganda identification by either using teacher predicted probabilities as additional features or soliciting guidance in a knowledge distillation framework. Experiments on the benchmark dataset demonstrate that leveraging guidance from discourse structures can significantly improve both precision and recall of propaganda content identification.

11:00-12:30 (East Foyer / Virtual)
**#45 'Don't Get Too Technical with Me': A Discourse Structure-Based Framework for Automatic Science Journalism**
*Ronald Cardenas, Bingsheng Yao, Dakuo Wang and Yufang Hou*
Science journalism refers to the task of reporting technical findings of a scientific paper as a less technical news article to the general public audience. We aim to design an automated system to support this real-world task (i.e., automatic science journalism ) by 1) introducing a newly-constructed and real-world dataset (SciTechNews), with tuples of a publicly-available scientific paper, its corresponding news article, and an expert-written short summary snippet; 2) proposing a novel technical framework that integrates a paper's discourse structure with its metadata to guide generation; and, 3) demonstrating with extensive automatic and human experiments that our model outperforms other baseline methods (e.g. Alpaca and ChatGPT) in elaborating a content plan meaningful for the target audience, simplify the information selected, and produce a coherent final report in a layman's style.

11:00-12:30 (East Foyer / Virtual)
**#46 When are Lemons Purple? The Concept Association Bias of Vision-Language Models**
*Yingtian Tang, Yutaro Yamada, Yoyo Minzhi Zhang and Ilker Yildirim*
Large-scale vision-language models such as CLIP have shown impressive performance on zero-shot image classification and image-to-text retrieval. However, such performance does not realize in tasks that require a finer-grained correspondence between vision and language, such as Visual Question Answering (VQA). We investigate why this is the case, and report an interesting phenomenon of vision-language models, which we call the Concept Association Bias (CAB), as a potential cause of the difficulty of applying these models to VQA and similar tasks. We find that models with CAB tend to treat input as a bag of concepts and attempt to fill in the other missing concept crossmodally, leading to an unexpected zero-shot prediction. We demonstrate CAB by showing that CLIP's zero-shot classification performance greatly suffers when there is a strong concept association between an object (e.g. eggplant) and an attribute (e.g. color purple). We also show that the strength of CAB predicts the performance on VQA. We observe that CAB is prevalent in vision-language models trained with contrastive losses, even when autoregressive losses are jointly employed. However, a model that solely relies on autoregressive loss seems to exhibit minimal or no signs of CAB.

11:00-12:30 (East Foyer / Virtual)
**#47 Faster Minimum Bayes Risk Decoding with Confidence-based Pruning**
*Julius Cheng and Andreas Vlachos*
Minimum Bayes risk (MBR) decoding outputs the hypothesis with the highest expected utility over the model distribution for some utility function. It has been shown to improve accuracy over beam search in conditional language generation problems and especially neural machine translation, in both human and automatic evaluations. However, the standard sampling-based algorithm for MBR is substantially more computationally expensive than beam search, requiring a large number of samples as well as a quadratic number of calls to the utility function, limiting its applicability. We describe an algorithm for MBR which gradually grows the number of samples used to estimate the utility while pruning hypotheses that are unlikely to have the highest utility according to confidence estimates obtained with bootstrap sampling. Our method requires fewer samples and drastically reduces the number of calls to the utility function compared to standard MBR while being statistically indistinguishable in terms of accuracy. We demonstrate the effectiveness of our approach in experiments on three language pairs, using chrF++ and COMET as utility/evaluation metrics.

11:00-12:30 (East Foyer / Virtual)
**#48 Why LLMs Hallucinate, and How to Get (Evidential) Closure: Perceptual, Intensional, and Extensional Learning for Faithful Natural Language Generation**
*Adam Bouyamourn*

We show that LLMs hallucinate because their output is not constrained to be synonymous with claims for which they have evidence: a condition that we call evidential closure. Information about the truth or falsity of sentences is not statistically identified in the standard neural language generation setup, and so cannot be conditioned on to generate new strings. We then show how to constrain LLMs to produce output that satisfies evidential closure. A multimodal LLM must learn about the external world (perceptual learning); it must learn a mapping from strings to states of the world (extensional learning); and, to achieve fluency when generalizing beyond a body of evidence, it must learn mappings from strings to their synonyms (intensional learning). The output of a unimodal LLM must be synonymous with strings in a validated evidence set. Finally, we present a heuristic procedure, Learn-Babble-Prune, that yields faithful output from an LLM by rejecting output that is not synonymous with claims for which the LLM has evidence.

11:00-12:30 (East Foyer / Virtual)
**#49 MaNtLE: Model-agnostic Natural Language Explainer**
*Rakesh R Menon, Kerem Zaman and Shashank Srivastava*
Understanding the internal reasoning behind the predictions of machine learning systems is increasingly vital, given their rising adoption and acceptance. While previous approaches, such as LIME generate algorithmic explanations by attributing importance to input features for individual examples, recent research indicates that practitioners prefer examining language explanations that explain sub-groups of examples (Lakkaraju et al., 2022). In this paper, we introduce MaNtLE, a model-agnostic natural language explainer that analyzes a set of classifier predictions and generates faithful natural language explanations of classifier rationale for structured classification tasks. MaNtLE uses multi-task training on thousands of synthetic classification tasks to generate faithful explanations. Our experiments indicate that, on average, MaNtLE-generated explanations are at least 11% more faithful compared to LIME and Anchors explanations across three tasks. Human evaluations demonstrate that users can better predict model behavior using explanations from MaNtLE compared to other techniques.

11:00-12:30 (East Foyer / Virtual)
**#50 MediaHG: Rethinking Eye-catchy Features in Social Media Headline Generation**
*Boning Zhang and Yang Yang*
An attractive blog headline on social media platforms can immediately grab readers and trigger more clicks. However, a good headline shall not only contract the main content but also be eye-catchy with domain platform features, which are decided by the website's users and objectives. With effective headlines, bloggers can obtain more site traffic and profits, while readers can have easier access to topics of interest. In this paper, we propose a disentanglement-based headline generation model: MediaHG (Social Media Headline Generation), which can balance the content and contextual features. Specifically, we first devise a sample module for various document views and generate the corresponding headline candidates. Then, we incorporate contrastive learning and auxiliary multi-task to choose the best domain-suitable headline, according to the disentangled budgets. Besides, our separated processing gains more flexible adaptation for other headline generation tasks with special domain features. Our model is built from the content and headlines of 70k hot posts collected from REDBook, a Chinese social media platform for daily sharing. Experimental results with language metrics ROUGE and human evaluation show the improvement in the headline generation task for the platform.

11:00-12:30 (East Foyer / Virtual)
**#51 Revisiting Block-based Quantisation: What is Important for Sub-8-bit LLM Inference?**
*Cheng Zhang, Jianyi Cheng, Ilia Shumailov, George Anthony Constantinides and Yiren Zhao*
The inference of Large language models (LLMs) requires immense computation and memory resources. To curtail these costs, quantisation has emerged as a promising solution, but existing LLM quantisation mainly focuses on 8-bit. In this work, we explore the statistical and learning properties of the LLM layer and attribute the bottleneck of LLM quantisation to numerical scaling offsets. To address this, we adapt block quantisations for LLMs, a family of methods that share scaling factors across packed numbers. Block quantisations efficiently reduce the numerical scaling offsets solely from an arithmetic perspective, without additional treatments in the computational path. Our nearly-lossless quantised 6-bit LLMs achieve a $19\times$ higher arithmetic density and $5\times$ memory density than the float32 baseline, surpassing the prior art 8-bit quantisation by $2.5\times$ in arithmetic density and $1.2\times$ in memory density, without requiring any data calibration or re-training. We also share our insights into sub-8-bit LLM quantisation, including the mismatch between activation and weight distributions, optimal fine-tuning strategies, and a lower quantisation granularity inherent in the statistical properties of LLMs. The latter two tricks enable nearly-lossless 4-bit LLMs on downstream tasks. Our code is open-sourced.

11:00-12:30 (East Foyer / Virtual)
**#52 Text-Transport: Toward Learning Causal Effects of Natural Language**
*Victoria Lin, Louis-Philippe Morency and Eli Ben-Michael*
As language technologies gain prominence in real-world settings, it is important to understand *how* changes to language affect reader perceptions. This can be formalized as the *causal effect* of varying a linguistic attribute (e.g., sentiment) on a reader's response to the text. In this paper, we introduce Text-Transport, a method for estimation of causal effects from natural language under any text distribution. Current approaches for valid causal effect estimation require strong assumptions about the data, meaning the data from which one *can* estimate valid causal effects often is not representative of the actual target domain of interest. To address this issue, we leverage the notion of distribution shift to describe an estimator that *transports* causal effects between domains, bypassing the need for strong assumptions in the target domain. We derive statistical guarantees on the uncertainty of this estimator, and we report empirical results and analyses that support the validity of Text-Transport across data settings. Finally, we use Text-Transport to study a realistic setting—hate speech on social media—in which causal effects do shift significantly between text domains, demonstrating the necessity of transport when conducting causal inference on natural language.

11:00-12:30 (East Foyer / Virtual)
**#53 Parameter-efficient Tuning for Large Language Model without Calculating Its Gradients**
*Feihu Jin, Jiajun Zhang and Chengqing Zong*
Fine-tuning all parameters of large language models (LLMs) requires significant computational resources and is time-consuming. Recent parameter-efficient tuning methods such as Adapter tuning, Prefix tuning, and LoRA allow for updating a small subset of parameters in large language models. However, they can only save approximately 30% of the training memory requirements, due to the problem that gradient computation and backpropagation are still necessary for these methods. This paper proposes a novel parameter-efficient tuning method for LLMs without calculating their gradients. Leveraging the discernible similarities between the parameter-efficient modules of the same task learned by both large and small language models, we put forward a strategy for transferring the parameter-efficient modules, originally derived from small language models to much larger ones. To ensure a smooth and effective adaptation process, we further introduce a Bridge model to guarantee dimensional consistency while also stimulating a dynamic interaction between the models. We demonstrate the effectiveness of our method using the T5 and GPT-2 series of language models on the SuperGLUE benchmark. Our method achieves comparable performance to both fine-tuning and parameter-efficient tuning on large language models without needing gradient-based optimization. Additionally, our method achieves up to 5.7x memory reduction compared to parameter-efficient tuning.

11:00-12:30 (East Foyer / Virtual)

**#54 Select, Prompt, Filter: Distilling Large Language Models for Summarizing Conversations**
*Minh-Quang Pham, Sathish Reddy Indurthi, Shamil Chollampatt and Marco Turchi*
Large language models (LLMs) like ChatGPT can be expensive to train, deploy, and use for specific natural language generation tasks such as text summarization and for certain domains. A promising alternative is to fine-tune relatively smaller language models (LMs) on a particular task using high-quality, in-domain datasets. However, it can be prohibitively expensive to get such high-quality training data. This issue has been mitigated by generating weakly supervised data via knowledge distillation (KD) of LLMs. We propose a three-step approach to distill ChatGPT and fine-tune smaller LMs for summarizing forum conversations. More specifically, we design a method to selectively sample a large unannotated corpus of forum conversation using a semantic similarity metric. Then, we use the same metric to retrieve suitable prompts for ChatGPT from a small annotated validation set in the same domain. The generated dataset is then filtered to remove low-quality instances. Our proposed select-prompt-filter KD approach leads to significant improvements of up to 6.6 ROUGE-2 score by leveraging sufficient in-domain pseudo-labeled data over a standard KD approach given the same size of training data.

11:00-12:30 (East Foyer / Virtual)
**#55 Leap-of-Thought: Accelerating Transformers via Dynamic Token Routing**
*Yeachan Kim, Junho Kim, Jun-Hyung Park, Mingyu Lee and SangKeun Lee*
Computational inefficiency in transformers has been a long-standing challenge, hindering the deployment in resource-constrained or real-time applications. One promising approach to mitigate this limitation is to progressively remove less significant tokens, given that the sequence length strongly contributes to the inefficiency. However, this approach entails a potential risk of losing crucial information due to the irrevocable nature of token removal. In this paper, we introduce Leap-of-Thought (LoT), a novel token reduction approach that dynamically routes tokens within layers. Unlike previous work that irrevocably discards tokens, LoT enables tokens to 'leap' across layers. This ensures that all tokens remain accessible in subsequent layers while reducing the number of tokens processed within layers. We achieve this by pairing the transformer with dynamic token routers, which learn to selectively process tokens essential for the task. Evaluation results clearly show that LoT achieves a substantial improvement in computational efficiency. Specifically, LoT attains up to 25x faster inference time without a significant loss in accuracy

11:00-12:30 (East Foyer / Virtual)
**#56 Prototype-based HyperAdapter for Sample-Efficient Multi-task Tuning**
*Hao Zhao, Jie Fu and Zhaofeng He*
Parameter-efficient fine-tuning (PEFT) has shown its effectiveness in adapting the pre-trained language models to downstream tasks while only updating a small number of parameters. Despite the success, most existing methods independently adapt to each task without considering knowledge transfer between tasks and are limited to low-data regimes. To overcome this issue, we propose Prototype-based HyperAdapter (PHA), a novel framework built on the adapter-tuning and hypernetwork. It introduces an instance-dense retriever and a prototypical hypernetwork to generate the conditional modules in a sample-efficient manner. This leads to comparable performance improvements against existing PEFT methods on multi-task learning and few-shot transfer learning. More importantly, when the available data size gets smaller, our method outperforms other strong baselines by a large margin. Based on our extensive empirical experiments across various datasets, we demonstrate that PHA strikes a better trade-off between trainable parameters, accuracy on stream tasks, and sample efficiency. Our code is publicly available at https://github.com/Bumble666/PHA

11:00-12:30 (East Foyer / Virtual)
**#57 A Study on Accessing Linguistic Information in Pre-Trained Language Models by Using Prompts**
*Marion Di Marco, Katharina Hämmerl and Alexander Fraser*
We study whether linguistic information in pre-trained multilingual language models can be accessed by human language: So far, there is no easy method to directly obtain linguistic information and gain insights into the linguistic principles encoded in such models. We use the technique of prompting and formulate linguistic tasks to test the LM's access to explicit grammatical principles and study how effective this method is at providing access to linguistic features. Our experiments on German, Icelandic and Spanish show that some linguistic properties can in fact be accessed through prompting, whereas others are harder to capture.

11:00-12:30 (East Foyer / Virtual)
**#58 Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation**
*Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng and Xuanjing Huang*
Large Language Models (LLMs) have made remarkable advancements in the field of natural language generation. However, the propensity of LLMs to generate inaccurate or non-factual content, termed "hallucinations", remains a significant challenge. Current hallucination detection methods often necessitate the retrieval of great numbers of relevant evidence, thereby increasing response times. We introduce a unique framework that leverages statistical decision theory and Bayesian sequential analysis to optimize the trade-off between costs and benefits during the hallucination detection process. This approach does not require a predetermined number of observations. Instead, the analysis proceeds in a sequential manner, enabling an expeditious decision towards "belief" or "disbelief" through a stop-or-continue strategy. Extensive experiments reveal that this novel framework surpasses existing methods in both efficiency and precision of hallucination detection. Furthermore, it requires fewer retrieval steps on average, thus decreasing response times.

11:00-12:30 (East Foyer / Virtual)
**#59 Universal Self-Adaptive Prompting**
*Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Martin Eisenschlos, Sercan O Arik and Tomas Pfister*
A hallmark of modern large language models (LLMs) is their impressive general zero-shot and few-shot abilities, often elicited through in-context learning (ICL) via prompting. However, while highly coveted and being the most general, zero-shot performances in LLMs are still typically weaker due to the lack of guidance and the difficulty of applying existing automatic prompt design methods in general tasks when ground-truth labels are unavailable. In this study, we address this by presenting Universal Self-Adaptive Prompting (USP), an automatic prompt design approach specifically tailored for zero-shot learning (while compatible with few-shot). Requiring only a small amount of unlabeled data and an inference-only LLM, USP is highly versatile: to achieve universal prompting, USP categorizes a possible NLP task into one of the three possible task types and then uses a corresponding selector to select the most suitable queries and zero-shot model-generated responses as pseudo-demonstrations, thereby generalizing ICL to the zero-shot setup in a fully automated way. We evaluate USP with PaLM and PaLM 2 models and demonstrate performances that are considerably stronger than standard zero-shot baselines and often comparable to or even superior to few-shot baselines across more than 40 natural language understanding, natural language generation, and reasoning tasks.

11:00-12:30 (East Foyer / Virtual)
**#60 Accelerating Toeplitz Neural Network with Constant-time Inference Complexity**
*Zhen Qin and Yiran Zhong*
Toeplitz Neural Networks (TNNs) have exhibited outstanding performance in various sequence modeling tasks. They outperform commonly used Transformer-based models while benefiting from log-linear space-time complexities. On the other hand, State Space Models (SSMs) achieve lower performance than TNNs in language modeling but offer the advantage of constant inference complexity. In this paper, we

aim to combine the strengths of TNNs and SSMs by converting TNNs to SSMs during inference, thereby enabling TNNs to achieve the same constant inference complexities as SSMs. To accomplish this, we formulate the conversion process as an optimization problem and provide a closed-form solution. We demonstrate how to transform the target equation into a Vandermonde linear system problem, which can be efficiently solved using the Discrete Fourier Transform (DFT). Notably, our method requires no training and maintains numerical stability. It can be also applied to any LongConv-based model. To assess its effectiveness, we conduct extensive experiments on language modeling tasks across various settings. Additionally, we compare our method to other gradient-descent solutions, highlighting the superior numerical stability of our approach. The source code is available at https://github.com/OpenNLPLab/ETSC-Exact-Toeplitz-to-SSM-Conversion.

11:00-12:30 (East Foyer / Virtual)
**#61 IBADR: an Iterative Bias-Aware Dataset Refinement Framework for Debiasing NLU models**
*Xiaoyue Wang, Xin Liu, Lijie Wang, Yaoxiang Wang, Jinsong Su and Hua Wu*
As commonly-used methods for debiasing natural language understanding (NLU) models, dataset refinement approaches heavily rely on manual data analysis, and thus maybe unable to cover all the potential biased features. In this paper, we propose IBADR, an Iterative Bias-Aware Dataset Refinement framework, which debiases NLU models without predefining biased features. We maintain an iteratively expanded sample pool. Specifically, at each iteration, we first train a shallow model to quantify the bias degree of samples in the pool. Then, we pair each sample with a bias indicator representing its bias degree, and use these extended samples to train a sample generator. In this way, this generator can effectively learn the correspondence relationship between bias indicators and samples. Furthermore, we employ the generator to produce pseudo samples with fewer biased features by feeding specific bias indicators. Finally, we incorporate the generated pseudo samples into the pool. Experimental results and in-depth analyses on two NLU tasks show that IBADR not only significantly outperforms existing dataset refinement approaches, achieving SOTA, but also is compatible with model-centric methods.

11:00-12:30 (East Foyer / Virtual)
**#62 Identifying Statements Crucial for Awareness of Interpretive Nonsense to Prevent Communication Breakdowns**
*Tomoyuki Maekawa and Michita Imai*
During remote conversations, communication breakdowns often occur when a listener misses certain statements. Our objective is to prevent such breakdowns by identifying Statements Crucial for Awareness of Interpretive Nonsense (SCAINs). If a listener misses a SCAIN, s/he may interpret subsequent statements differently from the speaker's intended meaning. To identify SCAINs, we adopt a unique approach where we create a dialogue by omitting two consecutive statements from the original dialogue and then generate text to make the following statement more specific. The novelty of the proposed method lies in simulating missing information by processing text with omissions. We validate the effectiveness of SCAINs through evaluation using a dialogue dataset. Furthermore, we demonstrate that SCAINs cannot be identified as merely important statements, highlighting the uniqueness of our proposed method.

11:00-12:30 (East Foyer / Virtual)
**#63 Outlier Dimensions Encode Task Specific Knowledge**
*William Rudman, Catherine Chen and Carsten Eickhoff*
Representations from large language models (LLMs) are known to be dominated by a small subset of dimensions with exceedingly high variance. Previous works have argued that although ablating these outlier dimensions in LLM representations hurts downstream performance, outlier dimensions are detrimental to the representational quality of embeddings. In this study, we investigate how fine-tuning impacts outlier dimensions and show that 1) outlier dimensions that occur in pre-training persist in fine-tuned models and 2) a single outlier dimension can complete downstream tasks with a minimal error rate. Our results suggest that outlier dimensions can encode crucial task-specific knowledge and that the value of a representation in a single outlier dimension drives downstream model decisions.

11:00-12:30 (East Foyer / Virtual)
**#64 Rather a Nurse than a Physician - Contrastive Explanations under Investigation**
*Oliver Eberle, Ilias Chalkidis, Laura Cabello and Stephanie Brandl*
Contrastive explanations, where one decision is explained *in contrast to another*, are supposed to be closer to how humans explain a decision than non-contrastive explanations, where the decision is not necessarily referenced to an alternative. This claim has never been empirically validated. We analyze four English text-classification datasets (SST2, DynaSent, BIOS and DBpedia-Animals). We fine-tune and extract explanations from three different models (RoBERTa, GTP-2, and T5), each in three different sizes and apply three post-hoc explainability methods (LRP, GradientxInput, GradNorm). We furthermore collect and release human rationale annotations for a subset of 100 samples from the BIOS dataset for contrastive and non-contrastive settings. A cross-comparison between model-based rationales and human annotations, both in contrastive and non-contrastive settings, yields a high agreement between the two settings for models as well as for humans. Moreover, model-based explanations computed in both settings align equally well with human rationales. Thus, we empirically find that humans do not necessarily explain in a contrastive manner.

11:00-12:30 (East Foyer / Virtual)
**#65 Conceptual structure coheres in human cognition but not in large language models**
*Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua and Timothy T. Rogers*
Neural network models of language have long been used as a tool for developing hypotheses about conceptual representation in the mind and brain. For many years, such use involved extracting vector-space representations of words and using distances among these to predict or understand human behavior in various semantic tasks. In contemporary language models, however, it is possible to interrogate the latent structure of conceptual representations using methods nearly identical to those commonly used with human participants. The current work uses three common techniques borrowed from cognitive psychology to estimate and compare lexical-semantic structure in both humans and a well-known large language model, the DaVinci variant of GPT-3. In humans, we show that conceptual structure is robust to differences in culture, language, and method of estimation. Structures estimated from the LLM behavior, while individually fairly consistent with those estimated from human behavior, depend much more upon the particular task used to generate behavior responses–responses generated by the very same model in the three tasks yield estimates of conceptual structure that cohere less with one another than do human structure estimates. The results suggest one important way that knowledge inhering in contemporary LLMs can differ from human cognition.

11:00-12:30 (East Foyer / Virtual)
**#66 BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology**
*Odhran O'Donoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essam Ghareeb and Samuel G Rodriques*
The ability to automatically generate accurate protocols for scientific experiments would represent a major step towards the automation of science. Large Language Models (LLMs) have impressive capabilities on a wide range of tasks, such as question answering and the generation of coherent text and code. However, LLMs can struggle with multi-step problems and long-term planning, which are crucial for designing scientific experiments. Moreover, evaluation of the accuracy of scientific protocols is challenging, because experiments can be described correctly in many different ways, require expert knowledge to evaluate, and cannot usually be executed automatically. Here we present an automatic evaluation framework for the task of planning experimental protocols, and we introduce BioProt: a dataset of biology protocols with corresponding pseudocode representations. To measure performance on generating scientific protocols, we use an LLM to convert a

natural language protocol into pseudocode, and then evaluate an LLM's ability to reconstruct the pseudocode from a high-level description and a list of admissible pseudocode functions. We evaluate GPT-3 and GPT-4 on this task and explore their robustness. We externally validate the utility of pseudocode representations of text by generating accurate novel protocols using retrieved pseudocode, and we run a generated protocol successfully in our biological laboratory. Our framework is extensible to the evaluation and improvement of language model

11:00-12:30 (East Foyer / Virtual)
**#67 Plan, Verify and Switch: Integrated Reasoning with Diverse X-of-Thoughts**
*Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu and Zheng Zhang*
As large language models (LLMs) have shown effectiveness with different prompting methods, such as Chain of Thought, Program of Thought, we find that these methods have formed a great complementarity to each other on math reasoning tasks. In this work, we propose XoT, an integrated problem solving framework by prompting LLMs with diverse reasoning thoughts. For each question, XoT always begins with selecting the most suitable method then executes each method iteratively. Within each iteration, XoT actively checks the validity of the generated answer and incorporates the feedback from external executors, allowing it to dynamically switch among different prompting methods. Through extensive experiments on 10 popular math reasoning datasets, we demonstrate the effectiveness of our proposed approach and thoroughly analyze the strengths of each module. Moreover, empirical results suggest that our framework is orthogonal to recent work that makes improvements on single reasoning methods and can further generalise to logical reasoning domain. By allowing method switching, XoT provides a fresh perspective on the collaborative integration of diverse reasoning thoughts in a unified framework.

11:00-12:30 (East Foyer / Virtual)
**#68 OpenAsp: A Benchmark for Multi-document Open Aspect-based Summarization**
*Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira and Ido Dagan*
The performance of automatic summarization models has improved dramatically in recent years. Yet, there is still a gap in meeting specific information needs of users in real-world scenarios, particularly when a targeted summary is sought, such as in the useful aspect-based summarization setting targeted in this paper. Previous datasets and studies for this setting have predominantly concentrated on a limited set of pre-defined aspects, focused solely on single document inputs, or relied on synthetic data. To advance research on more realistic scenarios, we introduce OpenAsp, a benchmark for multi-document open aspect-based summarization. This benchmark is created using a novel and cost-effective annotation protocol, by which an open aspect dataset is derived from existing generic multi-document summarization datasets. We analyze the properties of OpenAsp showcasing its high-quality content. Further, we show that the realistic open-aspect setting realized in OpenAsp poses a challenge for current state-of-the-art summarization models, as well as for large language models.

11:00-12:30 (East Foyer / Virtual)
**#69 Generative Adversarial Training with Perturbed Token Detection for Model Robustness**
*Jiahao Zhao and Wenji Mao*
Adversarial training is the dominant strategy towards model robustness. Current adversarial training methods typically apply perturbations to embedding representations, whereas actual text-based attacks introduce perturbations as discrete tokens. Thus there exists a gap between the continuous embedding representations and discrete text tokens that hampers the effectiveness of adversarial training. Moreover, the continuous representations of perturbations cannot be further utilized, resulting in the suboptimal performance. To bridge this gap for adversarial robustness, in this paper, we devise a novel generative adversarial training framework that integrates gradient-based learning, adversarial example generation and perturbed token detection. Our proposed framework consists of generative adversarial attack and adversarial training process. Specifically, in generative adversarial attack, the embeddings are shared between the classifier and the generative model, which enables the generative model to leverage the gradients from the classifier for generating perturbed tokens. Then, adversarial training process combines adversarial regularization with perturbed token detection to provide token-level supervision and improve the efficiency of sample utilization. Extensive experiments on five datasets from the AdvGLUE benchmark demonstrate that our framework significantly enhances the model robustness, surpassing the state-of-the-art results of ChatGPT by 10% in average accuracy.

11:00-12:30 (East Foyer / Virtual)
**#70 MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter**
*Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang and Tat-Seng Chua*
Language Models (LMs) have demonstrated impressive molecule understanding ability on various 1D text-related tasks. However, they inherently lack 2D graph perception — a critical ability of human professionals in comprehending molecules' topological structures. To bridge this gap, we propose MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. MolCA enables an LM (i.e., Galactica) to understand both text- and graph-based molecular contents via the cross-modal projector. Specifically, the cross-modal projector is implemented as a Q-Former to connect a graph encoder's representation space and an LM's text space. Further, MolCA employs a uni-modal adapter (i.e., LoRA) for the LM's efficient adaptation to downstream tasks. Unlike previous studies that couple an LM with a graph encoder via cross-modal contrastive learning, MolCA retains the LM's ability of open-ended text generation and augments it with 2D graph information. To showcase its effectiveness, we extensively benchmark MolCA on tasks of molecule captioning, IUPAC name prediction, and molecule-text retrieval, on which MolCA significantly outperforms the baselines.

11:00-12:30 (East Foyer / Virtual)
**#71 Polyglot or Not? Measuring Multilingual Encyclopedic Knowledge in Foundation Models**
*Tim Schott, Daniel Ryan Furman and Shreshta Bhat*
In this work, we assess the ability of foundation models to recall encyclopedic knowledge across a wide range of linguistic contexts. To support this, we: 1) produce a 20-language dataset that contains 303k factual associations paired with counterfactuals, 2) evaluate 5 models in a multilingual test, and 3) benchmark a diverse set of 24 models in an English-only test. Meta's LLaMA achieves the highest scores in both multilingual and English-only evaluations. Yet, an analysis of LLaMA's errors reveals significant limitations in its ability to recall facts in languages other than English, plus difficulties related to the location and gender of fact subjects. Overall, our findings suggest that today's foundation models are far from polyglots.

11:00-12:30 (East Foyer / Virtual)
**#72 Detecting and Mitigating Hallucinations in Multilingual Summarisation**
*Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti and Shay B. Cohen*
Hallucinations pose a significant challenge to the reliability of neural models for abstractive summarisation. While automatically generated summaries may be fluent, they often lack faithfulness to the original document. This issue becomes even more pronounced in low-resource languages, where summarisation requires cross-lingual transfer. With the existing faithful metrics focusing on English, even measuring the extent of this phenomenon in cross-lingual settings is hard. To address this, we first develop a novel metric, mFACT, evaluating the faithfulness of non-English summaries, leveraging translation-based transfer from multiple English faithfulness metrics. Through extensive experiments in multiple languages, we demonstrate that mFACT is best suited to detect hallucinations compared to alternative metrics. With mFACT, we assess a broad range of multilingual large language models, and find that they all tend to hallucinate often in languages different from English. We then propose a simple but effective method to reduce hallucinations in cross-lingual transfer, which weighs the loss of each

training example by its faithfulness score. This method drastically increases both performance and faithfulness according to both automatic and human evaluation when compared to strong baselines for cross-lingual transfer such as MAD-X. Our code and dataset are available at https://github.com/yfqiu-nlp/mfact-summ.

11:00-12:30 (East Foyer / Virtual)
**#73 FaMeSumm: Investigating and Improving Faithfulness of Medical Summarization**
*Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra and Rui Zhang*
Summaries of medical text shall be faithful by being consistent and factual with source inputs, which is an important but understudied topic for safety and efficiency in healthcare. In this paper, we investigate and improve faithfulness in summarization on a broad range of medical summarization tasks. Our investigation reveals that current summarization models often produce unfaithful outputs for medical input text. We then introduce FaMeSumm, a framework to improve faithfulness by fine-tuning pre-trained language models based on medical knowledge. FaMeSumm performs contrastive learning on designed sets of faithful and unfaithful summaries, and it incorporates medical terms and their contexts to encourage faithful generation of medical terms. We conduct comprehensive experiments on three datasets in two languages: health question and radiology report summarization datasets in English, and a patient-doctor dialogue dataset in Chinese. Results demonstrate that FaMeSumm is flexible and effective by delivering consistent improvements over mainstream language models such as BART, T5, mT5, and PEGASUS, yielding state-of-the-art performances on metrics for faithfulness and general quality. Human evaluation by doctors also shows that FaMeSumm generates more faithful outputs. Our code is available at https://github.com/psunlpgroup/FaMeSumm.

11:00-12:30 (East Foyer / Virtual)
**#74 Dynamic Top-k Estimation Consolidates Disagreement between Feature Attribution Methods**
*Jonathan Kamp, Lisa Beinborn and Antske Fokkens*
Feature attribution scores are used for explaining the prediction of a text classifier to users by highlighting a k number of tokens. In this work, we propose a way to determine the number of optimal k tokens that should be displayed from sequential properties of the attribution scores. Our approach is dynamic across sentences, method-agnostic, and deals with sentence length bias. We compare agreement between multiple methods and humans on an NLI task, using fixed k and dynamic k. We find that perturbation-based methods and Vanilla Gradient exhibit highest agreement on most method–method and method–human agreement metrics with a static k. Their advantage over other methods disappears with dynamic ks which mainly improve Integrated Gradient and GradientXInput. To our knowledge, this is the first evidence that sequential properties of attribution scores are informative for consolidating attribution signals for human interpretation.

11:00-12:30 (East Foyer / Virtual)
**#75 An Attribution Method for Siamese Encoders**
*Lucas Moeller, Dmitry Nikolaev and Sebastian Padó*
Despite the success of Siamese encoder models such as sentence transformers (ST), little is known about the aspects of inputs they pay attention to. A barrier is that their predictions cannot be attributed to individual features, as they compare two inputs rather than processing a single one. This paper derives a local attribution method for Siamese encoders by generalizing the principle of integrated gradients to models with multiple inputs. The output takes the form of feature-pair attributions and in case of STs it can be reduced to a token–token matrix. Our method involves the introduction of integrated Jacobians and inherits the advantageous formal properties of integrated gradients: it accounts for the model's full computation graph and is guaranteed to converge to the actual prediction. A pilot study shows that in case of STs few token pairs can dominate predictions and that STs preferentially focus on nouns and verbs. For accurate predictions, however, they need to attend to the majority of tokens and parts of speech.

11:00-12:30 (East Foyer / Virtual)
**#76 Regulation and NLP (RegNLP): Taming Large Language Models**
*Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordás and Gerasimos Spanakis*
The scientific innovation in Natural Language Processing (NLP) and more broadly in artificial intelligence (AI) is at its fastest pace to date. As large language models (LLMs) unleash a new era of automation, important debates emerge regarding the benefits and risks of their development, deployment and use. Currently, these debates have been dominated by often polarized narratives mainly led by the AI Safety and AI Ethics movements. This polarization, often amplified by social media, is swaying political agendas on AI regulation and governance and posing issues of regulatory capture. Capture occurs when the regulator advances the interests of the industry it is supposed to regulate, or of special interest groups rather than pursuing the general public interest. Meanwhile in NLP research, attention has been increasingly paid to the discussion of regulating risks and harms. This often happens without systematic methodologies or sufficient rooting in the disciplines that inspire an extended scope of NLP research, jeopardizing the scientific integrity of these endeavors. Regulation studies are a rich source of knowledge on how to systematically deal with risk and uncertainty, as well as with scientific evidence, to evaluate and compare regulatory options. This resource has largely remained untapped so far. In this paper, we argue how NLP research on these topics can benefit from proximity to regulatory studies and adjacent fields. We do so by discussing basic tenets of regulation, and risk and uncertainty, and by highlighting the shortcomings of current NLP discussions dealing with risk assessment. Finally, we advocate for the development of a new multidisciplinary research space on regulation and NLP (RegNLP), focused on connecting scientific knowledge to regulatory processes based on systematic methodologies.

11:00-12:30 (East Foyer / Virtual)
**#77 Self-ICL: Zero-Shot In-Context Learning with Self-Generated Demonstrations**
*Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen and Hsin-Hsi Chen*
Large language models (LLMs) have exhibited striking in-context learning (ICL) ability to adapt to target tasks with a few input-output demonstrations. For better ICL, different methods are proposed to select representative demonstrations from existing training corpora. However, such settings are not aligned with real-world practices, as end-users usually query LMs without access to demonstration pools. In this work, we introduce Self-ICL—a simple framework which bootstraps LMs' intrinsic capabilities to perform zero-shot ICL. Given a test input, Self-ICL first prompts the model to generate pseudo-inputs. Next, the model predicts pseudo-labels for the pseudo-inputs via zero-shot prompting. Finally, we perform ICL for the test input with the pseudo-input-label pairs as demonstrations. Evaluation on 23 BIG-Bench Hard tasks shows Self-ICL outperforms zero-shot baselines on both average accuracy and head-to-head comparison. Moreover, with zero-shot chain-of-thought, Self-ICL achieves results comparable to using real demonstrations. Additionally, we conduct a range of analyses to validate Self-ICL's effectiveness and provide insights for its behaviors under different settings.

11:00-12:30 (East Foyer / Virtual)
**#78 The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning**
*Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin and Minjoon Seo*
Language models (LMs) with less than 100B parameters are known to perform poorly on chain-of-thought (CoT) reasoning in contrast to large LMs when solving unseen tasks. In this work, we aim to equip smaller LMs with the step-by-step reasoning capability by instruction tuning with CoT rationales. In order to achieve this goal, we first introduce a new instruction-tuning dataset called the CoT Collection, which augments the existing Flan Collection (including only 9 CoT tasks) with additional 1.84 million rationales across 1,060 tasks. We show that

CoT fine-tuning Flan-T5 (3B & 11B) with CoT Collection enables smaller LMs to have better CoT capabilities on unseen tasks. On the BIG-Bench-Hard (BBH) benchmark, we report an average improvement of +4.34% (Flan-T5 3B) and +2.60% (Flan-T5 11B), in terms of zero-shot task accuracy. Furthermore, we show that instruction tuning with CoT Collection allows LMs to possess stronger few-shot learning capabilities on 4 domain-specific tasks, resulting in an improvement of +2.24% (Flan-T5 3B) and +2.37% (Flan-T5 11B), even outperforming ChatGPT utilizing demonstrations until the max length by a +13.98% margin. Our code, the CoT Collection data, and model checkpoints are publicly available.

11:00-12:30 (East Foyer / Virtual)
**#79 Cross-Modal Conceptualization in Bottleneck Models**
*Danis Alukaev, Semen Kiselev, Ilya Pershin, Bulat Ibragimov, Vladimir V. Ivanov, Alexey Kornaev and Ivan Titov*
Concept Bottleneck Models (CBMs) assume that training examples (e.g., x-ray images) are annotated with high-level concepts (e.g., types of abnormalities), and perform classification by first predicting the concepts, followed by predicting the label relying on these concepts. However, the primary challenge in employing CBMs lies in the requirement of defining concepts predictive of the label and annotating training examples with these concepts. In our approach, we adopt a more moderate assumption and instead use text descriptions (e.g., radiology reports), accompanying the images, to guide the induction of concepts. Our crossmodal approach treats concepts as discrete latent variables and promotes concepts that (1) are predictive of the label, and (2) can be predicted reliably from both the image and text. Through experiments conducted on datasets ranging from synthetic datasets (e.g., synthetic images with generated descriptions) to realistic medical imaging datasets, we demonstrate that crossmodal learning encourages the induction of interpretable concepts while also facilitating disentanglement.

11:00-12:30 (East Foyer / Virtual)
**#80 Reconstruct Before Summarize: An Efficient Two-Step Framework for Condensing and Summarizing Meeting Transcripts**
*Haochen Tan, Han Wu, Wei Shao, Xinyun Zhang, Mingjie Zhan, Zhaohui Hou, Ding Liang and Linqi Song*
Meetings typically involve multiple participants and lengthy conversations, resulting in redundant and trivial content. To overcome these challenges, we propose a two-step framework, Reconstruct before Summarize (RbS), for effective and efficient meeting summarization. RbS first leverages a self-supervised paradigm to annotate essential contents by reconstructing the meeting transcripts. Secondly, we propose a relative positional bucketing (RPB) algorithm to equip (conventional) summarization models to generate the summary. Despite the additional reconstruction process, our proposed RPB significantly compresses the input, leading to faster processing and reduced memory consumption compared to traditional summarization methods. We validate the effectiveness and efficiency of our method through extensive evaluations and analyses. On two meeting summarization datasets, AMI and ICSI, our approach outperforms previous state-of-the-art approaches without relying on large-scale pre-training or expert-grade annotating tools.

11:00-12:30 (East Foyer / Virtual)
**#81 Language Models with Rationality**
*Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze and Peter Clark*
While large language models (LLMs) are proficient at question-answering (QA), it is not always clear how (or even if) an answer follows from their latent "beliefs". This lack of interpretability is a growing impediment to widespread use of LLMs. To address this, our goals are to make model beliefs and their inferential relationships explicit, and to resolve inconsistencies that may exist, so that answers are supported by interpretable chains of reasoning drawn from a consistent network of beliefs. Our approach, which we call REFLEX, is to add a **rational, self-reflecting layer** on top of the LLM. First, given a question, we construct a **belief graph** using a backward-chaining process to materialize relevant model beliefs (including beliefs about answer candidates) and their inferential relationships. Second, we identify and minimize contradictions in that graph using a formal constraint reasoner. We find that REFLEX significantly improves consistency (by 8%-11% absolute) without harming overall answer accuracy, resulting in answers supported by faithful chains of reasoning drawn from a more consistent belief system. This suggests a new style of system architecture in which an LLM extended with a rational layer can provide an interpretable window into system beliefs, add a systematic reasoning capability, and repair latent inconsistencies present in the LLM.

11:00-12:30 (East Foyer / Virtual)
**#82 Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models**
*Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut and Mrinmaya Sachan*
Recent work has shown that language models (LMs) have strong multi-step (i.e., procedural) reasoning capabilities. However, it is unclear whether LMs perform these tasks by cheating with answers memorized from pretraining corpus, or, via a multi-step reasoning mechanism. In this paper, we try to answer this question by exploring a mechanistic interpretation of LMs for multi-step reasoning tasks. Concretely, we hypothesize that the LM implicitly embeds a reasoning tree resembling the correct reasoning process within it. We test this hypothesis by introducing a new probing approach (called MechanisticProbe) that recovers the reasoning tree from the model's attention patterns. We use our probe to analyze two LMs: GPT-2 on a synthetic task (k-th smallest element), and LLaMA on two simple language-based reasoning tasks (ProofWriter & AI2 Reasoning Challenge). We show that MechanisticProbe is able to detect the information of the reasoning tree from the model's attentions for most examples, suggesting that the LM indeed is going through a process of multi-step reasoning within its architecture in many cases.

11:00-12:30 (East Foyer / Virtual)
**#83 Improving Diversity of Demographic Representation in Large Language Models via Collective-Critiques and Self-Voting**
*Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel and Jilin Chen*
A crucial challenge for generative large language models (LLMs) is diversity: when a user's prompt is under-specified, models may follow implicit assumptions while generating a response, which may result in homogenization of the responses, as well as certain demographic groups being under-represented or even erased from the generated responses. In this paper, we formalize the problem diversity of representation in LLM generations. We present evaluation datasets and propose metrics to measure diversity in generated responses along people and culture axes. We find that LLMs understand the notion of diversity, and that they can reason and critique their own responses for that goal. This finding motivated a new prompting technique called collective-critique and self-voting (CCSV) to self-improve people diversity of LLMs by tapping into its diversity reasoning capabilities, without relying on handcrafted examples or prompt tuning. Extensive empirical experiments with both human and automated evaluations show that our proposed approach is effective at improving people and culture diversity, and outperforms all baseline methods by a large margin.

11:00-12:30 (East Foyer / Virtual)
**#84 Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness?**
*Kevin Liu, Stephen Casper, Dylan Hadfield-Menell and Jacob Andreas*
Neural language models (LMs) can be used to evaluate the truth of factual statements in two ways: they can be either queried for statement probabilities, or probed for internal representations of truthfulness. Past work has found that these two procedures sometimes disagree, and that probes tend to be more accurate than LM outputs. This has led some researchers to conclude that LMs "lie" or otherwise encode non-cooperative communicative intents. Is this an accurate description of today's LMs, or can query–probe disagreement arise in other ways? We

identify three different classes of disagreement, which we term confabulation, deception, and heterogeneity. In many cases, the superiority of probes is simply attributable to better calibration on uncertain answers rather than a greater fraction of correct, high-confidence answers. In some cases, queries and probes perform better on different subsets of inputs, and accuracy can further be improved by ensembling the two.

11:00-12:30 (East Foyer / Virtual)
**#85 What's "up" with vision-language models? Investigating their struggle with spatial reasoning**
*Amita Kamath, Jack Hessel and Kai-Wei Chang*
Recent vision-language (VL) models are powerful, but can they reliably distinguish "right" from "left"? We curate three new corpora to quantify model comprehension of such basic spatial relations. These tests isolate spatial reasoning more precisely than existing datasets like VQAv2, e.g., our What'sUp benchmark contains sets of photographs varying only the spatial relations of objects, keeping their identity fixed (see Figure 1: models must comprehend not only the usual case of a dog under a table, but also, the same dog on top of the same table). We evaluate 18 VL models, finding that all perform poorly, e.g., BLIP finetuned on VQAv2, which nears human parity on VQAv2, achieves 56% accuracy on our benchmarks vs. humans at 99%. We conclude by studying causes of this surprising behavior, finding: 1) that popular vision-language pretraining corpora like LAION-2B contain little reliable data for learning spatial relationships; and 2) that basic modeling interventions like up-weighting preposition-containing instances or fine-tuning on our corpora are not sufficient to address the challenges our benchmarks pose. We are hopeful that these corpora will facilitate further research, and we release our data and code at https://github.com/amitakamath/whatsup_vlms.

11:00-12:30 (East Foyer / Virtual)
**#86 When Language Models Fall in Love: Animacy Processing in Transformer Language Models**
*Michael Hanna, Yonatan Belinkov and Sandro Pezzelle*
Animacy—whether an entity is alive and sentient—is fundamental to cognitive processing, impacting areas such as memory, vision, and language. However, animacy is not always expressed directly in language: in English it often manifests indirectly, in the form of selectional constraints on verbs and adjectives. This poses a potential issue for transformer language models (LMs): they often train only on text, and thus lack access to extralinguistic information from which humans learn about animacy. We ask: how does this impact LMs' animacy processing—do they still behave as humans do? We answer this question using open-source LMs. Like previous studies, we find that LMs behave much like humans when presented with entities whose animacy is typical. However, we also show that even when presented with stories about atypically animate entities, such as a peanut in love, LMs adapt: they treat these entities as animate, though they do not adapt as well as humans. Even when the context indicating atypical animacy is very short, LMs pick up on subtle clues and change their behavior. We conclude that despite the limited signal through which LMs can learn about animacy, they are indeed sensitive to the relevant lexical semantic nuances available in English.

11:00-12:30 (East Foyer / Virtual)
**#87 StereoMap: Quantifying the Awareness of Human-like Stereotypes in Large Language Models**
*Sullam Jeoung, Yubin Ge and Jana Diesner*
Large Language Models (LLMs) have been observed to encode and perpetuate harmful associations present in the training data. We propose a theoretically grounded framework called StereoMap to gain insights into their perceptions of how demographic groups have been viewed by society. The framework is grounded in the Stereotype Content Model (SCM); a well-established theory from psychology. According to SCM, stereotypes are not all alike. Instead, the dimensions of Warmth and Competence serve as the factors that delineate the nature of stereotypes. Based on the SCM theory, StereoMap maps LLMs' perceptions of social groups (defined by socio-demographic features) using the dimensions of Warmth and Competence. Furthermore, the framework enables the investigation of keywords and verbalizations of reasoning of LLMs' judgments to uncover underlying factors influencing their perceptions. Our results show that LLMs exhibit a diverse range of perceptions towards these groups, characterized by mixed evaluations along the dimensions of Warmth and Competence. Furthermore, analyzing the reasonings of LLMs, our findings indicate that LLMs demonstrate an awareness of social disparities, often stating statistical data and research findings to support their reasoning. This study contributes to the understanding of how LLMs perceive and represent social groups, shedding light on their potential biases and the perpetuation of harmful associations.

11:00-12:30 (East Foyer / Virtual)
**#88 "Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms**
*Kunal Handa, Margarett Clapper, Jessica Boyle, Rose E Wang, Diyi Yang, David Yeager and Dorottya Demszky*
Teachers' growth mindset supportive language (GMSL)—rhetoric emphasizing that one's skills can be improved over time—has been shown to significantly reduce disparities in academic achievement and enhance students' learning outcomes. Although teachers espouse growth mindset principles, most find it difficult to adopt GMSL in their practice due the lack of effective coaching in this area. We explore whether large language models (LLMs) can provide automated, personalized coaching to support teachers' use of GMSL. We establish an effective coaching tool to reframe unsupportive utterances to GMSL by developing (i) a parallel dataset containing GMSL-trained teacher reframings of unsupportive statements with an accompanying annotation guide, (ii) a GMSL prompt framework to revise teachers' unsupportive language, and (iii) an evaluation framework grounded in psychological theory for evaluating GMSL with the help of students and teachers. We conduct a large-scale evaluation involving 174 teachers and 1,006 students, finding that both teachers and students perceive GMSL-trained teacher and model reframings as more effective in fostering a growth mindset and promoting challenge-seeking behavior, among other benefits. We also find that model-generated reframings outperform those from the GMSL-trained teachers. These results show promise for harnessing LLMs to provide automated GMSL feedback for teachers and, more broadly, LLMs' potentiality for supporting students' learning in the classroom. Our findings also demonstrate the benefit of large-scale human evaluations when applying LLMs in educational domains.

11:00-12:30 (East Foyer / Virtual)
**#89 Improving Summarization with Human Edits**
*Zonghai Yao, Benjamin J Schloss and Sai P Selvaraj*
Recent work has shown the promise of learning with human feedback paradigms to produce human-determined high-quality text. Existing works use human feedback to train large language models (LLMs) in general domain abstractive summarization and have obtained summary quality exceeding traditional likelihood training. In this paper, we focus on a less explored form of human feedback – Human Edits. We propose Sequence Alignment (un)Likelihood Training (SALT), a novel technique to use both the human-edited and model-generated data together in the training loop. In addition, we demonstrate simulating Human Edits with ground truth summaries coming from existing training data – Imitation edits, along with the model-generated summaries obtained after the training, to reduce the need for expensive human-edit data. In our experiments, we extend human feedback exploration from general domain summarization to medical domain summarization. Our results demonstrate the effectiveness of SALT in improving the summary quality with Human and Imitation Edits. Through additional experiments, we show that SALT outperforms the conventional RLHF method (designed for human preferences) – DPO, when applied to human-edit data. We hope the evidence in our paper prompts researchers to explore, collect, and better use different human feedback approaches scalably.

11:00-12:30 (East Foyer / Virtual)

**#90 Theory of Mind for Multi-Agent Collaboration via Large Language Models**
*Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Michael Lewis and Katia P. Sycara*
While Large Language Models (LLMs) have demonstrated impressive accomplishments in both reasoning and planning, their abilities in multi-agent collaborations remains largely unexplored. This study evaluates LLM-based agents in a multi-agent cooperative text game with Theory of Mind (ToM) inference tasks, comparing their performance with Multi-Agent Reinforcement Learning (MARL) and planning-based baselines. We observed evidence of emergent collaborative behaviors and high-order Theory of Mind capabilities among LLM-based agents. Our results reveal limitations in LLM-based agents' planning optimization due to systematic failures in managing long-horizon contexts and hallucination about the task state. We explore the use of explicit belief state representations to mitigate these issues, finding that it enhances task performance and the accuracy of ToM inferences for LLM-based agents.

11:00-12:30 (East Foyer / Virtual)
**#91 Characterizing Mechanisms for Factual Recall in Language Models**
*Qinan Yu, Jack Merullo and Ellie Pavlick*
Language Models (LMs) often must integrate facts they memorized in pretraining with new information that appears in a given context. These two sources can disagree, causing competition within the model, and it is unclear how an LM will resolve the conflict. On a dataset that queries for knowledge of world capitals, we investigate both distributional and mechanistic determinants of LM behavior in such situations. Specifically, we measure the proportion of the time an LM will use a counterfactual prefix (e.g., "The capital of Poland is London") to overwrite what it learned in pretraining ("Warsaw"). On Pythia and GPT2, the training frequency of both the query country ("Poland") and the in-context city ("London") highly affect the models' likelihood of using the counterfactual. We then use head attribution to identify individual attention heads that either promote the memorized answer or the in-context answer in the logits. By scaling up or down the value vector of these heads, we can control the likelihood of using the in-context answer on new data. This method can increase the rate of generating the in-context answer to 88% of the time simply by scaling a single head at runtime. Our work contributes to a body of evidence showing that we can often localize model behaviors to specific components and provides a proof of concept for how future methods might control model behavior dynamically at runtime.

11:00-12:30 (East Foyer / Virtual)
**#92 A Predictive Factor Analysis of Social Biases and Task-Performance in Pretrained Masked Language Models**
*Yi Zhou, Jose Camacho-Collados and Danushka Bollegala*
Various types of social biases have been reported with pretrained Masked Language Models (MLMs) in prior work. However, multiple underlying factors are associated with an MLM such as its model size, size of the training data, training objectives, the domain from which pretraining data is sampled, tokenization, and languages present in the pretrained corpora, to name a few. It remains unclear as to which of those factors influence social biases that are learned by MLMs. To study the relationship between model factors and the social biases learned by an MLM, as well as the downstream task performance of the model, we conduct a comprehensive study over 39 pretrained MLMs covering different model sizes, training objectives, tokenization methods, training data domains and languages. Our results shed light on important factors often neglected in prior literature, such as tokenization or model objectives.

11:00-12:30 (East Foyer / Virtual)
**#93 Enabling Large Language Models to Generate Text with Citations**
*Tianyu Gao, Howard Yen, Jiatong Yu and Danqi Chen*
Large language models (LLMs) have emerged as a widely-used tool for information seeking, but their generated outputs are prone to hallucination. In this work, our aim is to allow LLMs to generate text with citations, improving their factual correctness and verifiability. Existing work mainly relies on commercial search engines and human evaluation, making it challenging to reproduce and compare different modeling approaches. We propose ALCE, the first benchmark for Automatic LLMs' Citation Evaluation. ALCE collects a diverse set of questions and retrieval corpora and requires building end-to-end systems to retrieve supporting evidence and generate answers with citations. We develop automatic metrics along three dimensions—fluency, correctness, and citation quality—and demonstrate their strong correlation with human judgements. Our experiments with state-of-the-art LLMs and novel prompting strategies show that current systems have considerable room for improvement—For example, on the ELI5 dataset, even the best models lack complete citation support 50% of the time. Our analyses further highlight promising future directions, including developing better retrievers, advancing long-context LLMs, and improving the ability to synthesize information from multiple sources.

11:00-12:30 (East Foyer / Virtual)
**#94 EntSUMv2: Dataset, Models and Evaluation for More Abstractive Entity-Centric Summarization**
*Dhruv Mehra, Lingjue Xie, Ella Hofmann-Coyle, Mayank Kulkarni and Daniel Preotiuc-Pietro*
Entity-centric summarization is a form of controllable summarization that aims to generate a summary for a specific entity given a document. Concise summaries are valuable in various real-life applications, as they enable users to quickly grasp the main points of the document focusing on an entity of interest. This paper presents ENTSUMV2, a more abstractive version of the original entity-centric ENTSUM summarization dataset. In ENTSUMV2 the annotated summaries are intentionally made shorter to benefit more specific and useful entity-centric summaries for downstream users. We conduct extensive experiments on this dataset using multiple abstractive summarization approaches that employ supervised fine-tuning or large-scale instruction tuning. Additionally, we perform comprehensive human evaluation that incorporates metrics for measuring crucial facets. These metrics provide a more fine-grained interpretation of the current state-of-the-art systems and highlight areas for future improvement.

11:00-12:30 (East Foyer / Virtual)
**#95 Hyperpolyglot LLMs: Cross-Lingual Interpretability in Token Embeddings**
*Andrea W Wen-Yi and David Mimno*
Cross-lingual transfer learning is an important property of multilingual large language models (LLMs). But how do LLMs represent relationships between languages? Every language model has an input layer that maps tokens to vectors. This ubiquitous layer of language models is often overlooked. We find that similarities between these input embeddings are highly interpretable and that the geometry of these embeddings differs between model families. In one case (XLM-RoBERTa), embeddings encode language: tokens in different writing systems can be linearly separated with an average of 99.2% accuracy. Another family (mT5) represents cross-lingual semantic similarity: the 50 nearest neighbors for any token represent an average of 7.61 writing systems, and are frequently translations. This result is surprising given that there is no explicit parallel cross-lingual training corpora and no explicit incentive for translations in pre-training objectives. Our research opens the door for investigations in 1) The effect of pre-training and model architectures on representations of languages and 2) The applications of cross-lingual representations embedded in language models.

11:00-12:30 (East Foyer / Virtual)
**#96 A State-Vector Framework for Dataset Effects**
*Esmat Sahak, Zining Zhu and Frank Rudzicz*
The impressive success of recent deep neural network (DNN)-based systems is significantly influenced by the high-quality datasets used in

training. However, the effects of the datasets, especially how they interact with each other, remain underexplored. We propose a state-vector framework to enable rigorous studies in this direction. This framework uses idealized probing test results as the bases of a vector space. This framework allows us to quantify the effects of both standalone and interacting datasets. We show that the significant effects of some commonly-used language understanding datasets are characteristic and are concentrated on a few linguistic dimensions. Additionally, we observe some "spill-over" effects: the datasets could impact the models along dimensions that may seem unrelated to the intended tasks. Our state-vector framework paves the way for a systematic understanding of the dataset effects, a crucial component in responsible and robust model development.

11:00-12:30 (East Foyer / Virtual)
**#97 Deep Natural Language Feature Learning for Interpretable Prediction**
*Felipe Urrutia, Cristian Buc Calderon and Valentin Barriere*
We propose a general method to break down a main complex task into a set of intermediary easier sub-tasks, which are formulated in natural language as binary questions related to the final target task. Our method allows for representing each example by a vector consisting of the answers to these questions. We call this representation Natural Language Learned Features (NLLF). NLLF is generated by a small transformer language model (e.g., BERT) that has been trained in a Natural Language Inference (NLI) fashion, using weak labels automatically obtained from a Large Language Model (LLM). We show that the LLM normally struggles for the main task using in-context learning, but can handle these easiest subtasks and produce useful weak labels to train a BERT. The NLI-like training of the BERT allows for tackling zero-shot inference with any binary question, and not necessarily the ones seen during the training. We show that this NLLF vector not only helps to reach better performances by enhancing any classifier, but that it can be used as input of an easy-to-interpret machine learning model like a decision tree. This decision tree is interpretable but also reaches high performances, surpassing those of a pre-trained transformer in some cases. We have successfully applied this method to two completely different tasks: detecting incoherence in students' answers to open-ended mathematics exam questions, and screening abstracts for a systematic literature review of scientific papers on climate change and agroecology.

11:00-12:30 (East Foyer / Virtual)
**#98 The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions**
*Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji and Jiawei Han*
Recent progress in Large Language Models (LLMs) has produced models that exhibit remarkable performance across a variety of NLP tasks. However, it remains unclear whether the existing focus of NLP research accurately captures the genuine requirements of human users. This paper provides a comprehensive analysis of the divergence between academic research in NLP and the needs of real-world NLP applications via a large-scale collection of user-GPT conversations. We analyze a large-scale collection of real user queries to GPT. We compare these queries against existing NLP benchmark tasks and identify a significant gap between the tasks that users frequently request from LLMs and the tasks that are commonly studied in academic research. For example, we find that tasks such as "design" and "planning" are prevalent in user interactions but largely neglected or different from traditional NLP benchmarks. We investigate these overlooked tasks, dissect the practical challenges, and provide insights toward a roadmap to make LLMs better aligned with user needs.

11:00-12:30 (East Foyer / Virtual)
**#99 Evaluation of African American Language Bias in Natural Language Generation**
*Nicholas Deas, Jessica A Grieser, Shana Kleiner, Desmond U. Patton, Elsbeth Turcan and Kathleen McKeown*
While biases disadvantaging African American Language (AAL) have been uncovered in models for tasks such as speech recognition and toxicity detection, there has been little investigation of these biases for language generation models like ChatGPT. We evaluate how well LLMs understand AAL in comparison to White Mainstream English (WME), the encouraged "standard" form of English taught in American classrooms. We measure large language model performance on two tasks: a counterpart generation task, where a model generates AAL given WME and vice versa, and a masked span prediction (MSP) task, where models predict a phrase hidden from their input. Using a novel dataset of AAL texts from a variety of regions and contexts, we present evidence of dialectal bias for six pre-trained LLMs through performance gaps on these tasks.

11:00-12:30 (East Foyer / Virtual)
**#100 Increasing Probability Mass on Answer Choices Does Not Always Improve Accuracy**
*Sarah Wiegreffe, Matthew Finlayson, Oyvind Tafjord, Peter Clark and Ashish Sabharwal*
When pretrained language models (LMs) are applied to discriminative tasks such as multiple-choice questions, they place probability mass on vocabulary tokens that aren't among the given answer choices. Spreading probability mass across multiple surface forms with identical meaning (such as "bath" and "bathtub") is thought to cause an underestimation of a model's true performance, referred to as the "surface form competition" (SFC) hypothesis. This has motivated the introduction of various probability normalization methods. However, many core questions remain unanswered. How do we measure SFC? Are there direct ways of reducing it, and does doing so improve task performance? We propose a mathematical formalism for SFC which allows us to quantify and bound its impact for the first time. We identify a simple method for reducing it—namely, increasing probability mass on the given answer choices by a) including them in the prompt and b) using in-context learning with even just one example. We show this method eliminates the impact of SFC in the majority of instances. Our experiments on three diverse datasets and six LMs reveal several additional surprising findings. For example, both normalization and prompting methods for reducing SFC can be ineffective or even detrimental to task performance for some LMs. We conclude with practical insights for effectively prompting LMs for multiple-choice tasks.

11:00-12:30 (East Foyer / Virtual)
**#101 A Question Answering Framework for Decontextualizing User-facing Snippets from Scientific Documents**
*Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan and Kyle Lo*
Many real-world applications (e.g., note taking, search) require extracting a sentence or paragraph from a document and showing that snippet to a human outside of the source document. Yet, users may find snippets difficult to understand as they lack context from the original document. In this work, we use language models to rewrite snippets from scientific documents to be read on their own. First, we define the requirements and challenges for this user-facing decontextualization task, such as clarifying where edits occur and handling references to other documents. Second, we propose a framework that decomposes the task into three stages: question generation, question answering, and rewriting. Using this framework, we collect gold decontextualizations from experienced scientific article readers. We then conduct a range of experiments across state-of-the-art commercial and open-source language models to identify how to best provide missing-but-relevant information to models for our task. Finally, we develop QaDecontext, a simple prompting strategy inspired by our framework that improves over end-to-end prompting. We conclude with analysis that finds, while rewriting is easy, question generation and answering remain challenging for today's models.

11:00-12:30 (East Foyer / Virtual)
**#102 The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values**
*Hannah Rose Kirk, Andrew Michael Bean, Bertie Vidgen, Paul Rottger and Scott A. Hale*

Human feedback is increasingly used to steer the behaviours of Large Language Models (LLMs). However, it is unclear how to collect and incorporate feedback in a way that is efficient, effective and unbiased, especially for highly subjective human preferences and values. In this paper, we survey existing approaches for learning from human feedback, drawing on 95 papers primarily from the ACL and arXiv repositories. First, we summarise the past, pre-LLM trends for integrating human feedback into language models. Second, we give an overview of present techniques and practices, as well as the motivations for using feedback; conceptual frameworks for defining values and preferences; and how feedback is collected and from whom. Finally, we encourage a better future of feedback learning in LLMs by raising five unresolved conceptual and practical challenges.

11:00-12:30 (East Foyer / Virtual)
**#103 Background Summarization of Event Timelines**
*Adithya Pratapa, Kevin Small and Markus Dreyer*
Generating concise summaries of news events is a challenging natural language processing task. While journalists often curate timelines to highlight key sub-events, newcomers to a news event face challenges in catching up on its historical context. In this paper, we address this need by introducing the task of background news summarization, which complements each timeline update with a background summary of relevant preceding events. We construct a dataset by merging existing timeline datasets and asking human annotators to write a background summary for each timestep of each news event. We establish strong baseline performance using state-of-the-art summarization systems and propose a query-focused variant to generate background summaries. To evaluate background summary quality, we present a question-answering-based evaluation metric, Background Utility Score (BUS), which measures the percentage of questions about a current event timestep that a background summary answers. Our experiments show the effectiveness of instruction fine-tuned systems such as Flan-T5, in addition to strong zero-shot performance using GPT-3.5.

11:00-12:30 (East Foyer / Virtual)
**#104 Understanding the Inner-workings of Language Models Through Representation Dissimilarity**
*Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu and Henry Kvinge*
As language models are applied to an increasing number of real-world applications, understanding their inner workings has become an important issue in model trust, interpretability, and transparency. In this work we show that representation dissimilarity measures, which are functions that measure the extent to which two model's internal representations differ, can be a valuable tool for gaining insight into the mechanics of language models. Among our insights are: (i) an apparent asymmetry in the internal representations of model using SoLU and GeLU activation functions, (ii) evidence that dissimilarity measures can identify and locate generalization properties of models that are invisible via in-distribution test set performance, and (iii) new evaluations of how language model features vary as width and depth are increased. Our results suggest that dissimilarity measures are a promising set of tools for shedding light on the inner workings of language models.

11:00-12:30 (East Foyer / Virtual)
**#105 Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition**
*Sander V Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Liu Kost, Christopher R Carnahan and Jordan Lee Boyd-Graber*
Large Language Models (LLMs) are increasingly being deployed in interactive contexts that involve direct user engagement, such as chatbots and writing assistants. These deployments are increasingly plagued by prompt injection and jailbreaking (collectively, prompt hacking), in which models are manipulated to ignore their original instructions and instead follow potentially malicious ones. Although widely acknowledged as a significant security threat, there is a dearth of a large-scale resource and quantitative study on prompt hacking. To address this lacuna, we launch a global prompt hacking competition, which allows for free-form human input attacks. We elicit 600K+ adversarial prompts against three state-of-the-art LLMs. We describe the dataset, which empirically verifies that current LLMs can indeed be manipulated via prompt hacking. We also present a comprehensive ontology of the types of adversarial prompts.

11:00-12:30 (East Foyer / Virtual)
**#106 Gender Biases in Automatic Evaluation Metrics for Image Captioning**
*Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz and Nanyun Peng*
Model-based evaluation metrics (e.g., CLIPScore and GPTScore) have demonstrated decent correlations with human judgments in various language generation tasks. However, their impact on fairness remains largely unexplored. It is widely recognized that pretrained models can inadvertently encode societal biases, thus employing these models for evaluation purposes may inadvertently perpetuate and amplify biases. For example, an evaluation metric may favor the caption "a woman is calculating an account book" over "a man is calculating an account book," even if the image only shows male accountants. In this paper, we conduct a systematic study of gender biases in model-based automatic evaluation metrics for image captioning tasks. We start by curating a dataset comprising profession, activity, and object concepts associated with stereotypical gender associations. Then, we demonstrate the negative consequences of using these biased metrics, including the inability to differentiate between biased and unbiased generations, as well as the propagation of biases to generation models through reinforcement learning. Finally, we present a simple and effective way to mitigate the metric bias without hurting the correlations with human judgments. Our dataset and framework lay the foundation for understanding the potential harm of model-based evaluation metrics, and facilitate future works to develop more inclusive evaluation metrics.

11:00-12:30 (East Foyer / Virtual)
**#107 FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge**
*Shangbin Feng, Vidhisha Balachandran, Yuyang Bai and Yulia Tsvetkov*
Evaluating the factual consistency of automatically generated summaries is essential for the progress and adoption of reliable summarization systems. Despite recent advances, existing factuality evaluation models are not robust, being especially prone to entity and relation errors in new domains. We propose FactKB—a simple new approach to factuality evaluation that is generalizable across domains, in particular with respect to entities and relations. FactKB is based on language models pretrained using facts extracted from external knowledge bases. We introduce three types of complementary factuality pretraining objectives based on entity-specific facts, facts extracted from auxiliary knowledge about entities, and facts constructed compositionally through knowledge base walks. The resulting factuality evaluation model achieves state-of-the-art performance on two in-domain news summarization benchmarks as well as on three out-of-domain scientific literature datasets. Further analysis of FactKB shows improved ability to detect erroneous entities and relations in summaries and is robust and easily generalizable across domains.

11:00-12:30 (East Foyer / Virtual)
**#108 We Need to Talk About Reproducibility in NLP Model Comparison**
*Yan Xue, Xuefei Cao, Xingli Yang, Yu Wang, Ruibo Wang and Jihong Li*
NLPers frequently face reproducibility crisis in a comparison of various models of a real-world NLP task. Many studies have empirically showed that the standard splits tend to produce low reproducible and unreliable conclusions, and they attempted to improve the splits by using more random repetitions. However, the improvement on the reproducibility in a comparison of NLP models is limited attributed to a lack of

investigation on the relationship between the reproducibility and the estimator induced by a splitting strategy. In this paper, we formulate the reproducibility in a model comparison into a probabilistic function with regard to a conclusion. Furthermore, we theoretically illustrate that the reproducibility is qualitatively dominated by the signal-to-noise ratio (SNR) of a model performance estimator obtained on a corpus splitting strategy. Specifically, a higher value of the SNR of an estimator probably indicates a better reproducibility. On the basis of the theoretical motivations, we develop a novel mixture estimator of the performance of an NLP model with a regularized corpus splitting strategy based on a blocked $3 \times 2$ cross-validation. We conduct numerical experiments on multiple NLP tasks to show that the proposed estimator achieves a high SNR, and it substantially increases the reproducibility. Therefore, we recommend the NLP practitioners to use the proposed method to compare NLP models instead of the methods based on the widely-used standard splits and the random splits with multiple repetitions.

11:00-12:30 (East Foyer / Virtual)
**#109 GPT-RE: In-context Learning for Relation Extraction using Large Language Models**
*Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li and Sadao Kurohashi*
In spite of the potential for ground-breaking achievements offered by large language models (LLMs) (e.g., GPT-3) via in-context learning (ICL), they still lag significantly behind fully-supervised baselines (e.g., fine-tuned BERT) in relation extraction (RE). This is due to the two major shortcomings of ICL for RE: (1) low relevance regarding entity and relation in existing sentence-level demonstration retrieval approaches for ICL; and (2) the lack of explaining input-label mappings of demonstrations leading to poor ICL effectiveness. In this paper, we propose GPT-RE to successfully address the aforementioned issues by (1) incorporating task-aware representations in demonstration retrieval; and (2) enriching the demonstrations with gold label-induced reasoning logic. We evaluate GPT-RE on four widely-used RE datasets, and observe that GPT-RE achieves improvements over not only existing GPT-3 baselines, but also fully-supervised baselines as in Figure 1. Specifically, GPT-RE achieves SOTA performances on the Semeval and SciERC datasets, and competitive performances on the TACRED and ACE05 datasets. Additionally, a critical issue of LLMs revealed by previous work, the strong inclination to wrongly classify NULL examples into other pre-defined labels, is substantially alleviated by our method. We show an empirical analysis.

11:00-12:30 (East Foyer / Virtual)
**#110 Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection**
*Vyoma Raman, Eve Fleisig and Dan Klein*
The impact of AI models on marginalized communities has traditionally been measured by identifying performance differences between specified demographic subgroups. Though this approach aims to center vulnerable groups, it risks obscuring patterns of harm faced by intersectional subgroups or shared across multiple groups. To address this, we draw on theories of marginalization from disability studies and related disciplines, which state that people farther from the norm face greater adversity, to consider the "margins" in the domain of toxicity detection. We operationalize the "margins" of a dataset by employing outlier detection to identify text about people with demographic attributes distant from the "norm". We find that model performance is consistently worse for demographic outliers, with mean squared error (MSE) between outliers and non-outliers up to 70.4% worse across toxicity types. It is also worse for text outliers, with a MSE up to 68.4% higher for outliers than non-outliers. We also find text and demographic outliers to be particularly susceptible to errors in the classification of severe toxicity and identity attacks. Compared to analysis of disparities using traditional demographic breakdowns, we find that our outlier analysis frequently surfaces greater harms faced by a larger, more intersectional group, which suggests that outlier analysis is particularly beneficial for identifying harms against those groups.

11:00-12:30 (East Foyer / Virtual)
**#111 Incorporating Worker Perspectives into MTurk Annotation Practices for NLP**
*Olivia Huang, Eve Fleisig and Dan Klein*
Current practices regarding data collection for natural language processing on Amazon Mechanical Turk (MTurk) often rely on a combination of studies on data quality and heuristics shared among NLP researchers. However, without considering the perspectives of MTurk workers, these approaches are susceptible to issues regarding workers' rights and poor response quality. We conducted a critical literature review and a survey of MTurk workers aimed at addressing open questions regarding best practices for fair payment, worker privacy, data quality, and considering worker incentives. We found that worker preferences are often at odds with received wisdom among NLP researchers. Surveyed workers preferred reliable, reasonable payments over uncertain, very high payments; reported frequently lying on demographic questions; and expressed frustration at having work rejected with no explanation. We also found that workers view some quality control methods, such as requiring minimum response times or Master's qualifications, as biased and largely ineffective. Based on the survey results, we provide recommendations on how future NLP studies may better account for MTurk workers' experiences in order to respect workers' rights and improve data quality.

11:00-12:30 (East Foyer / Virtual)
**#112 MEGA: Multilingual Evaluation of Generative AI**
*Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali and Sunayana Sitaram*
Generative AI models have shown impressive performance on many Natural Language Processing tasks such as language understanding, reasoning, and language generation. An important question being asked by the AI community today is about the capabilities and limits of these models, and it is clear that evaluating generative AI is very challenging. Most studies on generative LLMs have been restricted to English and it is unclear how capable these models are at understanding and generating text in other languages. We present the first comprehensive benchmarking of generative LLMs - MEGA, which evaluates models on standard NLP benchmarks, covering 16 NLP datasets across 70 typologically diverse languages. We compare the performance of generative LLMs including Chat-GPT and GPT-4 to State of the Art (SOTA) non-autoregressive models on these tasks to determine how well generative models perform compared to the previous generation of LLMs. We present a thorough analysis of the performance of models across languages and tasks and discuss challenges in improving the performance of generative LLMs on low-resource languages. We create a framework for evaluating generative LLMs in the multilingual setting and provide directions for future progress in the field.

11:00-12:30 (East Foyer / Virtual)
**#113 PEFTDebias : Capturing debiasing information using PEFTs**
*Sumit Agarwal, Aditya Srikanth Veerubhotla and Srijan Bansal*
The increasing use of foundation models highlights the urgent need to address and eliminate implicit biases present in them that arise during pretraining. In this paper, we introduce PEFTDebias, a novel approach that employs parameter-efficient fine-tuning (PEFT) to mitigate the biases within foundation models. PEFTDebias consists of two main phases: an upstream phase for acquiring debiasing parameters along a specific bias axis, and a downstream phase where these parameters are incorporated into the model and frozen during the fine-tuning process. By evaluating on four datasets across two bias axes namely gender and race, we find that downstream biases can be effectively reduced with PEFTs. In addition, we show that these parameters possess axis-specific debiasing characteristics, enabling their effective transferability in mitigating biases in various downstream tasks.

11:00-12:30 (East Foyer / Virtual)

**#114 Reinforcement Replaces Supervision: Query focused Summarization using Deep Reinforcement Learning**
*Swaroop Nath, Pushpak Bhattacharyya and Harshad Khadilkar*
Query-focused Summarization (QfS) deals with systems that generate summaries from document(s) based on a query. Motivated by the insight that Reinforcement Learning (RL) provides a generalization to Supervised Learning (SL) for Natural Language Generation, and thereby performs better (empirically) than SL, we use an RL-based approach for this task of QfS. Additionally, we also resolve the conflict of employing RL in Transformers with Teacher Forcing. We develop multiple Policy Gradient networks, trained on various reward signals: ROUGE, BLEU, and Semantic Similarity, which lead to a *10*-point improvement over the State-of-the-Art approach on the ROUGE-L metric for a benchmark dataset (ELI5). We also show performance of our approach in zero-shot setting for another benchmark dataset (DebatePedia) – our approach leads to results comparable to baselines, which were specifically trained on DebatePedia. To aid the RL training, we propose a better semantic similarity reward, enabled by a novel Passage Embedding scheme developed using Cluster Hypothesis. Lastly, we contribute a gold-standard test dataset to further research in QfS and Long-form Question Answering (LfQA).

11:00-12:30 (East Foyer / Virtual)
**#115 Can Large Language Models Capture Dissenting Human Voices?**
*Noah Lee, Na Min An and James Thorne*
Large language models (LLMs) have shown impressive achievements in solving a broad range of tasks. Augmented by instruction fine-tuning, LLMs have also been shown to generalize in zero-shot settings as well. However, whether LLMs closely align with the human disagreement distribution has not been well-studied, especially within the scope of natural language inference (NLI). In this paper, we evaluate the performance and alignment of LLM distribution with humans using two different techniques to estimate the multinomial distribution: Monte Carlo Estimation (MCE) and Log Probability Estimation (LPE). As a result, we show LLMs exhibit limited ability in solving NLI tasks and simultaneously fail to capture human disagreement distribution. The inference and human alignment performances plunge even further on data samples with high human disagreement levels, raising concerns about their natural language understanding (NLU) ability and their representativeness to a larger human population.

11:00-12:30 (East Foyer / Virtual)
**#116 CRUSH4SQL: Collective Retrieval Using Schema Hallucination For Text2SQL**
*Mayank Kothyari, Dhruva Dhingra, Sunita Sarawagi and Soumen Chakrabarti*
Existing Text-to-SQL generators require the entire schema to be encoded with the user text. This is expensive or impractical for large databases with tens of thousands of columns. Standard dense retrieval techniques are inadequate for schema subsetting of a large structured database, where the correct semantics of retrieval demands that we rank sets of schema elements rather than individual documents. In response, we propose a two-stage process for effective coverage during retrieval. First, we use an LLM to hallucinate a minimal DB schema that it deems adequate to answer the query. We use the hallucinated schema to retrieve a subset of the actual schema, by composing the results from multiple dense retrievals. Remarkably, hallucination — generally considered a nuisance — turns out to be actually useful as a bridging mechanism. Since no existing benchmarks exist for schema subsetting on large databases, we introduce two benchmarks: (1) A semi-synthetic dataset of 4502 schema elements, by taking a union of schema on the well-known SPIDER dataset, and (2) A real-life benchmark called SocialDB sourced from an actual large data warehouse comprising of 17844 schema elements. We show that our method leads to significantly higher recall than SOTA retrieval-based augmentation methods.

11:00-12:30 (East Foyer / Virtual)
**#117 Small Language Models Fine-tuned to Coordinate Larger Language Models improve Complex Reasoning**
*Gurusha Juneja, Subhabrata Dutta, Soumen Chakrabarti, Sunny Manchanda and Tanmoy Chakraborty*
Large Language Models (LLMs) prompted to generate chain-of-thought (CoT) exhibit impressive reasoning capabilities. Recent attempts at prompt decomposition toward solving complex, multi-step reasoning problems depend on the ability of the LLM to simultaneously decompose and solve the problem. A significant disadvantage is that foundational LLMs are typically not available for fine-tuning, making adaptation computationally prohibitive. We believe (and demonstrate) that problem decomposition and solution generation are distinct capabilites, better addressed in separate modules, than by one monolithic LLM. We introduce DaSLaM, which uses a decomposition generator to decompose complex problems into subproblems that require fewer reasoning steps. These subproblems are answered by a solver. We use a relatively small (13B parameters) LM as the decomposition generator, which we train using policy gradient optimization to interact with a solver LM (regarded as black-box) and guide it through subproblems, thereby rendering our method solver-agnostic. Evaluation on multiple different reasoning datasets reveal that with our method, a 175 billion parameter LM (text-davinci-003) can produce competitive or even better performance, compared to its orders-of-magnitude larger successor, GPT-4. Additionally, we show that DaSLaM is not limited by the solver's capabilities as a function of scale; e.g., solver LMs with diverse sizes give significant performance improvement with our solver-agnostic decomposition technique. Exhaustive ablation studies evince the superiority of our modular finetuning technique over exorbitantly large decomposer LLMs, based on prompting alone.

11:00-12:30 (East Foyer / Virtual)
**#118 InterFair: Debiasing with Natural Language Feedback for Fair Interpretable Predictions**
*Bodhisattwa Prasad Majumder, Zexue He and Julian McAuley*
Debiasing methods in NLP models traditionally focus on isolating information related to a sensitive attribute (e.g., gender or race). We instead argue that a favorable debiasing method should use sensitive information 'fairly,' with explanations, rather than blindly eliminating it. This fair balance is often subjective and can be challenging to achieve algorithmically. We explore two interactive setups with a frozen predictive model and show that users able to provide feedback can achieve a better and *fairer* balance between task performance and bias mitigation. In one setup, users, by interacting with test examples, further decreased bias in the explanations (5-8%) while maintaining the same prediction accuracy. In the other setup, human feedback was able to disentangle associated bias and predictive information from the input leading to superior bias mitigation and improved task performance (4-5%) simultaneously.

11:00-12:30 (East Foyer / Virtual)
**#119 A Tale of Pronouns: Interpretability Informs Gender Bias Mitigation for Fairer Instruction-Tuned Machine Translation**
*Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza and Anne Lauscher*
Recent instruction fine-tuned models can solve multiple NLP tasks when prompted to do so, with machine translation (MT) being a prominent use case. However, current research often focuses on standard performance benchmarks, leaving compelling fairness and ethical considerations behind. In MT, this might lead to misgendered translations, resulting, among other harms, in the perpetuation of stereotypes and prejudices. In this work, we address this gap by investigating whether and to what extent such models exhibit gender bias in machine translation and how we can mitigate it. Concretely, we compute established gender bias metrics on the WinoMT corpus from English to German and Spanish. We discover that IFT models default to male-inflected translations, even disregarding female occupational stereotypes. Next, using interpretability methods, we unveil that models systematically overlook the pronoun indicating the gender of a target occupation in misgendered translations. Finally, based on this finding, we propose an easy-to-implement and effective bias mitigation solution based on few-shot learning that leads to significantly fairer translations.

11:00-12:30 (East Foyer / Virtual)
**#120 INFORM : Information eNtropy based multi-step reasoning FOR large language Models**
*Chuyue Zhou, Wangjie You, Juntao Li, Jing Ye, Kehai Chen and Min Zhang*
Large language models (LLMs) have demonstrated exceptional performance in reasoning tasks with dedicated Chain-of-Thought (CoT) prompts. Further enhancing CoT prompts with exquisite exemplars can significantly improve reasoning performance.However, the effectiveness of CoT prompts may fluctuate dramatically with different choices of in-context examples. Additionally, manual construction of rationale steps can be time-consuming, presenting challenges for the widespread adoption of CoT prompting. In this work, we propose a novel approach by introducing information entropy (IE) as a criteria on for CoT prompt selection. We extend this criterion to the CoT generation and inference stages, automatically generating CoT prompts with higher information entropy scores and adaptively determining the number of samples. These three stages together form our proposed information- entropy-based multi-step reasoning for large language models, named INFORM. Our experiments across seven reasoning benchmarks utilizing two language models(GPT-3.5-Turbo and text-davinci-003) demonstrate the superiority of INFORM both in performance and efficiency.

11:00-12:30 (East Foyer / Virtual)
**#121 Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4**
*Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Alexander Gupta, Joel Christoph, Jean-François Godbout and Reihaneh Rabbany*
Misinformation poses a critical societal challenge, and current approaches have yet to produce an effective solution. We propose focusing on generalization, uncertainty, and how to leverage recent large language models, in order to create more practical tools to evaluate information veracity in contexts where perfect classification is impossible. We first demonstrate that GPT-4 can outperform prior methods in multiple settings and languages. Next, we explore generalization, revealing that GPT-4 and RoBERTa-large exhibit differences in failure modes. Third, we propose techniques to handle uncertainty that can detect impossible examples and strongly improve outcomes. We also discuss results on other language models, temperature, prompting, versioning, explainability, and web retrieval, each one providing practical insights and directions for future research. Finally, we publish the LIAR-New dataset with novel paired English and French misinformation data and Possibility labels that indicate if there is sufficient context for veracity evaluation. Overall, this research lays the groundwork for future tools that can drive real-world progress to combat misinformation.

11:00-12:30 (East Foyer / Virtual)
**#122 Doolittle: Benchmarks and Corpora for Academic Writing Formalization**
*Shizhe Diao, Yongyu Lei, Liangming Pan, Tianqing Fang, Wangchunshu Zhou, Sedrick Scott Keh, Min-Yen Kan and Tong Zhang*
Improving the quality of academic writing is a meaningful but challenging task. Conventional methods of language refinement focus on narrow, specific linguistic features within isolated sentences, such as grammatical errors and improper word use. We propose a more general task, Academic Writing Formalization (AWF), to improve the overall quality of formal academic writing at the paragraph level. We formulate this language refinement task as a formal text style transfer task which transfers informal-academic text to formal-academic and contribute a large-scale non-parallel dataset, Doolittle, for this purpose. Concurrently, we apply a method named metric-oriented reinforcement learning (MORL) to two large language models (LLM) where we incorporate different levels of automatic feedback into the training process. Our experiments reveal that existing text transfer models and grammatical error correction models address certain aspects of AWF but still have a significant performance gap compared to human performance. Meanwhile, language models fine-tuned with our MORL method exhibit considerably improved performance, rivaling the latest chatbot ChatGPT, but still have a non-negligible gap compared to the ground truth formal-academic texts in Doolittle.

11:00-12:30 (East Foyer / Virtual)
**#123 Enhancing Biomedical Lay Summarisation with External Knowledge Graphs**
*Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton and Chenghua Lin*
Previous approaches for automatic lay summarisation are exclusively reliant on the source article that, given it is written for a technical audience (e.g., researchers), is unlikely to explicitly define all technical concepts or state all of the background information that is relevant for a lay audience. We address this issue by augmenting eLife, an existing biomedical lay summarisation dataset, with article-specific knowledge graphs, each containing detailed information on relevant biomedical concepts. Using both automatic and human evaluations, we systematically investigate the effectiveness of three different approaches for incorporating knowledge graphs within lay summarisation models, with each method targeting a distinct area of the encoder-decoder model architecture. Our results confirm that integrating graph-based domain knowledge can significantly benefit lay summarisation by substantially increasing the readability of generated text and improving the explanation of technical concepts.

11:00-12:30 (East Foyer / Virtual)
**#124 Do Transformers Parse while Predicting the Masked Word?**
*Haoyu Zhao, Abhishek Panigrahi, Rong Ge and Sanjeev Arora*
Pre-trained language models have been shown to encode linguistic structures like parse trees in their embeddings while being trained unsupervised. Some doubts have been raised whether the models are doing parsing or only some computation weakly correlated with it. Concretely: (a) Is it possible to explicitly describe transformers with realistic embedding dimensions, number of heads, etc. that are capable of doing parsing — or even approximate parsing? (b) Why do pre-trained models capture parsing structure? This paper takes a step toward answering these questions in the context of generative modeling with PCFGs. We show that masked language models like BERT or RoBERTa of moderate sizes can approximately execute the Inside-Outside algorithm for the English PCFG (Marcus et al., 1993). We also show that the Inside-Outside algorithm is optimal for masked language modeling loss on the PCFG-generated data. We conduct probing experiments on models pre-trained on PCFG-generated data to show that this not only allows recovery of approximate parse tree, but also recovers marginal span probabilities computed by the Inside-Outside algorithm, which suggests an implicit bias of masked language modeling towards this algorithm.

11:00-12:30 (East Foyer / Virtual)
**#125 Can LMs Generalize to Future Data? An Empirical Analysis on Text Summarization**
*Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu and Lidia S. Chao*
Recent pre-trained language models (PLMs) achieve promising results in existing abstractive summarization datasets. However, existing summarization benchmarks overlap in time with the standard pre-training corpora and finetuning datasets. Hence, the strong performance of PLMs may rely on the parametric knowledge that is memorized during pre-training and fine-tuning. Moreover, the knowledge memorized by PLMs may quickly become outdated, which affects the generalization performance of PLMs on future data. In this work, we propose TempoSum, a novel benchmark that contains data samples from 2010 to 2022, to understand the temporal generalization ability of abstractive summarization models. Through extensive human evaluation, we show that parametric knowledge stored in summarization models significantly affects the faithfulness of the generated summaries on future data. Moreover, existing faithfulness enhancement methods cannot reliably improve the faithfulness of summarization models on future data. Finally, we discuss several recommendations to the research community on how to evaluate and improve the temporal generalization capability of text summarization models.

11:00-12:30 (East Foyer / Virtual)
**#126 Towards Interpretable and Efficient Automatic Reference-Based Summarization Evaluation**
*Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong and Dragomir Radev*
Interpretability and efficiency are two important considerations for the adoption of neural automatic metrics. In this work, we develop strong-performing automatic metrics for reference-based summarization evaluation, based on a two-stage evaluation pipeline that first extracts basic information units from one text sequence and then checks the extracted units in another sequence. The metrics we developed include two-stage metrics that can provide high interpretability at both the fine-grained unit level and summary level, and one-stage metrics that achieve a balance between efficiency and interpretability. We make the developed tools publicly available at https://github.com/Yale-LILY/AutoACU.

## Findings 5

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems**
*Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych and Mrinmaya Sachan*
While automatic dialogue tutors hold great potential in making education personalized and more accessible, research on such systems has been hampered by a lack of sufficiently large and high-quality datasets. Collecting such datasets remains challenging, as recording tutoring sessions raises privacy concerns and crowdsourcing leads to insufficient data quality. To address this, we propose a framework to generate such dialogues by pairing human teachers with a Large Language Model (LLM) prompted to represent common student errors. We describe how we use this framework to collect MathDial, a dataset of 3k one-to-one teacher-student tutoring dialogues grounded in multi-step math reasoning problems. While models like GPT-3 are good problem solvers, they fail at tutoring because they generate factually incorrect feedback or are prone to revealing solutions to students too early. To overcome this, we let teachers provide learning opportunities to students by guiding them using various scaffolding questions according to a taxonomy of teacher moves. We demonstrate MathDial and its extensive annotations can be used to finetune models to be more effective tutors (and not just solvers). We confirm this by automatic and human evaluation, notably in an interactive setting that measures the trade-off between student solving success and telling solutions. The dataset is released publicly.

11:00-12:30 (East Foyer / Virtual)
**VISIT: Visualizing and Interpreting the Semantic Information Flow of Transformers**
*Shahar Katz and Yonatan Belinkov*
Recent advances in interpretability suggest we can project weights and hidden states of transformer-based language models (LMs) to their vocabulary, a transformation that makes them more human interpretable. In this paper, we investigate LM attention heads and memory values, the vectors the models dynamically create and recall while processing a given input. By analyzing the tokens they represent through this projection, we identify patterns in the information flow inside the attention mechanism. Based on our discoveries, we create a tool to visualize a forward pass of Generative Pre-trained Transformers (GPTs) as an interactive flow graph, with nodes representing neurons or hidden states and edges representing the interactions between them. Our visualization simplifies huge amounts of data into easy-to-read plots that can reflect the models' internal processing, uncovering the contribution of each component to the models' final prediction. Our visualization also unveils new insights about the role of layer norms as semantic filters that influence the models' output, and about neurons that are always activated during forward passes and act as regularization vectors.

11:00-12:30 (East Foyer / Virtual)
**Crosslingual Transfer Learning for Low-Resource Languages Based on Multilingual Colexification Graphs**
*Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei and Hinrich Schuetze*
In comparative linguistics, colexification refers to the phenomenon of a lexical form conveying two or more distinct meanings. Existing work on colexification patterns relies on annotated word lists, limiting scalability and usefulness in NLP. In contrast, we identify colexification patterns of more than 2,000 concepts across 1,335 languages directly from an unannotated parallel corpus. We then propose simple and effective methods to build multilingual graphs from the colexification patterns: **ColexNet** and **ColexNet+**. ColexNet's nodes are concepts and its edges are colexifications. In ColexNet+, concept nodes are additionally linked through intermediate nodes, each representing an ngram in one of 1,334 languages. We use ColexNet+ to train $\overrightarrow{\text{ColexNet+}}$, high-quality multilingual embeddings that are well-suited for transfer learning. In our experiments, we first show that ColexNet achieves high recall on CLICS, a dataset of crosslingual colexifications. We then evaluate $\overrightarrow{\text{ColexNet+}}$ on roundtrip translation, sentence retrieval and sentence classification and show that our embeddings surpass several transfer learning baselines. This demonstrates the benefits of using colexification as a source of information in multilingual NLP.

11:00-12:30 (East Foyer / Virtual)
**Enhancing Scalability of Pre-trained Language Models via Efficient Parameter Sharing**
*Peiyu Liu, Ze-Feng Gao, Yushuo Chen, Xin Zhao and Ji-Rong Wen*
In this paper, we propose a highly parameter-efficient approach to scaling pre-trained language models (PLMs) to a deeper model depth. Unlike prior work that shares all parameters or uses extra blocks, we design a more capable parameter-sharing architecture based on matrix product operator (MPO), an efficient tensor decomposition method to factorize the parameter matrix into a set of local tensors. Based on such a decomposition, we share the important local tensor across all layers for reducing the model size and meanwhile keep layer-specific tensors (also using Adapters) for enhancing the adaptation flexibility. To improve the model training, we further propose a stable initialization algorithm tailored for the MPO-based architecture. Extensive experiments have demonstrated the effectiveness of our proposed model in enhancing scalability and achieving higher performance (i.e., with fewer parameters than BERT-base, we successfully scale the model depth by a factor of 4x and even achieve 0.1 points higher than BERT-large for GLUE score). The code to reproduce the results of this paper can be found at https://github.com/RUCAIBox/MPOBERT-code.

11:00-12:30 (East Foyer / Virtual)
**Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data**
*Kashun Shum, Shizhe Diao and Tong Zhang*
Chain-of-thought (CoT) advances the reasoning abilities of large language models (LLMs) and achieves superior performance in complex reasoning tasks. However, most CoT studies rely on carefully designed human-annotated rational chains to prompt LLMs, posing challenges for real-world applications where labeled data is available without rational chains. This paper proposes a new strategy, AutomateCoT (Automatic Prompt Augmentation and Selection with Chain-of-Thought), that can bypass human engineering of CoT by automatically augmenting rational chains from a small labeled dataset, and then pruning low-quality chains to construct a candidate pool of machine-generated rationale chains based on the labels. Finally, it selects the optimal combination of several rationale chains from the pool for CoT prompting by employing a variance-reduced policy gradient strategy to estimate the significance of each example. Automate-CoT enables a quick adaptation of the

CoT technique to different tasks. Experimental results demonstrate the effectiveness of our method, where competitive results are achieved on arithmetic reasoning (+2.7%), commonsense reasoning (+3.4%), symbolic reasoning (+3.2%), and non-reasoning tasks (+2.5%).

11:00-12:30 (East Foyer / Virtual)
**Code-Switching with Word Senses for Pretraining in Neural Machine Translation**
*Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Z. Pan and Roberto Navigli*
Lexical ambiguity is a significant and pervasive challenge in Neural Machine Translation (NMT), with many state-of-the-art (SOTA) NMT systems struggling to handle polysemous words (Campolungo et al., 2022). The same holds for the NMT pretraining paradigm of denoising synthetic "code-switched" text (Pan et al., 2021; Iyer et al., 2023), where word senses are ignored in the noising stage – leading to harmful sense biases in the pretraining data that are subsequently inherited by the resulting models. In this work, we introduce Word Sense Pretraining for Neural Machine Translation (WSP-NMT) - an end-to-end approach for pretraining multilingual NMT models leveraging word sense-specific information from Knowledge Bases. Our experiments show significant improvements in overall translation quality. Then, we show the robustness of our approach to scale to various challenging data and resource-scarce scenarios and, finally, report fine-grained accuracy improvements on the DiBiMT disambiguation benchmark. Our studies yield interesting and novel insights into the merits and challenges of integrating word sense information and structured knowledge in multilingual pretraining for NMT.

11:00-12:30 (East Foyer / Virtual)
**Accuracy is not enough: Evaluating Personalization in Summarizers**
*Rahul Vansh, Darsh Rank, Sourish Dasgupta and Tanmoy Chakraborty*
Text summarization models are evaluated in terms of their accuracy and quality using various measures such as ROUGE, BLEU, METEOR, BERTScore, PYRAMID, readability, and several other recently proposed ones. The central objective of all accuracy measures is to evaluate the model's ability to capture *saliency* accurately. Since saliency is subjective w.r.t the readers' preferences, there cannot be a fit-all summary for a given document. This means that in many use-cases, summarization models need to be personalized w.r.t user-profiles. However, to our knowledge, there is no measure to evaluate the *degree-of-personalization* of a summarization model. In this paper, we first establish that existing accuracy measures cannot evaluate the degree of personalization of any summarization model, and then propose a novel measure, called $EGISES$, for automatically computing the same. Using the PENS dataset released by Microsoft Research, we analyze the degree of personalization of ten different state-of-the-art summarization models (both extractive and abstractive), five of which are explicitly trained for personalized summarization, and the remaining are appropriated to exhibit personalization. We conclude by proposing a generalized accuracy measure, called $P\text{-}Accuracy$, for designing accuracy measures that should also take personalization into account and demonstrate the robustness and reliability of the measure through meta-evaluation.

11:00-12:30 (East Foyer / Virtual)
**Understanding HTML with Large Language Models**
*Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin V Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel and Aleksandra Faust*
Large language models (LLMs) have shown exceptional performance on a variety of natural language tasks. Yet, their capabilities for HTML understanding – i.e., parsing the raw HTML of a webpage, with applications to automation of web-based tasks, crawling, and browser-assisted retrieval – have not been fully explored. We contribute HTML understanding models (fine-tuned LLMs) and an in-depth analysis of their capabilities under three tasks: (i) Semantic Classification of HTML elements, (ii) Description Generation for HTML inputs, and (iii) Autonomous Web Navigation of HTML pages. While previous work has developed dedicated architectures and training procedures for HTML understanding, we show that LLMs pretrained on standard natural language corpora transfer remarkably well to HTML understanding tasks. For instance, when fine-tuned on data from the MiniWoB benchmark, LLMs successfully complete 50% more tasks using 192x less data compared to the previous best supervised model. We create and open-source a large-scale HTML dataset distilled and auto-labeled from CommonCrawl

11:00-12:30 (East Foyer / Virtual)
**Compositional Generalization for Data-to-Text Generation**
*Xinnuo Xu, Ivan Titov and Mirella Lapata*
Data-to-text generation involves transforming structured data, often represented as predicate-argument tuples, into coherent textual descriptions. Despite recent advances, systems still struggle when confronted with unseen combinations of predicates, producing unfaithful descriptions (e.g.,hallucinations or omissions). We refer to this issue as compositional generalisation, and it encouraged us to create a benchmark for assessing the performance of different approaches on this specific problem. Furthermore, we propose a novel model that addresses compositional generalization by clustering predicates into groups. Our model generates text in a sentence-by-sentence manner, relying on one cluster of predicates at a time. This approach significantly outperforms T5-baselines across all evaluation metrics. Notably, it achieved a 31% improvement over T5 in terms of a metric focused on maintaining faithfulness to the input.

11:00-12:30 (East Foyer / Virtual)
**Context Quality Matters in Training Fusion-in-Decoder for Extractive Open-Domain Question Answering**
*Kosuke Akimoto, Kunihiro Takeoka and Masafumi Oyamada*
Retrieval-augmented generation models augment knowledge encoded in a language model by providing additional relevant external knowledge (context) during generation. Although it has been shown that the quantity and quality of context impact the performance of retrieval-augmented generation models during inference, limited research explores how these characteristics affect model training. This paper explores how context quantity and quality during model training affect the performance of Fusion-in-Decoder (FiD), the state-of-the-art retrieval-augmented generation model, in extractive open-domain question answering tasks. Experimental results suggest that FiD models overfit to context quality during training and show suboptimal performance when evaluated on different context quality. Through the experimental results, we also reveal FiD models trained with different context quality have different cross-attention distribution patterns. Specifically, as context quality during training increases, FiD models tend to attend more uniformly to each passage in context. Finally, based on these observations, we propose a method to mitigate overfitting to specific context quality by introducing bias to the cross-attention distribution, which we demonstrate to be effective in improving the performance of FiD models on different context quality.

11:00-12:30 (East Foyer / Virtual)
**ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models**
*Alex Mei, Sharon Levy and William Yang Wang*
As large language models are integrated into society, robustness toward a suite of prompts is increasingly important to maintain reliability in a high-variance environment.Robustness evaluations must comprehensively encapsulate the various settings in which a user may invoke an intelligent system. This paper proposes ASSERT, Automated Safety Scenario Red Teaming, consisting of three methods – semantically aligned augmentation, target bootstrapping, and adversarial knowledge injection. For robust safety evaluation, we apply these methods in the critical domain of AI safety to algorithmically generate a test suite of prompts covering diverse robustness settings – semantic equivalence, related scenarios, and adversarial. We partition our prompts into four safety domains for a fine-grained analysis of how the domain affects

model performance. Despite dedicated safeguards in existing state-of-the-art models, we find statistically significant performance differences of up to 11% in absolute classification accuracy among semantically related scenarios and error rates of up to 19% absolute error in zero-shot adversarial settings, raising concerns for users' physical safety.

11:00-12:30 (East Foyer / Virtual)
**Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation**
*Miao Li, Eduard Hovy and Jey Han Lau*
We present PeerSum, a novel dataset for generating meta-reviews of scientific papers. The meta-reviews can be interpreted as abstractive summaries of reviews, multi-turn discussions and the paper abstract. These source documents have a rich inter-document relationship with an explicit hierarchical conversational structure, cross-references and (occasionally) conflicting information. To introduce the structural inductive bias into pre-trained language models, we introduce RAMMER (Relationship-aware Multi-task Meta-review Generator), a model that uses sparse attention based on the conversational structure and a multi-task training objective that predicts metadata features (e.g., review ratings). Our experimental results show that RAMMER outperforms other strong baseline models in terms of a suite of automatic evaluation metrics. Further analyses, however, reveal that RAMMER and other models struggle to handle conflicts in source documents, suggesting meta-review generation is a challenging task and a promising avenue for further research.

11:00-12:30 (East Foyer / Virtual)
**Frugal Prompting for Dialog Models**
*Bishal Santra, Sakya Basak, Abhinandan De, Manish Gupta and Pawan Goyal*
The use of large language models (LLMs) in natural language processing (NLP) tasks is rapidly increasing, leading to changes in how researchers approach problems in the field. To fully utilize these models' abilities, a better understanding of their behavior for different input protocols is required. With LLMs, users can directly interact with the models through a text-based interface to define and solve various tasks. Hence, understanding the conversational abilities of these LLMs, which may not have been specifically trained for dialog modeling, is also important. This study examines different approaches for building dialog systems using LLMs by considering various aspects of the prompt. As part of prompt tuning, we experiment with various ways of providing instructions, exemplars, current query and additional context. The research also analyzes the representations of dialog history that have the optimal usable-information density. Based on the findings, the paper suggests more compact ways of providing dialog history information while ensuring good performance and reducing model's inference-API costs. The research contributes to a better understanding of how LLMs can be effectively used for building interactive systems.

11:00-12:30 (East Foyer / Virtual)
**GlotLID: Language Identification for Low-Resource Languages**
*Amir Hossein Kargaran, Ayyoob Imani, François Yvon and Hinrich Schuetze*
Several recent papers have published good solutions for language identification (LID) for about 300 high-resource and medium-resource languages. However, there is no LID available that (i) covers a wide range of low-resource languages, (ii) is rigorously evaluated and reliable and (iii) efficient and easy to use. Here, we publish GlotLID-M, an LID model that satisfies the desiderata of wide coverage, reliability and efficiency. It identifies 1665 languages, a large increase in coverage compared to prior work. In our experiments, GlotLID-M outperforms four baselines (CLD3, FT176, OpenLID and NLLB) when balancing F1 and false positive rate (FPR). We analyze the unique challenges that low-resource LID poses: incorrect corpus metadata, leakage from high-resource languages, difficulty separating closely related languages, handling of macrolanguage vs varieties and in general noisy data. We hope that integrating GlotLID-M into dataset creation pipelines will improve quality and enhance accessibility of NLP technology for low-resource languages and cultures. GlotLID-M model, code, and list of data sources are available: https://github.com/cisnlp/GlotLID.

11:00-12:30 (East Foyer / Virtual)
**What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study**
*Aman Madaan, Katherine Hermann and Amir Yazdanbakhsh*
The effectiveness of Chain-of-thought prompting (CoT) has been widely recognized, but the underlying mechanisms behind its success, the reason why it just works for a wide range of tasks, remains an open question. To investigate this, we employ a counterfactual prompting approach, systematically manipulating elements of examples used in a few-shot prompt, and testing the consequences on model behavior. This allows us to understand the relative contributions of prompt elements such as symbols (digits, entities) and patterns (equations, sentence structure) on in-context learning. Our experiments with three different large language models (LLMs) reveal several key findings. First, the specific symbols used in the prompt do not significantly impact the model's performance. However, consistent patterns in examples and specifying text in style frequently found on the web are crucial. Second, our findings suggest that the necessity of accurate few-shot examples depends on their role in communicating task understanding. We identify tasks where inaccurate few-shot examples hurt and, surprisingly, tasks where they improve performance. Additionally, we find that the intermediate steps in CoT may not necessarily facilitate learning how to solve a task, but instead efficiently convey task understanding (what) to the model. Furthermore, CoT leverages LLMs to fill in missing commonsense information, particularly helping difficult reasoning problems and long-tail questions.

11:00-12:30 (East Foyer / Virtual)
**BioDEX: Large-Scale Biomedical Adverse Drug Event Extraction for Real-World Pharmacovigilance**
*Karel D'Oosterlinck, François Remy, Johannes Deleu, Thomas Demeester, Chris Develder, Klim Zaporojets, Aneiss Ghodsi, Simon Ellershaw, Jack Collins and Christopher Potts*
Timely and accurate extraction of Adverse Drug Events (ADE) from biomedical literature is paramount for public safety, but involves slow and costly manual labor. We set out to improve drug safety monitoring (pharmacovigilance, PV) through the use of Natural Language Processing (NLP). We introduce BioDEX, a large-scale resource for Biomedical adverse Drug Event eXtraction, rooted in the historical output of drug safety reporting in the U.S. BioDEX consists of 65k abstracts and 19k full-text biomedical papers with 256k associated document-level safety reports created by medical experts. The core features of these reports include the reported weight, age, and biological sex of a patient, a set of drugs taken by the patient, the drug dosages, the reactions experienced, and whether the reaction was life threatening. In this work, we consider the task of predicting the core information of the report given its originating paper. We estimate human performance to be 72.0% F1, whereas our best model achieves 59.1% F1 (62.3 validation), indicating significant headroom. We also begin to explore ways in which these models could help professional PV reviewers. Our code and data are available at https://github.com/KarelDO/BioDEX.

11:00-12:30 (East Foyer / Virtual)
**Mind the Gap: Automated Corpus Creation for Enthymeme Detection and Reconstruction in Learner Arguments**
*Maja Stahl, Nick Düsterhus, Mei-Hua Chen and Henning Wachsmuth*
Writing strong arguments can be challenging for learners. It requires to select and arrange multiple argumentative discourse units (ADUs) in a logical and coherent way as well as to decide which ADUs to leave implicit, so called enthymemes. However, when important ADUs are missing, readers might not be able to follow the reasoning or understand the argument's main point. This paper introduces two new tasks for learner arguments: to identify gaps in arguments (enthymeme detection) and to fill such gaps (enthymeme reconstruction). Approaches to both tasks may help learners improve their argument quality. We study how corpora for these tasks can be created automatically by deleting ADUs

from an argumentative text that are central to the argument and its quality, while maintaining the text's naturalness. Based on the ICLEv3 corpus of argumentative learner essays, we create 40,089 argument instances for enthymeme detection and reconstruction. Through manual studies, we provide evidence that the proposed corpus creation process leads to the desired quality reduction, and results in arguments that are similarly natural to those written by learners. Finally, first baseline approaches to enthymeme detection and reconstruction demonstrate the corpus' usefulness.

11:00-12:30 (East Foyer / Virtual)
**Boosting Prompt-Based Self-Training With Mapping-Free Automatic Verbalizer for Multi-Class Classification**
*Yookyung Kho, Jaehee Kim and Pilsung Kang*
Recently, prompt-based fine-tuning has garnered considerable interest as a core technique for few-shot text classification task. This approach reformulates the fine-tuning objective to align with the Masked Language Modeling (MLM) objective. Leveraging unlabeled data, prompt-based self-training has shown greater effectiveness in binary and three-class classification. However, prompt-based self-training for multi-class classification has not been adequately investigated, despite its significant applicability to real-world scenarios. Moreover, extending current methods to multi-class classification suffers from the verbalizer that extracts the predicted value of manually pre-defined single label word for each class from MLM predictions. Consequently, we introduce a novel, efficient verbalizer structure, named Mapping-free Automatic Verbalizer (MAV). Comprising two fully connected layers, MAV serves as a trainable verbalizer that automatically extracts the requisite word features for classification by capitalizing on all available information from MLM predictions. Experimental results on five multi-class classification datasets indicate MAV's superior self-training efficacy.

11:00-12:30 (East Foyer / Virtual)
**From Chaos to Clarity: Claim Normalization to Empower Fact-Checking**
*Megha Sundriyal, Tanmoy Chakraborty and Preslav Nakov*
With the rise of social media, users are exposed to many misleading claims. However, the pervasive noise inherent in these posts presents a challenge in identifying precise and prominent claims that require verification. Extracting the important claims from such posts is arduous and time-consuming, yet it is an underexplored problem. Here, we aim to bridge this gap. We introduce a novel task, Claim Normalization (aka ClaimNorm), which aims to decompose complex and noisy social media posts into more straightforward and understandable forms, termed normalized claims. We propose CACN, a pioneering approach that leverages chain-of-thought and claim check-worthiness estimation, mimicking human reasoning processes, to comprehend intricate claims. Moreover, we capitalize on the in-context learning capabilities of large language models to provide guidance and to improve claim normalization. To evaluate the effectiveness of our proposed model, we meticulously compile a comprehensive real-world dataset, CLAN, comprising more than 6k instances of social media posts alongside their respective normalized claims. Our experiments demonstrate that CACN outperforms several baselines across various evaluation measures. Finally, our rigorous error analysis validates CACN's capabilities and pitfalls.

11:00-12:30 (East Foyer / Virtual)
**Mitigating Biases in Hate Speech Detection from A Causal Perspective**
*Zhehao Zhang, Jiaao Chen and Diyi Yang*
Nowadays, many hate speech detectors are built to automatically detect hateful content. However, their training sets are sometimes skewed towards certain stereotypes (e.g., race or religion-related). As a result, the detectors are prone to depend on some shortcuts for predictions. Previous works mainly focus on token-level analysis and heavily rely on human experts' annotations to identify spurious correlations, which is not only costly but also incapable of discovering higher-level artifacts. In this work, we use grammar induction to find grammar patterns for hate speech and analyze this phenomenon from a causal perspective. Concretely, we categorize and verify different biases based on their spuriousness and influence on the model prediction. Then, we propose two mitigation approaches including Multi-Task Intervention and Data-Specific Intervention based on these confounders. Experiments conducted on 9 hate speech datasets demonstrate the effectiveness of our approaches.

11:00-12:30 (East Foyer / Virtual)
**DiQAD: A Benchmark Dataset for Open-domain Dialogue Quality Assessment**
*Yukun Zhao, Lingyong Yan, Weiwei Sun, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren and Dawei Yin*
Dialogue assessment plays a critical role in the development of open-domain dialogue systems. Existing work are uncapable of providing an end-to-end and human-epistemic assessment dataset, while they only provide sub-metrics like coherence or the dialogues are conversed between annotators far from real user settings. In this paper, we release a large-scale dialogue quality assessment dataset (DiQAD), for automatically assessing open-domain dialogue quality. Specifically, we (1) establish the assessment criteria based on the dimensions conforming to human judgements on dialogue qualities, and (2) annotate large-scale dialogues that conversed between real users based on these annotation criteria, which contains around 100,000 dialogues. We conduct several experiments and report the performances of the baselines as the benchmark on DiQAD. The dataset is openly accessible at https://github.com/yukunZhao/Dataset_Dialogue_quality_evaluation.

11:00-12:30 (East Foyer / Virtual)
**Let's Synthesize Step by Step: Iterative Dataset Synthesis with Large Language Models by Extrapolating Errors from Small Models**
*Ruida Wang, Wangchunshu Zhou and Mrinmaya Sachan*
*Data Synthesis* is a promising way to train a small model with very little labeled data. One approach for data synthesis is to leverage the rich knowledge from large language models to synthesize pseudo training examples for small models, making it possible to achieve both data and compute efficiency at the same time. However, a key challenge in data synthesis is that the synthesized dataset often suffers from a large distributional discrepancy from the *real task* data distribution. Thus, in this paper, we propose *Synthesis Step by Step* (**S3**), a data synthesis framework that shrinks this distribution gap by iteratively extrapolating the errors made by a small model trained on the synthesized dataset on a small real-world validation dataset using a large language model. Extensive experiments on multiple NLP tasks show that our approach improves the performance of a small model by reducing the gap between the synthetic dataset and the real data, resulting in significant improvement compared to several baselines: 9.48% improvement compared to ZeroGen and 2.73% compared to GoldGen, and at most 15.17% improvement compared to the small model trained on human-annotated data.

11:00-12:30 (East Foyer / Virtual)
**Solving the Right Problem is Key for Translational NLP: A Case Study in UMLS Vocabulary Insertion**
*Bernal Jimenez Gutierrez, Yuqing Mao, Vinh Nguyen, Kin Wah Fung, Yu Su and Olivier Bodenreider*
As the immense opportunities enabled by large language models become more apparent, NLP systems will be increasingly expected to excel in real-world settings. However, in many instances, powerful models alone will not yield translational NLP solutions, especially if the formulated problem is not well aligned with the real-world task. In this work, we study the case of UMLS vocabulary insertion, an important real-world task in which hundreds of thousands of new terms, referred to as atoms, are added to the UMLS, one of the most comprehensive open-source biomedical knowledge bases. Previous work aimed to develop an automated NLP system to make this time-consuming, costly, and error-prone task more efficient. Nevertheless, practical progress in this direction has been difficult to achieve due to a problem formu-

lation and evaluation gap between research output and the real-world task. In order to address this gap, we introduce a new formulation for UMLS vocabulary insertion which mirrors the real-world task, datasets which faithfully represent it and several strong baselines we developed through re-purposing existing solutions. Additionally, we propose an effective rule-enhanced biomedical language model which enables important new model behavior, outperforms all strong baselines and provides measurable qualitative improvements to editors who carry out the UVI task. We hope this case study provides insight into the considerable importance of problem formulation for the success of translational NLP solutions.

11:00-12:30 (East Foyer / Virtual)
**MoqaGPT : Zero-Shot Multi-modal Open-domain Question Answering with Large Language Model**
*Le Zhang, Yihong Wu, Fengran Mo, Jian-Yun Nie and Aishwarya Agrawal*
Multi-modal open-domain question answering typically requires evidence retrieval from databases across diverse modalities, such as images, tables, passages, etc. Even Large Language Models (LLMs) like GPT-4 fall short in this task. To enable LLMs to tackle the task in a zero-shot manner, we introduce MoqaGPT, a straightforward and flexible framework. Using a divide-and-conquer strategy that bypasses intricate multi-modality ranking, our framework can accommodate new modalities and seamlessly transition to new models for the task. Built upon LLMs, MoqaGPT retrieves and extracts answers from each modality separately, then fuses this multi-modal information using LLMs to produce a final answer. Our methodology boosts performance on the MMCoQA dataset, improving F1 by +37.91 points and EM by +34.07 points over the supervised baseline. On the MultiModalQA dataset, MoqaGPT surpasses the zero-shot baseline, improving F1 by 9.5 points and EM by 10.1 points, and significantly closes the gap with supervised methods. Our codebase is available at https://github.com/lezhang7/MOQAGPT.

11:00-12:30 (East Foyer / Virtual)
**Goodtriever: Adaptive Toxicity Mitigation with Retrieval-augmented Models**
*Luiza Amador Pozzobon, Beyza Ermis, Patrick Lewis and Sara Hooker*
Considerable effort has been dedicated to mitigating toxicity, but existing methods often require drastic modifications to model parameters or the use of computationally intensive auxiliary models. Furthermore, previous approaches have often neglected the crucial factor of language's evolving nature over time. In this work, we present a comprehensive perspective on toxicity mitigation that takes into account its changing nature. We introduce Goodtriever, a flexible methodology that matches the current state-of-the-art toxicity mitigation while achieving 43% relative latency reduction during inference and being more computationally efficient. By incorporating a retrieval-based approach at decoding time, Goodtriever enables toxicity-controlled text generation. Our research advocates for an increased focus on adaptable mitigation techniques, which better reflect the data drift models face when deployed in the wild.

11:00-12:30 (East Foyer / Virtual)
**ReLM: Leveraging Language Models for Enhanced Chemical Reaction Prediction**
*Yaorui Shi, An Zhang, Enzhi Zhang, Zhiyuan Liu and Xiang Wang*
Predicting chemical reactions, a fundamental challenge in chemistry, involves forecasting the resulting products from a given reaction process. Conventional techniques, notably those employing Graph Neural Networks (GNNs), are often limited by insufficient training data and their inability to utilize textual information, undermining their applicability in real-world applications. In this work, we propose **ReLM**, a novel framework that leverages the chemical knowledge encoded in language models (LMs) to assist GNNs, thereby enhancing the accuracy of real-world chemical reaction predictions. To further enhance the model's robustness and interpretability, we incorporate the confidence score strategy, enabling the LMs to self-assess the reliability of their predictions. Our experimental results demonstrate that ReLM improves the performance of state-of-the-art GNN-based methods across various chemical reaction datasets, especially in out-of-distribution settings. Codes are available at https://github.com/syr-cn/ReLM.

11:00-12:30 (East Foyer / Virtual)
**Don't Add, don't Miss: Effective Content Preserving Generation from Pre-Selected Text Spans**
*Aviv Slobodkin, Avi Caciularu, Eran Hirsch and Ido Dagan*
The recently introduced Controlled Text Reduction (CTR) task isolates the text generation step within typical summarization-style tasks. It does so by challenging models to generate coherent text conforming to pre-selected content within the input text ("highlights"). This framing enables increased modularity in summarization-like tasks, allowing to couple a single CTR model with various content-selection setups and modules. However, there are currently no reliable CTR models, while the performance of the existing baseline for the task is mediocre, falling short of practical utility. Here, we address this gap by introducing a high-quality, open-source CTR model that tackles two prior key limitations: inadequate enforcement of the content-preservation constraint, and suboptimal silver training data. Addressing these, we amplify the content-preservation constraint in both training, via RL, and inference, via a controlled decoding strategy. Further, we substantially improve the silver training data quality via GPT-4 distillation. Overall, pairing the distilled dataset with the highlight-adherence strategies yields marked gains over the current baseline, of up to 30 ROUGE-L points, providing a reliable CTR model for downstream use.

11:00-12:30 (East Foyer / Virtual)
**Language-Agnostic Bias Detection in Language Models with Bias Probing**
*Abdullatif Köksal, Omer Faruk Yalcin, Ahmet Akbiyik, M. Tahir Kilavuz, Anna Korhonen and Hinrich Schuetze*
Pretrained language models (PLMs) are key components in NLP, but they contain strong social biases. Quantifying these biases is challenging because current methods focusing on fill-the-mask objectives are sensitive to slight changes in input. To address this, we propose a bias probing technique called LABDet, for evaluating social bias in PLMs with a robust and language-agnostic method. For nationality as a case study, we show that LABDet "surfaces" nationality bias by training a classifier on top of a frozen PLM on non-nationality sentiment detection. We find consistent patterns of nationality bias across monolingual PLMs in six languages that align with historical and political context. We also show for English BERT that bias surfaced by LABDet correlates well with bias in the pretraining data; thus, our work is one of the few studies that directly links pretraining data to PLM behavior. Finally, we verify LABDet's reliability and applicability to different templates and languages through an extensive set of robustness checks. We publicly share our code and dataset in https://github.com/akoksal/LABDet.

11:00-12:30 (East Foyer / Virtual)
**Using In-Context Learning to Improve Dialogue Safety**
*Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu and Dilek Hakkani-Tur*
While large neural-based conversational models have become increasingly proficient dialogue agents, recent work has highlighted safety issues with these systems. For example, these systems can be goaded into generating toxic content, often perpetuating social biases or stereotypes. We investigate a retrieval-based approach for reducing bias and toxicity in responses from chatbots. It uses in-context learning to steer a model towards safer generations. Concretely, to generate a response to an unsafe dialogue context, we retrieve demonstrations of safe responses to similar dialogue contexts. We find our method performs competitively with existing approaches to dialogue safety without requiring training. We also show, using automatic and human evaluation, that reductions in toxicity obtained using our approach are not at the cost engagingness or coherency. Finally, we note our method can be used in compliment to existing dialogue safety approaches, such as RLHF.

11:00-12:30 (East Foyer / Virtual)

**Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training**
*Max Müller-Eberstein, Rob van der Goot, Barbara Plank and Ivan Titov*
Representational spaces learned via language modeling are fundamental to Natural Language Processing (NLP), however there has been limited understanding regarding how and when during training various types of linguistic information emerge and interact. Leveraging a novel information theoretic probing suite, which enables direct comparisons of not just task performance, but their representational subspaces, we analyze nine tasks covering syntax, semantics and reasoning, across 2M pre-training steps and five seeds. We identify critical learning phases across tasks and time, during which subspaces emerge, share information, and later disentangle to specialize. Across these phases, syntactic knowledge is acquired rapidly after 0.5% of full training. Continued performance improvements primarily stem from the acquisition of open-domain knowledge, while semantics and reasoning tasks benefit from later boosts to long-range contextualization and higher specialization. Measuring cross-task similarity further reveals that linguistically related tasks share information throughout training, and do so more during the critical phase of learning than before or after. Our findings have implications for model interpretability, multi-task learning, and learning from limited data.

11:00-12:30 (East Foyer / Virtual)
**Methodological Insights in Detecting Subtle Semantic Shifts with Contextualized and Static Language Models**
*Sanne Hoeken, Özge Alacam, Antske Fokkens and Pia Sommerauer*
In this paper, we investigate automatic detection of subtle semantic shifts between social communities of different political convictions in Dutch and English. We perform a methodological study comparing methods using static and contextualized language models. We investigate the impact of specializing contextualized models through fine-tuning on target corpora, word sense disambiguation and sentiment. We furthermore propose a new approach using masked token prediction, that relies on behavioral information, specifically the most probable substitutions, instead of geometrical comparison of representations. Our results show that methods using static models and our masked token prediction method can detect differences in connotation of politically loaded terms, whereas methods that rely on measuring the distance between contextualized representations are not providing clear signals, even in synthetic scenarios of extreme shifts.

11:00-12:30 (East Foyer / Virtual)
**Revisiting Entropy Rate Constancy in Text**
*Vivek Verma, Nicholas Tomlin and Dan Klein*
The uniform information density (UID) hypothesis states that humans tend to distribute information roughly evenly across an utterance or discourse. Early evidence in support of the UID hypothesis came from Genzel and Charniak (2002), which proposed an entropy rate constancy principle based on the probability of English text under $n$-gram language models. We re-evaluate the claims of Genzel and Charniak (2002) with neural language models, failing to find clear evidence in support of entropy rate constancy. We conduct a range of experiments across datasets, model sizes, and languages and discuss implications for the uniform information density hypothesis and linguistic theories of efficient communication more broadly.

11:00-12:30 (East Foyer / Virtual)
**Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models**
*Ziqiao Ma, Jacob Sansom, Run Peng and Joyce Chai*
Large Language Models (LLMs) have generated considerable interest and debate regarding their potential emergence of Theory of Mind (ToM). Several recent inquiries reveal a lack of robust ToM in these models and pose a pressing demand to develop new benchmarks, as current ones primarily focus on different aspects of ToM and are prone to shortcuts and data leakage. In this position paper, we seek to answer two road-blocking questions: (1) How can we taxonomize a holistic landscape of machine ToM? (2) What is a more effective evaluation protocol for machine ToM? Following psychological studies, we taxonomize machine ToM into 7 mental state categories and delineate existing benchmarks to identify under-explored aspects of ToM. We argue for a holistic and situated evaluation of ToM to break ToM into individual components and treat LLMs as an agent who is physically situated in environments and socially situated in interactions with humans. Such situated evaluation provides a more comprehensive assessment of mental states and potentially mitigates the risk of shortcuts and data leakage. We further present a pilot study in a grid world setup as a proof of concept. We hope this position paper can facilitate future research to integrate ToM with LLMs and offer an intuitive means for researchers to better position their work in the landscape of ToM.

11:00-12:30 (East Foyer / Virtual)
**Dolphin: A Challenging and Diverse Benchmark for Arabic NLG**
*El Moatez Billah Nagoudi, AbdelRahim A. Elmadany, Ahmed Oumar El-Shangiti and Muhammad Abdul-Mageed*
We present Dolphin, a novel benchmark that addresses the need for a natural language generation (NLG) evaluation framework dedicated to the wide collection of Arabic languages and varieties. The proposed benchmark encompasses a broad range of 13 different NLG tasks, including dialogue generation, question answering, machine translation, summarization, among others. Dolphin comprises a substantial corpus of 40 diverse and representative public datasets across 50 test splits, carefully curated to reflect real-world scenarios and the linguistic richness of Arabic. It sets a new standard for evaluating the performance and generalization capabilities of Arabic and multilingual models, promising to enable researchers to push the boundaries of current methodologies. We provide an extensive analysis of Dolphin, highlighting its diversity and identifying gaps in current Arabic NLG research. We also offer a public leaderboard that is both interactive and modular and evaluate several Arabic and multilingual models on our benchmark, allowing us to set strong baselines against which researchers can compare.

11:00-12:30 (East Foyer / Virtual)
**Auto-Instruct: Automatic Instruction Generation and Ranking for Black-Box Language Models**
*Zhihan Zhang, Shuohang Wang, Wenhao Yu, Yichong Xu, Dan Iter, Qingkai Zeng, Yang Liu, Chenguang Zhu and Meng Jiang*
Large language models (LLMs) can perform a wide range of tasks by following natural language instructions, without the necessity of task-specific fine-tuning. Unfortunately, the performance of LLMs is greatly influenced by the quality of these instructions, and manually writing effective instructions for each task is a laborious and subjective process. In this paper, we introduce Auto-Instruct, a novel method to automatically improve the quality of instructions provided to LLMs. Our method leverages the inherent generative ability of LLMs to produce diverse candidate instructions for a given task, and then ranks them using a scoring model trained on a variety of 575 existing NLP tasks. In experiments on 118 out-of-domain tasks, Auto-Instruct surpasses both human-written instructions and existing baselines of LLM-generated instructions. Furthermore, our method exhibits notable generalizability even with other LLMs that are not incorporated into its training process.

11:00-12:30 (East Foyer / Virtual)
**Exploiting Contrastive Learning and Numerical Evidence for Confusing Legal Judgment Prediction**
*Leilei Gan, Baokui Li, Kun Kuang, Yating Zhang, Lei Wang, Anh Tuan Luu, Yi Yang and Fei Wu*
Given the fact description text of a legal case, legal judgment prediction (LJP) aims to predict the case's charge, applicable law article, and term of penalty. A core problem of LJP is distinguishing confusing legal cases where only subtle text differences exist. Previous studies fail to distinguish different classification errors with a standard cross-entropy classification loss and ignore the numbers in the fact description for predicting the term of penalty. To tackle these issues, in this work, first, in order to exploit the numbers in legal cases for predicting the

term of penalty of certain charges, we enhance the representation of the fact description with extracted crime amounts which are encoded by a pre-trained numeracy model. Second, we propose a moco-based supervised contrastive learning to learn distinguishable representations and explore the best strategy to construct positive example pairs to benefit all three subtasks of LJP simultaneously. Extensive experiments on real-world datasets show that the proposed method achieves new state-of-the-art results, particularly for confusing legal cases. Ablation studies also demonstrate the effectiveness of each component.

11:00-12:30 (East Foyer / Virtual)
**Definitions Matter: Guiding GPT for Multi-label Classification**
*Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy and Paolo Rosso*
Large language models have recently risen in popularity due to their ability to perform many natural language tasks without requiring any fine-tuning. In this work, we focus on two novel ideas: (1) generating definitions from examples and using them for zero-shot classification, and (2) investigating how an LLM makes use of the definitions. We thoroughly analyze the performance of GPT-3 model for fine-grained multi-label conspiracy theory classification of tweets using zero-shot labeling. In doing so, we asses how to improve the labeling by providing minimal but meaningful context in the form of the definitions of the labels. We compare descriptive noun phrases, human-crafted definitions, introduce a new method to help the model generate definitions from examples, and propose a method to evaluate GPT-3's understanding of the definitions. We demonstrate that improving definitions of class labels has a direct consequence on the downstream classification results.

11:00-12:30 (East Foyer / Virtual)
**Improving Question Generation with Multi-level Content Planning**
*Zehua Xia, Qi Gou, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li and Nguyen Cam-Tu*
This paper addresses the problem of generating questions from a given context and an answer, specifically focusing on questions that require multi-hop reasoning across an extended context. Previous studies have suggested that key phrase selection is essential for question generation (QG), yet it is still challenging to connect such disjointed phrases into meaningful questions, particularly for long context. To mitigate this issue, we propose MultiFactor, a novel QG framework based on multi-level content planning. Specifically, MultiFactor includes two components: FA-Model, which simultaneously selects key phrases and generates full answers, and Q-Model which takes the generated full answer as an additional input to generate questions. Here, full answer generation is introduced to connect the short answer with the selected key phrases, thus forming an answer-aware summary to facilitate QG. Both FA-Model and Q-Model are formalized as simple-yet-effective Phrase-Enhanced Transformers, our joint model for phrase selection and text generation. Experimental results show that our method outperforms strong baselines on two popular QG datasets. Our code is available at https://github.com/zeaver/MultiFactor.

11:00-12:30 (East Foyer / Virtual)
**Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can Large Language Models pass the test?**
*Paola Merlo*
How do we evaluate Large Language Models (LLMs) and determine the aspects and limits of their intelligent behaviour? It is currently conjectured that shortcomings of LLMs in multi-linguality and reasoning are due to a lack of ability to generalize. It has been argued that, instead, humans are better at generalization because they have a tendency at extracting rules from complex data. We propose a method to evaluate LLMs ability to rule-based generalization. When exposed to tests of analytic intelligence, for example the visual RAVEN IQ test, human problem-solvers identify the relevant objects in the picture and their relevant attributes and reason based on rules applied to them. Based on the induced rules, they are able to provide a generalisation and a solution to the test. An analogous language task has recently been proposed (called BLM) for LLM. In this paper, we argue that we can use this task to investigate what linguistic reasoning LLM develop, by asking them to solve some simple variants of the BLM task. We find that current state-of-the-art generative models, such as ChatGPT, can handle the task in the sense that they easily understand the instructions and can provide step-by-step reasoning that shows that it can solve two of the main cognitive hurdles: correspondence finding (object and attribute identification) and item novelty. However, overall they cannot find the correct answer, even with considerable help. In particular, they never identify the structure of the problem, exhibiting, we hypothesize, a lack of goal and subgoal management abilities, an ability that has been argued to measure differential abilities in humans. We argue that this finding supports the usefulness of the task as a method to test the limits and specific properties of generalisation ability in Large Language Models, providing an intrinsic evaluation method inspired by tests of human intelligence.

11:00-12:30 (East Foyer / Virtual)
**Re-Examining Summarization Evaluation across Multiple Quality Criteria**
*Ori Ernst, Ori Shapira, Ido Dagan and Ran Levy*
The common practice for assessing automatic evaluation metrics is to measure the correlation between their induced system rankings and those obtained by reliable human evaluation, where a higher correlation indicates a better metric. Yet, an intricate setting arises when an NLP task is evaluated by multiple Quality Criteria (QCs), like for text summarization where prominent criteria including relevance, consistency, fluency and coherence. In this paper, we challenge the soundness of this methodology when multiple QCs are involved, concretely for the summarization case. First, we show that the allegedly best metrics for certain QCs actually do not perform well, failing to detect even drastic summary corruptions with respect to the considered QC. To explain this, we show that some of the high correlations obtained in the multi-QC setup are spurious. Finally, we propose a procedure that may help detecting this effect. Overall, our findings highlight the need for further investigating metric evaluation methodologies for the multiple-QC case.

11:00-12:30 (East Foyer / Virtual)
**Detrimental Contexts in Open-Domain Question Answering**
*Philhoon Oh and James Thorne*
For knowledge intensive NLP tasks, it has been widely accepted that accessing more information is a contributing factor to improvements in the model's end-to-end performance. However, counter-intuitively, too much context can have a negative impact on the model when evaluated on common question answering (QA) datasets. In this paper, we analyze how passages can have a detrimental effect on retrieve-then-read architectures used in question answering. Our empirical evidence indicates that the current read architecture does not fully leverage the retrieved passages and significantly degrades its performance when using the whole passages compared to utilizing subsets of them. Our findings demonstrate that model accuracy can be improved by 10% on two popular QA datasets by filtering out detrimental passages. Additionally, these outcomes are attained by utilizing existing retrieval methods without further training or data. We further highlight the challenges associated with identifying the detrimental passages. First, even with the correct context, the model can make an incorrect prediction, posing a challenge in determining which passages are most influential. Second, evaluation typically considers lexical matching, which is not robust to variations of correct answers. Despite these limitations, our experimental results underscore the pivotal role of identifying and removing these detrimental passages for the context-efficient retrieve-then-read pipeline.

11:00-12:30 (East Foyer / Virtual)
**An Empirical Study of Multimodal Model Merging**
*Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal and Lijuan Wang*

Model merging (e.g., via interpolation or task arithmetic) fuses multiple models trained on different tasks to generate a multi-task solution. The technique has been proven successful in previous studies, where the models are trained on similar tasks and with the same initialization. In this paper, we expand on this concept to a multimodal setup by merging transformers trained on different modalities. Furthermore, we conduct our study for a novel goal where we can merge vision, language, and cross-modal transformers of a modality-specific architecture to create a parameter-efficient modality-agnostic architecture. Through comprehensive experiments, we systematically investigate the key factors impacting model performance after merging, including initialization, merging mechanisms, and model architectures. We also propose two metrics that assess the distance between weights to be merged and can serve as an indicator of the merging outcomes. Our analysis leads to an effective training recipe for matching the performance of the modality-agnostic baseline (i.e., pre-trained from scratch) via model merging. Our method also outperforms naive merging significantly on various tasks, with improvements of 3% on VQA, 7% on COCO retrieval, 25% on NLVR2, 14% on Flickr30k and 3% on ADE20k.

11:00-12:30 (East Foyer / Virtual)
**Is Explanation the Cure? Misinformation Mitigation in the Short Term and Long Term**
*Yi-Li Hsu, Shih-Chieh Dai, Aiping Xiong and Lun-Wei Ku*
With advancements in natural language processing (NLP) models, automatic explanation generation has been proposed to mitigate misinformation on social media platforms in addition to adding warning labels to identified fake news. While many researchers have focused on generating good explanations, how these explanations can really help humans combat fake news is under-explored. In this study, we compare the effectiveness of a warning label and the state-of- the-art counterfactual explanations generated by GPT-4 in debunking misinformation. In a two-wave, online human-subject study, participants (N = 215) were randomly assigned to a control group in which false contents are shown without any intervention, a warning tag group in which the false claims were labeled, or an explanation group in which the false contents were accompanied by GPT-4 generated explanations. Our results show that both interventions significantly decrease participants' self-reported belief in fake claims in an equivalent manner for the short-term and long-term. We discuss the implications of our findings and directions for future NLP-based misinformation debunking strategies.

11:00-12:30 (East Foyer / Virtual)
**Annotation Sensitivity: Training Data Collection Methods Affect Model Performance**
*Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma and Frauke Kreuter*
When training data are collected from human annotators, the design of the annotation instrument, the instructions given to annotators, the characteristics of the annotators, and their interactions can impact training data. This study demonstrates that design choices made when creating an annotation instrument also impact the models trained on the resulting annotations. We introduce the term annotation sensitivity to refer to the impact of annotation data collection methods on the annotations themselves and on downstream model performance and predictions. We collect annotations of hate speech and offensive language in five experimental conditions of an annotation instrument, randomly assigning annotators to conditions. We then fine-tune BERT models on each of the five resulting datasets and evaluate model performance on a holdout portion of each condition. We find considerable differences between the conditions for 1) the share of hate speech/offensive language annotations, 2) model performance, 3) model predictions, and 4) model learning curves. Our results emphasize the crucial role played by the annotation instrument which has received little attention in the machine learning literature. We call for additional research into how and why the instrument impacts the annotations to inform the development of best practices in instrument design.

11:00-12:30 (East Foyer / Virtual)
**Towards Multilingual Interlinear Morphological Glossing**
*Shu Okabe and François Yvon*
Interlinear Morphological Glosses are annotations produced in the context of language documentation. Their goal is to identify morphs occurring in an L1 sentence and to explicit their function and meaning, with the further support of an associated translation in L2. We study here the task of automatic glossing, aiming to provide linguists with adequate tools to facilitate this process. Our formalisation of glossing uses a latent variable Conditional Random Field (CRF), which labels the L1 morphs while simultaneously aligning them to L2 words. In experiments with several under-resourced languages, we show that this approach is both effective and data-efficient and mitigates the problem of annotating unknown morphs. We also discuss various design choices regarding the alignment process and the selection of features. We finally demonstrate that it can benefit from multilingual (pre-)training, achieving results which outperform very strong baselines.

11:00-12:30 (East Foyer / Virtual)
**Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing**
*Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat and Isabelle Augenstein*
Dual use, the intentional, harmful reuse of technology and scientific artefacts, is an ill-defined problem within the context of Natural Language Processing (NLP). As large language models (LLMs) have advanced in their capabilities and become more accessible, the risk of their intentional misuse becomes more prevalent. To prevent such intentional malicious use, it is necessary for NLP researchers and practitioners to understand and mitigate the risks of their research. Hence, we present an NLP-specific definition of dual use informed by researchers and practitioners in the field. Further, we propose a checklist focusing on dual-use in NLP, that can be integrated into existing conference ethics-frameworks. The definition and checklist are created based on a survey of NLP researchers and practitioners.

11:00-12:30 (East Foyer / Virtual)
**InstructSafety: A Unified Framework for Building Multidimensional and Explainable Safety Detector through Instruction Tuning**
*Zhexin Zhang, Jiale Cheng, Hao Sun, Jiawen Deng and Minlie Huang*
Safety detection has been an increasingly important topic in recent years and it has become even more necessary to develop reliable safety detection systems with the rapid development of large language models. However, currently available safety detection systems have limitations in terms of their versatility and interpretability. In this paper, we first introduce InstructSafety, a safety detection framework that unifies 7 common sub-tasks for safety detection. These tasks are unified into a similar form through different instructions. We then conduct a comprehensive survey of existing safety detection datasets and process 39 human-annotated datasets for instruction tuning. We also construct adversarial samples to enhance the model's robustness. After fine-tuning Flan-T5 on the collected data, we have developed Safety-Flan-T5, a multidimensional and explainable safety detector. We conduct comprehensive experiments on a variety of datasets and tasks, and demonstrate the strong performance of Safety-Flan-T5 in comparison to supervised baselines and served APIs (Perspective API, ChatGPT and InstructGPT). We will release the processed data, fine-tuned Safety-Flan-T5 and related code for public use.

11:00-12:30 (East Foyer / Virtual)
**In-Image Neural Machine Translation with Segmented Pixel Sequence-to-Sequence Model**
*Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo and Bin Wang*
In-Image Machine Translation (IIMT) aims to convert images containing texts from one language to another. Traditional approaches for this task are cascade methods, which utilize optical character recognition (OCR) followed by neural machine translation (NMT) and text rendering. However, the cascade methods suffer from compounding errors of OCR and NMT, leading to a decrease in translation quality. In this paper, we propose an end-to-end model instead of the OCR, NMT and text rendering pipeline. Our neural architecture adopts encoder-decoder

paradigm with segmented pixel sequences as inputs and outputs. Through end-to-end training, our model yields improvements across various dimensions, (i) it achieves higher translation quality by avoiding error propagation, (ii) it demonstrates robustness for out domain data, and (iii) it displays insensitivity to incomplete words. To validate the effectiveness of our method and support for future research, we construct our dataset containing 4M pairs of De-En images and train our end-to-end model. The experimental results show that our approach outperforms both cascade method and current end-to-end model.

11:00-12:30 (East Foyer / Virtual)
**Approximating Two-Layer Feedforward Networks for Efficient Transformers**
*Róbert Csordás, Kazuki Irie and Jürgen Schmidhuber*
How to reduce compute and memory requirements of neural networks (NNs) without sacrificing performance? Many recent works use sparse Mixtures of Experts (MoEs) to build resource-efficient large language models (LMs). Here we introduce several novel perspectives on MoEs, presenting a general framework that *unifies* various methods to *approximate two-layer NNs* (e.g., feedforward blocks of Transformers), including product-key memories (PKMs). Leveraging insights from this framework, we propose methods to improve both MoEs and PKMs. Unlike prior work that compares MoEs with dense baselines under the *compute-equal* condition, our evaluation condition is *parameter-equal*, which is crucial to properly evaluate LMs. We show that our MoEs are competitive with the *dense* Transformer-XL on both the WikiText-103 and enwiki8 datasets at two different scales, while being much more resource efficient. This demonstrates that MoEs are relevant not only to extremely large LMs but also to any-scale resource-efficient LMs. Our code is public.

11:00-12:30 (East Foyer / Virtual)
**CReTIHC: Designing Causal Reasoning Tasks about Temporal Interventions and Hallucinated Confoundings**
*Changwoo Chun, SongEun Lee, Jaehyung Seo and Heuiseok Lim*
Large language models (LLMs) have demonstrated impressive capabilities in natural language processing. However, their ability to establish causal relationships, particularly in the context of temporal interventions and language hallucinations, remains challenging. This paper presents **CReTIHC**, a novel dataset designed to test and enhance the causal reasoning abilities of LLMs. The dataset is constructed using a unique approach that incorporates elements of verbal hallucinations and temporal interventions through the reengineering of existing causal inference datasets. This transformation creates complex scenarios that push LLMs to critically evaluate the information presented and identify cause-and-effect relationships. The CReTIHC dataset serves as a pioneering tool for improving LLM's causal inference capabilities, paving the way for a more nuanced understanding of causal relationships in natural language processing (NLP) tasks. The whole dataset is publicly accessible at: (https://github.com/ChangwooChun/CReTIHC)

11:00-12:30 (East Foyer / Virtual)
**Balaur: Language Model Pretraining with Lexical Semantic Relations**
*Andrei Mircea and Jackie C. K. Cheung*
Lexical semantic relations (LSRs) characterize meaning relationships between words and play an important role in systematic generalization on lexical inference tasks. Notably, several tasks that require knowledge of hypernymy still pose a challenge for pretrained language models (LMs) such as BERT, underscoring the need to better align their linguistic behavior with our knowledge of LSRs. In this paper, we propose Balaur, a model that addresses this challenge by modeling LSRs directly in the LM's hidden states throughout pretraining. Motivating our approach is the hypothesis that the internal representations of LMs can provide an interface to their observable linguistic behavior, and that by controlling one we can influence the other. We validate our hypothesis and demonstrate that Balaur generally improves the performance of large transformer-based LMs on a comprehensive set of hypernymy-informed tasks, as well as on the original LM objective. Code and data are made available at https://github.com/mirandrom/balaur

11:00-12:30 (East Foyer / Virtual)
**Explicit Alignment and Many-to-many Entailment Based Reasoning for Conversational Machine Reading**
*Yangyang Luo, Shiyu Tian, Caixia Yuan and Xiaojie Wang*
Conversational Machine Reading (CMR) requires answering a user's initial question through multi-turn dialogue interactions based on a given document. Although there exist many effective methods, they largely neglected the alignment between the *document* and the *user-provided information*, which significantly affects the intermediate decision-making and subsequent follow-up question generation. To address this issue, we propose a pipeline framework that (1) aligns the aforementioned two sides in an explicit way, (2) makes decisions using a lightweight many-to-many entailment reasoning module, and (3) directly generates follow-up questions based on the document and previously asked questions. Our proposed method achieves state-of-the-art in micro-accuracy and ranks the first place on the public leaderboard of the CMR benchmark dataset ShARC.

11:00-12:30 (East Foyer / Virtual)
**T-Projection: High Quality Annotation Projection for Sequence Labeling Tasks**
*Iker García-Ferrero, Rodrigo Agerri and German Rigau*
In the absence of readily available labeled data for a given sequence labeling task and language, annotation projection has been proposed as one of the possible strategies to automatically generate annotated data. Annotation projection has often been formulated as the task of transporting, on parallel corpora, the labels pertaining to a given span in the source language into its corresponding span in the target language. In this paper we present T-Projection, a novel approach for annotation projection that leverages large pretrained text2text language models and state-of-the-art machine translation technology. T-Projection decomposes the label projection task into two subtasks: (i) A candidate generation step, in which a set of projection candidates using a multilingual T5 model is generated and, (ii) a candidate selection step, in which the generated candidates are ranked based on translation probabilities. We conducted experiments on intrinsic and extrinsic tasks in 5 Indo-European and 8 low-resource African languages. We demostrate that T-projection outperforms previous annotation projection methods by a wide margin. We believe that T-Projection can help to automatically alleviate the lack of high-quality training data for sequence labeling tasks. Code and data are publicly available.

11:00-12:30 (East Foyer / Virtual)
**Towards General Error Diagnosis via Behavioral Testing in Machine Translation**
*Junjie Wu, Lemao Liu and Dit-Yan Yeung*
Behavioral testing offers a crucial means of diagnosing linguistic errors and assessing capabilities of NLP models. However, applying behavioral testing to machine translation (MT) systems is challenging as it generally requires human efforts to craft references for evaluating the translation quality of such systems on newly generated test cases. Existing works in behavioral testing of MT systems circumvent this by evaluating translation quality without references, but this restricts diagnosis to specific types of errors, such as incorrect translation of single numeric or currency words. In order to diagnose general errors, this paper proposes a new Bilingual Translation Pair Generation based Behavior Testing (BTPGBT) framework for conducting behavioral testing of MT systems. The core idea of BTPGBT is to employ a novel bilingual translation pair generation (BTPG) approach that automates the construction of high-quality test cases and their pseudoreferences. Experimental results on various MT systems demonstrate that BTPGBT could provide comprehensive and accurate behavioral testing results for general error diagnosis, which further leads to several insightful findings. Our code and data are available at https: //github.com/wujunjie1998/BTPGBT.

11:00-12:30 (East Foyer / Virtual)
**Towards Detecting Contextual Real-Time Toxicity for In-Game Chat**
*Zachary Yang, Nicolas Grenon-Godbout and Reihaneh Rabbany*
Real-time toxicity detection in online environments poses a significant challenge, due to the increasing prevalence of social media and gaming platforms. We introduce ToxBuster, a simple and scalable model that reliably detects toxic content in real-time for a line of chat by including chat history and metadata. ToxBuster consistently outperforms conventional toxicity models across popular multiplayer games, including Rainbow Six Siege, For Honor, and DOTA 2. We conduct an ablation study to assess the importance of each model component and explore ToxBuster's transferability across the datasets. Furthermore, we showcase ToxBuster's efficacy in post-game moderation, successfully flagging 82.1% of chat-reported players at a precision level of 90.0%. Additionally, we show how an additional 6% of unreported toxic players can be proactively moderated.

11:00-12:30 (East Foyer / Virtual)
**FREDSum: A Dialogue Summarization Corpus for French Political Debates**
*Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter and Michalis Vazirgiannis*
Recent advances in deep learning, and especially the invention of encoder-decoder architectures, have significantly improved the performance of abstractive summarization systems. While the majority of research has focused on written documents, we have observed an increasing interest in the summarization of dialogues and multi-party conversations over the past few years. In this paper, we present a dataset of French political debates for the purpose of enhancing resources for multi-lingual dialogue summarization. Our dataset consists of manually transcribed and annotated political debates, covering a range of topics and perspectives. We highlight the importance of high-quality transcription and annotations for training accurate and effective dialogue summarization models, and emphasize the need for multilingual resources to support dialogue summarization in non-English languages. We also provide baseline experiments using state-of-the-art methods, and encourage further research in this area to advance the field of dialogue summarization. Our dataset will be made publicly available for use by the research community, enabling further advances in multilingual dialogue summarization.

11:00-12:30 (East Foyer / Virtual)
**AutoPlan: Automatic Planning of Interactive Decision-Making Tasks With Large Language Models**
*Siqi Ouyang and Lei Li*
Recent large language models (LLMs) are promising for making decisions in grounded environments. However, LLMs frequently fail in complex decision-making tasks due to the misalignment between the pre-trained knowledge in LLMs and the actual rules in the environment. Existing methods require either costly gradient computation or lengthy in-context demonstrations. In this paper, we propose AutoPlan, an approach to guide LLM-based agents to accomplish interactive decision-making tasks. AutoPlan augments the LLM prompt with a task-solving plan and optimizes it through iterative experience collection and reflection. Our experiments show that AutoPlan, though using no in-context demonstrations, achieves success rates on par with the baselines using human-written demonstrations on ALFWorld and even outperforms them by 8% on HotpotQA. The code is available at https://github.com/owaski/AutoPlan.

11:00-12:30 (East Foyer / Virtual)
**Mitigating Data Imbalance and Representation Degeneration in Multilingual Machine Translation**
*Wen Lai, Alexandra Chronopoulou and Alexander Fraser*
Despite advances in multilingual neural machine translation (MNMT), we argue that there are still two major challenges in this area: data imbalance and representation degeneration. The data imbalance problem refers to the imbalance in the amount of parallel corpora for all language pairs, especially for long-tail languages (i.e., very low-resource languages). The representation degeneration problem refers to the problem of encoded tokens tending to appear only in a small subspace of the full space available to the MNMT model. To solve these two issues, we propose Bi-ACL, a framework which only requires target-side monolingual data and a bilingual dictionary to improve the performance of the MNMT model. We define two modules, named bidirectional autoencoder and bidirectional contrastive learning, which we combine with an online constrained beam search and a curriculum learning sampling strategy. Extensive experiments show that our proposed method is more effective than strong baselines both in long-tail languages and in high-resource languages. We also demonstrate that our approach is capable of transferring knowledge between domains and languages in zero-shot scenarios.

11:00-12:30 (East Foyer / Virtual)
**Are Structural Concepts Universal in Transformer Language Models? Towards Interpretable Cross-Lingual Generalization**
*Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang and Xuanjing Huang*
Large language models (LLMs) have exhibited considerable cross-lingual generalization abilities, whereby they implicitly transfer knowledge across languages. However, the transfer is not equally successful for all languages, especially for low-resource ones, which poses an ongoing challenge. It is unclear whether we have reached the limits of implicit cross-lingual generalization and if explicit knowledge transfer is viable. In this paper, we investigate the potential for explicitly aligning conceptual correspondence between languages to enhance cross-lingual generalization. Using the syntactic aspect of language as a testbed, our analyses of 43 languages reveal a high degree of alignability among the spaces of structural concepts within each language for both encoder-only and decoder-only LLMs. We then propose a meta-learning-based method to learn to align conceptual spaces of different languages, which facilitates zero-shot and few-shot generalization in concept classification and also offers insights into the cross-lingual in-context learning phenomenon. Experiments on syntactic analysis tasks show that our approach achieves competitive results with state-of-the-art methods and narrows the performance gap between languages, particularly benefiting those with limited resources.

11:00-12:30 (East Foyer / Virtual)
**Efficiently Enhancing Zero-Shot Performance of Instruction Following Model via Retrieval of Soft Prompt**
*Seonghyeon Ye, Joel Jang, Doyoung Kim, Yongrae Jo and Minjoon Seo*
Enhancing the zero-shot performance of instruction-following models requires heavy computation, either by scaling the total number of training datasets or the model size. In this work, we explore how retrieval of soft prompts obtained through prompt tuning can efficiently assist hard prompts in zero-shot task generalization. Specifically, we train soft prompt embeddings for each prompt through prompt tuning, store the samples of the training instances mapped with the prompt embeddings, and retrieve the corresponding prompt embedding of the training instance closest to the query instance during inference. While only adding 0.007% additional parameters, retrieval of soft prompt enhances the performance of T0 on unseen tasks by outperforming it on 10 out of 11 datasets as well as improving the mean accuracy of T0 on BIG-bench benchmark by 2.39% points. Also, we report an interesting finding that retrieving source embeddings trained on similar answer choice formats is more important than those on similar task types.

11:00-12:30 (East Foyer / Virtual)
**Closed Boundary Learning for Classification Tasks with the Universum Class**
*Hanzhang Zhou, Zijian Feng and Kezhi Mao*
The Universum class, often known as the *other* class or the*miscellaneous* class, is defined as a collection of samples that do not belong

to any class of interest. It is a typical class that exists in many classification-based tasks in NLP, such as relation extraction, named entity recognition, sentiment analysis, etc. The Universum class exhibits very different properties, namely heterogeneity and lack of representativeness in training data; however, existing methods often treat the Universum class equally with the classes of interest, leading to problems such as overfitting, misclassification, and diminished model robustness. In this work, we propose a closed boundary learning method that applies closed decision boundaries to classes of interest and designates the area outside all closed boundaries in the feature space as the space of the Universum class. Specifically, we formulate closed boundaries as arbitrary shapes, propose the inter-class rule-based probability estimation for the Universum class to cater to its unique properties, and propose a boundary learning loss to adjust decision boundaries based on the balance of misclassified samples inside and outside the boundary. In adherence to the natural properties of the Universum class, our method enhances both accuracy and robustness of classification models, demonstrated by improvements on six state-of-the-art works across three different tasks. Our code is available at https://github.com/hzzhou01/Closed-Boundary-Learning.

11:00-12:30 (East Foyer / Virtual)
**Mitigating Framing Bias with Polarity Minimization Loss**
*Yejin Bang, Nayeon Lee and Pascale Fung*
Framing bias plays a significant role in exacerbating political polarization by distorting the perception of actual events. Media outlets with divergent political stances often use polarized language in their reporting of the same event. We propose a new loss function that encourages the model to minimize the polarity difference between the polarized input articles to reduce framing bias. Specifically, our loss is designed to jointly optimize the model to map polarity ends bidirectionally. Our experimental results demonstrate that incorporating the proposed polarity minimization loss leads to a substantial reduction in framing bias when compared to a BART-based multi-document summarization model. Notably, we find that the effectiveness of this approach is most pronounced when the model is trained to minimize the polarity loss associated with informational framing bias (i.e., skewed selection of information to report).

11:00-12:30 (East Foyer / Virtual)
**ASPIRO: Any-shot Structured Parsing-error-Induced ReprOmpting for Consistent Data-to-Text Generation**
*Martin Vejvar and Yasutaka Fujimoto*
We present ASPIRO, an approach for structured data verbalisation into short template sentences in zero to few-shot settings. Unlike previous methods, our approach prompts Large Language Models (LLMs) to directly produce entity-agnostic templates, rather than relying on LLMs to faithfully copy the given example entities, or validating/crafting the templates manually. We incorporate LLM re-prompting, triggered by algorithmic parsing checks, as well as the PARENT metric induced consistency validation to identify and rectify template generation problems in real-time. ASPIRO, compared to direct LLM output, averages 66% parsing error rate reduction in generated verbalisations of RDF triples on the DART dataset. Our best 5-shot text-davinci-003 setup, scoring BLEU of 50.62, METEOR of 45.16, BLEURT of 0.82, NUBIA of 0.87, and PARENT of 0.8962 on the Rel2Text dataset, competes effectively with recent fine-tuned pretrained language models.

11:00-12:30 (East Foyer / Virtual)
**Adaptive Hinge Balance Loss for Document-Level Relation Extraction**
*Jize Wang, Xinyi Le, Xiaodi Peng and Cailian Chen*
Document-Level Relation Extraction aims at predicting relations between entities from multiple sentences. A common practice is to select multi-label classification thresholds to decide whether a relation exists between an entity pair. However, in the document-level task, most entity pairs do not express any relations, resulting in a highly imbalanced distribution between positive and negative classes. We argue that the imbalance problem affects threshold selection and may lead to incorrect "no-relation" predictions. In this paper, we propose to down-weight the easy negatives by utilizing a distance between the classification threshold and the predicted score of each relation. Our novel Adaptive Hinge Balance Loss measures the difficulty of each relation class with the distance, putting more focus on hard, misclassified relations, i.e. the minority positive relations. Experiment results on Re-DocRED demonstrate the superiority of our approach over other balancing methods. Source codes are available at https://github.com/Jize-W/HingeABL.

11:00-12:30 (East Foyer / Virtual)
**The Truth, The Whole Truth, and Nothing but the Truth: A New Benchmark Dataset for Hebrew Text Credibility Assessment**
*Ben Hagag and Reut Tsarfaty*
In the age of information overload, it is more important than ever to discern fact from fiction. From the internet to traditional media, we are constantly confronted with a deluge of information, much of which comes from politicians and other public figures who wield significant influence. In this paper, we introduce HeTrue: a new, publicly available dataset for evaluating the credibility of statements made by Israeli public figures and politicians. This dataset consists of 1021 statements, manually annotated by Israeli professional journalists, for their credibility status. Using this corpus, we set out to assess whether the credibility of statements can be predicted based on the text alone. To establish a baseline, we compare text-only methods with others using additional data like metadata, context, and evidence. Furthermore, we develop several credibility assessment models, including a feature-based model that utilizes linguistic features, and state-of-the-art transformer-based models with contextualized embeddings from a pre-trained encoder. Empirical results demonstrate improved performance when models integrate statement and context, outperforming those relying on the statement text alone. Our best model, which also integrates evidence, achieves a 48.3 F1 Score, suggesting that HeTrue is a challenging benchmark, calling for further work on this task.

11:00-12:30 (East Foyer / Virtual)
**The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation**
*Dung Manh Nguyen, Le Hai Nam, Anh T. V. Dau, Anh Minh Nguyen, Khanh Nghiem, Jin L.C. Guo and Nghi D. Q. Bui*
We present The Vault, an open-source dataset of high quality code-text pairs in multiple programming languages for training large language models to understand and generate code. We propose methods for thoroughly extracting samples that use both rules and deep learning to ensure that they contain high-quality pairs of code and text, resulting in a dataset of 43 million high-quality code-text pairs. We thoroughly evaluated this dataset and discovered that when used to train common code language models (such as CodeT5, CodeBERT, and CodeGen), it outperforms the same models train on other datasets such as CodeSearchNet. These evaluations included common coding tasks such as code generation, code summarization, and code search. The Vault can be used by researchers and practitioners to train a wide range of big language models that understand code. Alternatively, researchers can use our data cleaning methods and scripts to improve their own datasets. We anticipate that using The Vault to train large language models will improve their ability to understand and generate code, propelling AI research and software development forward. We are releasing our source code and a framework to make it easier for others to replicate our results.

11:00-12:30 (East Foyer / Virtual)
**Don't waste a single annotation: improving single-label classifiers through soft labels**
*Ben Peng Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva and Xingyi Song*
In this paper, we address the limitations of the common data annotation and training methods for objective single-label classification tasks. Typically, when annotating such tasks annotators are only asked to provide a single label for each sample and annotator disagreement is discarded when a final hard label is decided through majority voting. We challenge this traditional approach, acknowledging that determining

the appropriate label can be difficult due to the ambiguity and lack of context in the data samples. Rather than discarding the information from such ambiguous annotations, our soft label method makes use of them for training. Our findings indicate that additional annotator information, such as confidence, secondary label and disagreement, can be used to effectively generate soft labels. Training classifiers with these soft labels then leads to improved performance and calibration on the hard label test set.

11:00-12:30 (East Foyer / Virtual)
**NERetrieve: Dataset for Next Generation Named Entity Recognition and Retrieval**
*Uri Katz, Matan Vetzler, Amir David Nissan Cohen and Yoav Goldberg*
Recognizing entities in texts is a central need in many information-seeking scenarios, and indeed, Named Entity Recognition (NER) is arguably one of the most successful examples of a widely adopted NLP task and corresponding NLP technology. Recent advances in large language models (LLMs) appear to provide effective solutions (also) for NER tasks that were traditionally handled with dedicated models, often matching or surpassing the abilities of the dedicated models. Should NER be considered a solved problem? We argue to the contrary: the capabilities provided by LLMs are not the end of NER research, but rather an exciting beginning. They allow taking NER to the next level, tackling increasingly more useful, and increasingly more challenging, variants. We present three variants of the NER task, together with a dataset to support them. The first is a move towards more fine-grained—and intersectional—entity types. The second is a move towards zero-shot recognition and extraction of these fine-grained types based on entity-type labels. The third, and most challenging, is the move from the recognition setup to a novel retrieval setup, where the query is a zero-shot entity type, and the expected result is all the sentences from a large, pre-indexed corpus that contain entities of these types, and their corresponding spans. We show that all of these are far from being solved. We provide a large, silver-annotated corpus of 4 million paragraphs covering 500 entity types, to facilitate research towards all of these three goals.

11:00-12:30 (East Foyer / Virtual)
**Emergent Inabilities? Inverse Scaling Over the Course of Pretraining**
*James Michaelov and Ben Bergen*
Does inverse scaling only occur as a function of model size, or can it also occur over the course of training? We carry out an exploratory study investigating whether the performance of language models on specific tasks can decrease (while general performance remains high) during training on the language modeling task. We find 8 tasks on which Pythia 12B (Biderman et al., 2023) shows decreased performance over the course of training. Five of these tasks (TruthfulQA-MC1, TruthfulQA-MC2, Hindsight Neglect, Memo Trap, and Pattern Match Suppression) additionally show a consistent relationship whereby larger language models show a greater decrease in performance the more they are trained, despite showing standard (positive) scaling overall. This highlights the importance of testing performance at all relevant benchmarks any time models are trained on additional data, even if their overall performance improves.

11:00-12:30 (East Foyer / Virtual)
**INGENIOUS: Using Informative Data Subsets for Efficient Pre-Training of Language Models**
*H S V N S Kowndinya Renduchintala, Krishnateja Killamsetty, Sumit Bhatia, Milan Aggarwal, Ganesh Ramakrishnan, Rishabh K Iyer and Balaji Krishnamurthy*
A salient characteristic of pre-trained language models (PTLMs) is a remarkable improvement in their generalization capability and emergence of new capabilities with increasing model capacity and pre-training dataset size. Consequently, we are witnessing the development of enormous models pushing the state-of-the-art. It is, however, imperative to realize that this inevitably leads to prohibitively long training times, extortionate computing costs, and a detrimental environmental impact. Significant efforts are underway to make PTLM training more efficient through innovations in model architectures, training pipelines, and loss function design, with scant attention being paid to optimizing the utility of training data. The key question that we ask is whether it is possible to train PTLMs by employing only highly informative subsets of the training data while maintaining downstream performance? Building upon the recent progress in informative data subset selection, we show how we can employ submodular optimization to select highly representative subsets of the training corpora and demonstrate that the proposed framework can be applied to efficiently train multiple PTLMs (BERT, BioBERT, GPT-2) using only a fraction of data. Further, we perform a rigorous empirical evaluation to show that the resulting models achieve up to $\sim$ 99% of the performance of the fully-trained models. We made our framework publicly available at https://github.com/Efficient-AI/ingenious.

11:00-12:30 (East Foyer / Virtual)
**Unsupervised Candidate Answer Extraction through Differentiable Masker-Reconstructor Model**
*Zhuoer Wang, Yicheng Wang, Ziwei Zhu and James Caverlee*
Question generation is a widely used data augmentation approach with extensive applications, and extracting qualified candidate answers from context passages is a critical step for most question generation systems. However, existing methods for candidate answer extraction are reliant on linguistic rules or annotated data that face the partial annotation issue and challenges in generalization. To overcome these limitations, we propose a novel unsupervised candidate answer extraction approach that leverages the inherent structure of context passages through a Differentiable Masker-Reconstructor (DMR) Model with the enforcement of self-consistency for picking up salient information tokens. We curated two datasets with exhaustively-annotated answers and benchmark a comprehensive set of supervised and unsupervised candidate answer extraction methods. We demonstrate the effectiveness of the DMR model by showing its performance is superior among unsupervised methods and comparable to supervised methods.

11:00-12:30 (East Foyer / Virtual)
**Citance-Contextualized Summarization of Scientific Papers**
*Shahbaz Syed, Ahmad Dawar Hakimi, Khalid Al-Khatib and Martin Potthast*
Current approaches to automatic summarization of scientific papers generate informative summaries in the form of abstracts. However, abstracts are not intended to show the relationship between a paper and the references cited in it. We propose a new contextualized summarization approach that can generate an informative summary conditioned on a given sentence containing the citation of a reference (a so-called "citance"). This summary outlines content of the cited paper relevant to the citation location. Thus, our approach extracts and models the citances of a paper, retrieves relevant passages from cited papers, and generates abstractive summaries tailored to each citance. We evaluate our approach using **Webis-Context-SciSumm-2023**, a new dataset containing 540K computer science papers and 4.6M citances therein.

11:00-12:30 (East Foyer / Virtual)
**Sources of Hallucination by Large Language Models on Inference Tasks**
*Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson and Mark Steedman*
Large Language Models (LLMs) are claimed to be capable of Natural Language Inference (NLI), necessary for applied tasks like question answering and summarization. We present a series of behavioral studies on several LLM families (LLaMA, GPT-3.5, and PaLM) which probe their behavior using controlled experiments. We establish two biases originating from pretraining which predict much of their behavior, and show that these are major sources of hallucination in generative LLMs. First, memorization at the level of sentences: we show that, regardless of the premise, models falsely label NLI test samples as entailing when the hypothesis is attested in training data, and that entities are used as "indices' to access the memorized data. Second, statistical patterns of usage learned at the level of corpora: we further show a similar effect when the premise predicate is less frequent than that of the hypothesis in the training data, a bias following from previous studies. We

demonstrate that LLMs perform significantly worse on NLI test samples which do not conform to these biases than those which do, and we offer these as valuable controls for future LLM evaluation.

11:00-12:30 (East Foyer / Virtual)
**Novel Slot Detection With an Incremental Setting**
*Chen Liang, Hongliang Li, Changhao Guan, Qingbin Liu, Jian Liu, Jinan Xu and Zhe Zhao*
Current dialogue systems face diverse user requests and rapid change domains, making quickly adapt to scenarios with previous unseen slot types become a major challenge. Recently, researchers have introduced novel slot detection (NSD) to discover potential new types. However, dialogue system with NSD does not bring practical improvements due to the system still cannot handle novel slots in subsequent interactions. In this paper, we define incremental novel slot detection (INSD), which separates the dialogue system to deal with novel types as two major phrases: 1) model discovers unknown slots, 2) training model to possess the capability to handle new classes. We provide an effective model to extract novel slots with set prediction strategy and propose a query-enhanced approach to overcome catastrophic forgetting during the process of INSD. We construct two INSD datasets to evaluate our method and experimental results show that our approach exhibits superior performance.

11:00-12:30 (East Foyer / Virtual)
**Prompt-Based Editing for Text Style Transfer**
*Guoqing Luo, Yu Tong Han, Lili Mou and Mauajama Firdaus*
Prompting approaches have been recently explored in text style transfer, where a textual prompt is used to query a pretrained language model (PLM) to generate style-transferred texts word by word in an autoregressive manner. However, such a generation process is less controllable and early prediction errors may affect future word predictions. In this paper, we propose a prompt-based editing approach to text style transfer. Specifically, we prompt a PLM for style classification and use the classification probability to compute a style score. Then, we perform discrete search with word-level editing to maximize a comprehensive scoring function for the style-transfer task. In this way, we transform a prompt-based generation problem into a classification one, which does not suffer from the error accumulation problem and is more controllable than the autoregressive generation of sentences. In our experiments, we performed both automatic and human evaluation on three style-transfer benchmark datasets, and show that our approach largely outperforms the existing systems that have 20 times more parameters. Additional empirical analyses further demonstrate the effectiveness of our approach.

11:00-12:30 (East Foyer / Virtual)
**Women Wearing Lipstick: Measuring the Bias Between an Object and Its Related Gender**
*Ahmed Sabir and Lluís Padró*
In this paper, we investigate the impact of objects on gender bias in image captioning systems. Our results show that only gender-specific objects have a strong gender bias (e.g., women-lipstick). In addition, we propose a visual semantic-based gender score that measures the degree of bias and can be used as a plug-in for any image captioning system. Our experiments demonstrate the utility of the gender score, since we observe that our score can measure the bias relation between a caption and its related gender; therefore, our score can be used as an additional metric to the existing Object Gender Co-Occ approach.

11:00-12:30 (East Foyer / Virtual)
**GRACE: Discriminator-Guided Chain-of-Thought Reasoning**
*Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee and Lu Wang*
In the context of multi-step reasoning, e.g., with chain-of-thought, language models (LMs) can easily assign a high likelihood to incorrect steps. As a result, decoding strategies that optimize for solution likelihood often yield incorrect solutions. To address this issue, we propose Guiding chain-of-thought ReAsoning with a CorrectnEss Discriminator (GRACE), a stepwise decoding approach that steers the decoding process towards producing correct reasoning steps. GRACE employs a discriminator trained with a contrastive loss over correct and incorrect steps, which is used during decoding to score next-step candidates based on their correctness. Importantly, GRACE only requires sampling from the LM, without the need for LM training or fine-tuning. Using models from FLAN-T5 and LLaMA families, we evaluate GRACE over four math and two symbolic reasoning tasks, where it exhibits substantial performance gains compared to greedy decoding, verifiers, and self-consistency in most settings. When further combined with self-consistency, GRACE outperforms all the baselines by sizeable margins. Human and LLM evaluations over GSM8K show that GRACE not only improves the final answer accuracy but also the correctness of the intermediate reasoning.

11:00-12:30 (East Foyer / Virtual)
**Entity-Based Evaluation of Political Bias in Automatic Summarization**
*Karen Zhou and Chenhao Tan*
Growing literature has shown that NLP systems may encode social biases; however, the *political* bias of summarization models remains relatively unknown. In this work, we use an entity replacement method to investigate the portrayal of politicians in automatically generated summaries of news articles. We develop an entity-based computational framework to assess the sensitivities of several extractive and abstractive summarizers to the politicians Donald Trump and Joe Biden. We find consistent differences in these summaries upon entity replacement, such as reduced emphasis of Trump's presence in the context of the same article and a more individualistic representation of Trump with respect to the collective US government (i.e., administration). These summary dissimilarities are most prominent when the entity is heavily featured in the source article. Our characterization provides a foundation for future studies of bias in summarization and for normative discussions on the ideal qualities of automatic summaries.

11:00-12:30 (East Foyer / Virtual)
**Leveraging GPT-4 for Automatic Translation Post-Editing**
*Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadalla and Arul Menezes*
While Neural Machine Translation (NMT) represents the leading approach to Machine Translation (MT), the outputs of NMT models still require translation post-editing to rectify errors and enhance quality under critical settings. In this work, we formalize the task of direct translation post-editing with Large Language Models (LLMs) and explore the use of GPT-4 to automatically post-edit NMT outputs across several language pairs. Our results demonstrate that GPT-4 is adept at translation post-editing, producing meaningful and trustworthy edits to translations that help improve its general quality as well as remove different classes of major errors in translations. In particular, human evaluations on assessing edit trustworthiness show that GPT-4 exhibits a large improvement over the prior state-of-the-art LLM. Notably, we improve upon state-of-the-art performance on WMT-22 English-Chinese, English-German, Chinese-English and German-English language pairs using GPT-4 based post-editing, as evaluated by state-of-the-art MT quality metrics. However, we also show that GPT-4 could produce hallucinated edits, thereby urging caution in its use as an expert translation post-editor.

11:00-12:30 (East Foyer / Virtual)
**BYOC: Personalized Few-Shot Classification with Co-Authored Class Descriptions**
*Arth Bohra, Govert Verkes, Artem Harutyunyan, Pascal Weinberger and Giovanni Campagna*

Text classification is a well-studied and versatile building block for many NLP applications. Yet, existing approaches require either large annotated corpora to train a model with or, when using large language models as a base, require carefully crafting the prompt as well as using a long context that can fit many examples. As a result, it is not possible for end-users to build classifiers for themselves. To address this issue, we propose a novel approach to few-shot text classification using an LLM. Rather than few-shot examples, the LLM is prompted with descriptions of the salient features of each class. These descriptions are coauthored by the user and the LLM interactively: while the user annotates each few-shot example, the LLM asks relevant questions that the user answers. Examples, questions, and answers are summarized to form the classification prompt. Our experiments show that our approach yields high accuracy classifiers, within 79% of the performance of models trained with significantly larger datasets while using only 1% of their training sets. Additionally, in a study with 30 participants, we show that end-users are able to build classifiers to suit their specific needs. The personalized classifiers show an average accuracy of 90%, which is 15% higher than the state-of-the-art approach.

11:00-12:30 (East Foyer / Virtual)
**Hallucination Detection for Grounded Instruction Generation**
*Lingjun Zhao, Khanh Xuan Nguyen and Hal Daumé III*
We investigate the problem of generating instructions to guide humans to navigate in simulated residential environments. A major issue with current models is hallucination: they generate references to actions or objects that are inconsistent with what a human follower would perform or encounter along the described path. We develop a model that detects these hallucinated references by adopting a model pre-trained on a large corpus of image-text pairs, and fine-tuning it with a contrastive loss that separates correct instructions from instructions containing synthesized hallucinations. Our final model outperforms several baselines, including using word probability estimated by the instruction-generation model, and supervised models based on LSTM and Transformer.

11:00-12:30 (East Foyer / Virtual)
**HoneyBee: Progressive Instruction Finetuning of Large Language Models for Materials Science**
*Yu Song, Santiago Miret, Huan Zhang and Bang Liu*
We propose an instruction-based process for trustworthy data curation in materials science (MatSci-Instruct), which we then apply to finetune a LLaMa-based language model targeted for materials science (HoneyBee). MatSci-Instruct helps alleviate the scarcity of relevant, high-quality materials science textual data available in the open literature, and HoneyBee is the first billion-parameter language model specialized to materials science. In MatSci-Instruct we improve the trustworthiness of generated data by prompting multiple commercially available large language models for generation with an Instructor module (e.g. Chat-GPT) and verification from an independent Verifier module (e.g. Claude). Using MatSci-Instruct, we construct a dataset of multiple tasks and measure the quality of our dataset along multiple dimensions, including accuracy against known facts, relevance to materials science, as well as completeness and reasonableness of the data. Moreover, we iteratively generate more targeted instructions and instruction-data in a finetuning-evaluation-feedback loop leading to progressively better performance for our finetuned HoneyBee models. Our evaluation on the MatSci-NLP benchmark shows HoneyBee's outperformance of existing language models on materials science tasks and iterative improvement in successive stages of instruction-data refinement. We study the quality of HoneyBee's language modeling through automatic evaluation and analyze case studies to further understand the model's capabilities and limitations. Our code and relevant datasets are publicly available at https://github.com/BangLab-UdeM-Mila/NLP4MatSci-HoneyBee.

11:00-12:30 (East Foyer / Virtual)
**Investigating Online Community Engagement through Stancetaking**
*Jai Aggarwal, Brian Diep, Julia Watson and Suzanne Stevenson*
Much work has explored lexical and semantic variation in online communities, and drawn connections to community identity and user engagement patterns. Communities also express identity through the sociolinguistic concept of stancetaking. Large-scale computational work on stancetaking has explored community similarities in their preferences for stance markers – words that serve to indicate aspects of a speaker's stance – without considering the stance-relevant properties of the contexts in which stance markers are used. We propose representations of stance contexts for 1798 Reddit communities and show how they capture community identity patterns distinct from textual or marker similarity measures. We also relate our stance context representations to broader inter- and intra-community engagement patterns, including cross-community posting patterns and social network properties of communities. Our findings highlight the strengths of using rich properties of stance as a way of revealing community identity and engagement patterns in online multi-community spaces.

11:00-12:30 (East Foyer / Virtual)
**On Surgical Fine-tuning for Language Encoders**
*Abhilasha Lodha, Gayatri Vyankatesh Belapurkar, Saloni Chalkapurkar, Yuanming Tao, Reshmi Ghosh, Samyadeep Basu, Dmitrii M Petrov and Soundararajan Srinivasan*
Fine-tuning all the layers of a pre-trained neural language encoder (either using all the parameters or using parameter-efficient methods) is often the de-facto way of adapting it to a new task. We show evidence that for different downstream language tasks, fine-tuning only a subset of layers is sufficient to obtain performance that is close to and often better than fine-tuning all the layers in the language encoder. We propose an efficient metric based on the diagonal of the Fisher information matrix (FIM score), to select the candidate layers for selective fine-tuning. We show, empirically on GLUE and SuperGLUE tasks and across distinct language encoders, that this metric can effectively select layers leading to a strong downstream performance. Our work highlights that task-specific information corresponding to a given downstream task is often localized within a few layers, and tuning only those is sufficient for strong performance. Additionally, we demonstrate the robustness of the FIM score to rank layers in a manner that remains constant during the optimization process.

11:00-12:30 (East Foyer / Virtual)
**Linguistic Compression in Single-Sentence Human-Written Summaries**
*Fangcong Yin and Marten van Schijndel*
Summarizing texts involves significant cognitive efforts to compress information. While advances in automatic summarization systems have drawn attention from the NLP and linguistics communities to this topic, there is a lack of computational studies of linguistic patterns in human-written summaries. This work presents a large-scale corpus study of human-written single-sentence summaries. We analyzed the linguistic compression patterns from source documents to summaries at different granularities, and we found that summaries are generally written with morphological expansion, increased lexical diversity, and similar positional arrangements of specific words compared to the source across different genres. We also studied how linguistic compressions of different factors affect reader judgments of quality through a human study, with the results showing that the use of morphological and syntactic changes by summary writers matches reader preferences while lexical diversity and word specificity preferences are not aligned between summary writers and readers.

11:00-12:30 (East Foyer / Virtual)
**Measuring and Mitigating Constraint Violations of In-Context Learning for Utterance-to-API Semantic Parsing**
*Shufan Wang, Sébastien Jean, Sailik Sengupta, James Gung, Nikolaos Pappas and Yi Zhang*
In executable task-oriented semantic parsing, the system aims to translate users' utterances in natural language to machine-interpretable programs (API calls) that can be executed according to pre-defined API specifications. With the popularity of Large Language Models (LLMs),

in-context learning offers a strong baseline for such scenarios, especially in data-limited regimes. However, LLMs are known to hallucinate and therefore pose a formidable challenge in constraining generated content. Thus, it remains uncertain if LLMs can effectively perform task-oriented utterance-to-API generation, where respecting the API's structural and task-specific constraints is crucial. In this work, we seek to measure, analyze and mitigate such constraints violations. First, we identify the categories of various constraints in obtaining API-semantics from task-oriented utterances, and define fine-grained metrics that complement traditional ones. Second, we leverage these metrics to conduct a detailed error analysis of constraints violations seen in state-of-the-art LLMs, which motivates us to investigate two popular mitigation strategies– Semantic-Retrieval of Demonstrations (SRD) and API-aware Constrained Decoding (API-CD). Our experiments show that these strategies are effective at reducing constraints violations and improving the quality of the generated API calls, but require careful consideration given their implementation complexity and latency.

11:00-12:30 (East Foyer / Virtual)
**Learn Your Tokens: Word-Pooled Tokenization for Language Modeling**
*Avijit Thawani, Saurabh Ghanekar, Xiaoyuan Zhu and Jay Pujara*
Language models typically tokenize text into subwords, using a deterministic, hand-engineered heuristic of combining characters into longer surface-level strings such as 'ing' or whole words. Recent literature has repeatedly shown the limitations of such a tokenization strategy, particularly for documents not written in English and for representing numbers. On the other extreme, byte/character-level language models are much less restricted but suffer from increased sequence description lengths and a subsequent quadratic expansion in self-attention computation. Recent attempts to compress and limit these context lengths with fixed size convolutions is helpful but completely ignores the word boundary. This paper considers an alternative 'learn your tokens' scheme which utilizes the word boundary to pool bytes/characters into word representations, which are fed to the primary language model, before again decoding individual characters/bytes per word in parallel. We find that our moderately expressive and moderately fast end-to-end tokenizer outperform by over '300%' both subwords and byte/character models over the intrinsic language modeling metric of next-word prediction across datasets. It particularly outshines on rare words, outperforming by a factor of 30! We extensively study the language modeling setup for all three categories of tokenizers and theoretically analyze how our end-to-end models can also be a strong trade-off in efficiency and robustness.

11:00-12:30 (East Foyer / Virtual)
**PARROT: Zero-Shot Narrative Reading Comprehension via Parallel Reading**
*Chao Zhao, Anvesh Rao Vijjini and Snigdha Chaturvedi*
Narrative comprehension is a challenging task that requires a deep understanding of the foundational elements of narratives. Acquiring this skill requires extensive annotated data. To mitigate the burden of data annotation, we present Parrot, a zero-shot approach for narrative reading comprehension through parallel reading, which involves two parallel narratives that tell the same story. By leveraging one narrative as a source of supervision signal to guide the understanding of the other, Parrot abstracts the textual content and develops genuine narrative understanding. Evaluation conducted on two narrative comprehension benchmarks demonstrates that Parrot surpasses previous zero-shot approaches and achieves comparable performance to fully supervised models. The code will be available at https://github.com/zhaochaocs/Parrot.

11:00-12:30 (East Foyer / Virtual)
**SYMPTOMIFY: Transforming Symptom Annotations with Language Model Knowledge Harvesting**
*Bosung Kim and Ndapa Nakashole*
Given the high-stakes nature of healthcare decision-making, we aim to improve the efficiency of human annotators rather than replacing them with fully automated solutions. We introduce a new comprehensive resource, SYMPTOMIFY, a dataset of annotated vaccine adverse reaction reports detailing individual vaccine reactions. The dataset, consisting of over 800k reports, surpasses previous datasets in size. Notably, it features reasoning-based explanations alongside background knowledge obtained via language model knowledge harvesting. We evaluate performance across various methods and learning paradigms, paving the way for future comparisons and benchmarking.

11:00-12:30 (East Foyer / Virtual)
**Are NLP Models Good at Tracing Thoughts: An Overview of Narrative Understanding**
*Lixing Zhu, Runcong Zhao, Lin Gui and Yulan He*
Narrative understanding involves capturing the author's cognitive processes, providing insights into their knowledge, intentions, beliefs, and desires. Although large language models (LLMs) excel in generating grammatically coherent text, their ability to comprehend the author's thoughts remains uncertain. This limitation hinders the practical applications of narrative understanding. In this paper, we conduct a comprehensive survey of narrative understanding tasks, thoroughly examining their key features, definitions, taxonomy, associated datasets, training objectives, evaluation metrics, and limitations. Furthermore, we explore the potential of expanding the capabilities of modularized LLMs to address novel narrative understanding tasks. By framing narrative understanding as the retrieval of the author's imaginative cues that outline the narrative structure, our study introduces a fresh perspective on enhancing narrative comprehension.

11:00-12:30 (East Foyer / Virtual)
**DelucionQA: Detecting Hallucinations in Domain-specific Question Answering**
*Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez and Zhe Feng*
Hallucination is a well-known phenomenon in text generated by large language models (LLMs). The existence of hallucinatory responses is found in almost all application scenarios e.g., summarization, question-answering (QA) etc. For applications requiring high reliability (e.g., customer-facing assistants), the potential existence of hallucination in LLM-generated text is a critical problem. The amount of hallucination can be reduced by leveraging information retrieval to provide relevant background information to the LLM. However, LLMs can still generate hallucinatory content for various reasons (e.g., prioritizing its parametric knowledge over the context, failure to capture the relevant information from the context, etc.). Detecting hallucinations through automated methods is thus paramount. To facilitate research in this direction, we introduce a sophisticated dataset, DelucionQA, that captures hallucinations made by retrieval-augmented LLMs for a domain-specific QA task. Furthermore, we propose a set of hallucination detection methods to serve as baselines for future works from the research community. Analysis and case study are also provided to share valuable insights on hallucination phenomena in the target scenario.

11:00-12:30 (East Foyer / Virtual)
**Adversarial Robustness for Large Language NER models using Disentanglement and Word Attributions**
*Xiaomeng Jin, Bhanukiran Vinzamuri, Sriram Venkatapathy, Heng Ji and Pradeep Natarajan*
Large language models (LLM's) have been widely used for several applications such as question answering, text classification and clustering. While the preliminary results across the aforementioned tasks looks promising, recent work has dived deep into LLM's performing poorly for complex Named Entity Recognition (NER) tasks in comparison to fine-tuned pre-trained language models (PLM's). To enhance wider adoption of LLM's, our paper investigates the robustness of such LLM NER models and its instruction fine-tuned variants to adversarial attacks. In particular, we propose a novel attack which relies on disentanglement and word attribution techniques where the former aids in learning an embedding capturing both entity and non-entity influences separately, and the latter aids in identifying important words across both components. This is in stark contrast to most techniques which primarily leverage non-entity words for perturbations limiting the space

being explored to synthesize effective adversarial examples. Adversarial training results based on our method improves the F1 score over original LLM NER model by 8% and 18% on CoNLL-2003 and Ontonotes 5.0 datasets respectively.

### Industry 5

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**BUSTER: a "BUSiness Transaction Entity Recognition" dataset**
*Andrea Zugarini, Andrew Zamai, Marco Ernandes and Leonardo Rigutini*
Albeit Natural Language Processing has seen major breakthroughs in the last few years, transferring such advances into real-world business cases can be challenging. One of the reasons resides in the displacement between popular benchmarks and actual data. Lack of supervision, unbalanced classes, noisy data and long documents often affect real problems in vertical domains such as finance, law and health. To support industry-oriented research, we present BUSTER, a BUSiness Transaction Entity Recognition dataset. The dataset consists of 3779 manually annotated documents on financial transactions. We establish several baselines exploiting both general-purpose and domain-specific language models. The best performing model is also used to automatically annotate 6196 documents, which we release as an additional silver corpus to BUSTER.

## Lunch
12:30-14:30 - Location: Resort World

## Business Meeting: All Attendees Welcome
13:45-14:30 - Location: East & Central

## Session 7: Plenary - Keynote Speaker: Emily Mower Provost
14:30-15:30 - Location: East & Central

## Coffee Break
15:30-16:00 - Location: West Foyer

## Session 8: Plenary - Panel
16:00-17:00 - Location: East & Central

## Social Event - 18:30-23:45

# Main Conference: Sunday, December 10, 2023

# Session 9: Oral & Poster - 09:00-10:30

## Semantics 1

09:00-10:30 (East Ballroom)

09:00-09:15 (East Ballroom)
**WiCE: Real-World Entailment for Claims in Wikipedia**
*Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez and Greg Durrett*
Textual entailment models are increasingly applied in settings like fact-checking, presupposition verification in question answering, or summary evaluation. However, these represent a significant domain shift from existing entailment datasets, and models underperform as a result. We propose WiCE, a new fine-grained textual entailment dataset built on natural claim and evidence pairs extracted from Wikipedia. In addition to standard claim-level entailment, WiCE provides entailment judgments over sub-sentence units of the claim, and a minimal subset of evidence sentences that support each subclaim. To support this, we propose an automatic claim decomposition strategy using GPT-3.5 which we show is also effective at improving entailment models' performance on multiple datasets at test time. Finally, we show that real claims in our dataset involve challenging verification and retrieval problems that existing models fail to address.

09:15-09:30 (East Ballroom)
**Understanding Computational Models of Semantic Change: New Insights from the Speech Community**
*Filip Miletić, Anne Przewozny-Desriaux and Ludovic Tanguy*
We investigate the descriptive relevance of widely used semantic change models in linguistic descriptions of present-day speech communities. We focus on the sociolinguistic issue of contact-induced semantic shifts in Quebec English, and analyze 40 target words using type-level and token-level word embeddings, empirical linguistic properties, and – crucially – acceptability ratings and qualitative remarks by 15 speakers from Montreal. Our results confirm the overall relevance of the computational approaches, but also highlight practical issues and the complementary nature of different semantic change estimates. To our knowledge, this is the first study to substantively engage with the speech community being described using semantic change models.

09:30-09:45 (East Ballroom)
**What do Deck Chairs and Sun Hats Have in Common? Uncovering Shared Properties in Large Concept Vocabularies**
*Amit Gajbhiye, Zied Bouraoui, Na Li, Usashi Chatterjee, Luis Espinosa-Anke and Steven Schockaert*
Concepts play a central role in many applications. This includes settings where concepts have to be modelled in the absence of sentence context. Previous work has therefore focused on distilling decontextualised concept embeddings from language models. But concepts can be modelled from different perspectives, whereas concept embeddings typically mostly capture taxonomic structure. To address this issue, we propose a strategy for identifying what different concepts, from a potentially large concept vocabulary, have in common with others. We then represent concepts in terms of the properties they share with the other concepts. To demonstrate the practical usefulness of this way of modelling concepts, we consider the task of ultra-fine entity typing, which is a challenging multi-label classification problem. We show that by augmenting the label set with shared properties, we can improve the performance of the state-of-the-art models for this task.

09:45-10:00 (East Ballroom)
**AdaSent: Efficient Domain-Adapted Sentence Embeddings for Few-Shot Classification**
*Yongxin Huang, Kexin Wang, Sourav Dutta, Raj Nath Patel, Goran Glavaš and Iryna Gurevych*
Recent work has found that few-shot sentence classification based on pre-trained Sentence Encoders (SEs) is efficient, robust, and effective. In this work, we investigate strategies for domain-specialization in the context of few-shot sentence classification with SEs. We first establish that unsupervised Domain-Adaptive Pre-Training (DAPT) of a base Pre-trained Language Model (PLM) (i.e., not an SE) substantially improves the accuracy of few-shot sentence classification by up to 8.4 points. However, applying DAPT on SEs, on the one hand, disrupts the effects of their (general-domain) Sentence Embedding Pre-Training (SEPT). On the other hand, applying general-domain SEPT on top of a domain-adapted base PLM (i.e., after DAPT) is effective but inefficient, since the computationally expensive SEPT needs to be executed on top of a DAPT-ed PLM of each domain. As a solution, we propose AdaSent, which decouples SEPT from DAPT by training a SEPT adapter on the base PLM. The adapter can be inserted into DAPT-ed PLMs from any domain. We demonstrate AdaSent's effectiveness in extensive experiments on 17 different few-shot sentence classification datasets. AdaSent matches or surpasses the performance of full SEPT on DAPT-ed PLM, while substantially reducing the training costs. The code for AdaSent is available.

10:00-10:15 (East Ballroom)
**Ditto: A Simple and Efficient Approach to Improve Sentence Embeddings**
*Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, Chong Deng, Hai Yu, Jiaqing Liu, Yukun Ma and Chong Zhang*
Prior studies diagnose the anisotropy problem in sentence representations from pre-trained language models, e.g., BERT, without fine-tuning. Our analysis reveals that the sentence embeddings from BERT suffer from a bias towards uninformative words, limiting the performance in semantic textual similarity (STS) tasks. To address this bias, we propose a simple and efficient unsupervised approach, Diagonal Attention Pooling (Ditto), which weights words with model-based importance estimations and computes the weighted average of word representations from pre-trained models as sentence embeddings. Ditto can be easily applied to any pre-trained language model as a postprocessing operation. Compared to prior sentence embedding approaches, Ditto does not add parameters nor requires any learning. Empirical evaluations demonstrate that our proposed Ditto can alleviate the anisotropy problem and improve various pre-trained models on the STS benchmarks.

10:15-10:30 (East Ballroom)
**Connecting degree and polarity: An artificial language learning study**
*Lisa Bylinina, Alexey Tikhonov and Ekaterina Garmash*
We investigate a new linguistic generalisation in pre-trained language models (taking BERT Devlin et al. 2019 as a case study). We focus on degree modifiers (expressions like slightly, very, rather, extremely) and test the hypothesis that the degree expressed by a modifier (low, medium or high degree) is related to the modifier's sensitivity to sentence polarity (whether it shows preference for affirmative or negative sentences or neither). To probe this connection, we apply the Artificial Language Learning experimental paradigm from psycholinguistics to a neural language model. Our experimental results suggest that BERT generalizes in line with existing linguistic observations that relate de- gree semantics to polarity sensitivity, including the main one: low degree semantics is associated with preference towards positive polarity.

## Sentiment/Stylistic Analysis

09:00-10:30 (Central 1 Ballroom)

09:00-09:15 (Central 1 Ballroom)
**Argument-based Detection and Classification of Fallacies in Political Debates**
*Pierpaolo Goffredo, Mariana Chaves Espinoza, Serena Villata and Elena Cabrio*
Fallacies are arguments that employ faulty reasoning. Given their persuasive and seemingly valid nature, fallacious arguments are often used in political debates. Employing these misleading arguments in politics can have detrimental consequences for society, since they can lead to inaccurate conclusions and invalid inferences from the public opinion and the policymakers. Automatically detecting and classifying fallacious arguments represents therefore a crucial challenge to limit the spread of misleading or manipulative claims and promote a more informed and healthier political discourse. Our contribution to address this challenging task is twofold. First, we extend the ElecDeb60To16 dataset of U.S. presidential debates annotated with fallacious arguments, by incorporating the most recent Trump-Biden presidential debate. We include updated token-level annotations, incorporating argumentative components (i.e., claims and premises), the relations between these components (i.e., support and attack), and six categories of fallacious arguments (i.e., Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogans). Second, we perform the twofold task of fallacious argument detection and classification by defining neural network architectures based on Transformers models, combining text, argumentative features, and engineered features. Our results show the advantages of complementing transformer-generated text representations with non-text features.

09:15-09:30 (Central 1 Ballroom)
**Why Should This Article Be Deleted? Transparent Stance Detection in Multilingual Wikipedia Editor Discussions**
*Lucie-Aimée Kaffee, Arnav Arora and Isabelle Augenstein*
The moderation of content on online platforms is usually non-transparent. On Wikipedia, however, this discussion is carried out publicly and editors are encouraged to use the content moderation policies as explanations for making moderation decisions. Currently, only a few comments explicitly mention those policies – 20% of the English ones, but as few as 2% of the German and Turkish comments. To aid in this process of understanding how content is moderated, we construct a novel multilingual dataset of Wikipedia editor discussions along with their reasoning in three languages. The dataset contains the stances of the editors (keep, delete, merge, comment), along with the stated reason, and a content moderation policy, for each edit decision. We demonstrate that stance and corresponding reason (policy) can be predicted jointly with a high degree of accuracy, adding transparency to the decision-making process. We release both our joint prediction models and the multilingual content moderation dataset for further research on automated transparent content moderation.

09:30-09:45 (Central 1 Ballroom)
**Identification of Multimodal Stance Towards Frames of Communication**
*Maxwell Weinzierl and Sanda Harabagiu*
Frames of communication are often evoked in multimedia documents. When an author decides to add an image to a text, one or both of the modalities may evoke a communication frame. Moreover, when evoking the frame, the author also conveys her/his stance towards the frame. Until now, determining if the author is in favor of, against or has no stance towards the frame was performed automatically only when processing texts. This is due to the absence of stance annotations on multimedia documents. In this paper we introduce MMVax-Stance, a dataset of 11,300 multimedia documents retrieved from social media, which have stance annotations towards 113 different frames of communication. This dataset allowed us to experiment with several models of multimedia stance detection, which revealed important interactions between texts and images in the inference of stance towards communication frames. When inferring the text/image relations, a set of 46,606 synthetic examples of multimodal documents with known stance was generated. This greatly impacted the quality of identifying multimedia stance, yielding an improvement of 20% in F1-score.

09:45-10:00 (Central 1 Ballroom)
**EXPLAIN, EDIT, GENERATE: Rationale-Sensitive Counterfactual Data Augmentation for Multi-hop Fact Verification**
*Yingjie Zhu, Jiasheng Si, Yibo Zhao, Haiyang Zhu, Deyu Zhou and Yulan He*
Automatic multi-hop fact verification task has gained significant attention in recent years. Despite impressive results, these well-designed models perform poorly on out-of-domain data. One possible solution is to augment the training data with counterfactuals, which are generated by minimally altering the causal features of the original data. However, current counterfactual data augmentation techniques fail to handle multi-hop fact verification due to their incapability to preserve the complex logical relationships within multiple correlated texts. In this paper, we overcome this limitation by developing a rationale-sensitive method to generate linguistically diverse and label-flipping counterfactuals while preserving logical relationships. In specific, the diverse and fluent counterfactuals are generated via an Explain-Edit-Generate architecture. Moreover, the checking and filtering modules are proposed to regularize the counterfactual data with logical relations and flipped labels. Experimental results show that the proposed approach outperforms the SOTA baselines and can generate linguistically diverse counterfactual data without disrupting their logical relationships.

10:00-10:15 (Central 1 Ballroom)
**Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimoda Emotion Recognition**
*Dongyuan Li, Yusong Wang, Kotaro Funakoshi and Manabu Okumura*
Multimodal emotion recognition aims to recognize emotions for each utterance from multiple modalities, which has received increasing attention for its application in human-machine interaction. Current graph-based methods fail to simultaneously depict global contextual features and local diverse uni-modal features in a dialogue. Furthermore, with the number of graph layers increasing, they easily fall into over-smoothing. In this paper, we propose a method for joint modality fusion and graph contrastive learning for multimodal emotion recognition (Joyful), where multimodality fusion, contrastive learning, and emotion recognition are jointly optimized. Specifically, we first design a new multimodal fusion mechanism that can provide deep interaction and fusion between the global contextual and uni-modal specific features. Then, we introduce a graph contrastive learning framework with inter- and intra-view contrastive losses to learn more distinguishable representations for samples with different sentiments. Extensive experiments on three benchmark datasets indicate that Joyful achieved state-of-the-art (SOTA) performance compared with all baselines. Code is released on Github (https://anonymous.4open.science/r/MERC-7F88).

10:15-10:30 (Central 1 Ballroom)
**Can Authorship Representation Learning Capture Stylistic Features?**
*Nicholas Oliver Andrews, Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera-Soto and Marcus Bishop*
Automatically disentangling an author's style from the content of their writing is a longstanding and possibly insurmountable problem in computational linguistics. At the same time, the availability of large text corpora furnished with author labels has recently enabled learning authorship representations in a purely data-driven manner for authorship attribution, a task that ostensibly depends to a greater extent on encoding writing style than encoding content. However, success on this surrogate task does not ensure that such representations capture writing style since authorship could also be correlated with other latent variables, such as topic. In an effort to better understand the nature of the information these representations convey, and specifically to validate the hypothesis that they chiefly encode writing style, we systematically

probe these representations through a series of targeted experiments. The results of these experiments suggest that representations learned for the surrogate authorship prediction task are indeed sensitive to writing style. As a consequence, authorship representations may be expected to be robust to certain kinds of data shift, such as topic drift over time. Additionally, our findings may open the door to downstream applications that require stylistic representations, such as style transfer.

## Speech & Multimodality 1

09:00-10:30 (Central 3 Ballroom)

09:00-09:15 (Central 3 Ballroom)
**A Video Is Worth 4096 Tokens: Verbalize Story Videos To Understand Them In Zero Shot**
*Aanisha Bhattacharyya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah and Changyou Chen*
Multimedia content, such as advertisements and story videos, exhibit a rich blend of creativity and multiple modalities. They incorporate elements like text, visuals, audio, and storytelling techniques, employing devices like emotions, symbolism, and slogans to convey meaning. There is a dearth of large annotated training datasets in the multimedia domain hindering the development of supervised learn-ing models with satisfactory performance for real-world applications. On the other hand, the rise of large language models (LLMs) has witnessed remarkable zero-shot performance in various natural language processing (NLP) tasks, such as emotion classification, question-answering, and topic classification. To leverage such advanced techniques to bridge this performance gap in multimedia understanding, we propose verbalizing long videos to generate their descriptions in natural language, followed by performing video-understanding tasks on the generated story as opposed to the original video. Through extensive experiments on fifteen video-understanding tasks, we demonstrate that our method, despite being zero-shot, achieves significantly better results than supervised baselines for video understanding. Furthermore, to alleviate a lack of story understanding benchmarks, we publicly release the first dataset on a crucial task in computational social science on persuasion strategy identification.

09:15-09:30 (Central 3 Ballroom)
**Balance Act: Mitigating Hubness in Cross-Modal Retrieval with Query and Gallery Banks**
*Yimu Wang, Xiangru Jian and Bo Xue*
In this work, we present a post-processing solution to address the hubness problem in cross-modal retrieval, a phenomenon where a small number of gallery data points are frequently retrieved, resulting in a decline in retrieval performance. We first theoretically demonstrate the necessity of incorporating both the gallery and query data for addressing hubness as hubs always exhibit high similarity with gallery and query data. Second, building on our theoretical results, we propose a novel framework, Dual Bank Normalization (DBNorm). While previous work has attempted to alleviate hubness by only utilizing the query samples, DBNorm leverages two banks constructed from the query and gallery samples to reduce the occurrence of hubs during inference. Next, to complement DBNorm, we introduce two novel methods, dual inverted softmax and dual dynamic inverted softmax, for normalizing similarity based on the two banks. Specifically, our proposed methods reduce the similarity between hubs and queries while improving the similarity between non-hubs and queries. Finally, we present extensive experimental results on diverse language-grounded benchmarks, including text-image, text-video, and text-audio, demonstrating the superior performance of our approaches compared to previous methods in addressing hubness and boosting retrieval performance.

09:30-09:45 (Central 3 Ballroom)
**Three Stream Based Multi-level Event Contrastive Learning for Text-Video Event Extraction**
*Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Dehai Min, Yongrui Chen and Guilin Qi*
Text-video based multimodal event extraction refers to identifying event information from the given text-video pairs. Existing methods predominantly utilize video appearance features (VAF) and text sequence features (TSF) as input information. Some of them employ contrastive learning to align VAF with the event types extracted from TSF. However, they disregard the motion representations in videos and the optimization of contrastive objective could be misguided by the background noise from RGB frames. We observe that the same event triggers correspond to similar motion trajectories, which are hardly affected by the background noise. Moviated by this, we propose a Three Stream Multimodal Event Extraction framework (TSEE) that simultaneously utilizes the features of text sequence and video appearance, as well as the motion representations to enhance the event extraction capacity. Firstly, we extract the optical flow features (OFF) as motion representations from videos to incorporate with VAF and TSF. Then we introduce a Multi-level Event Contrastive Learning module to align the embedding space between OFF and event triggers, as well as between event triggers and types. Finally, a Dual Querying Text module is proposed to enhance the interaction between modalities. Experimental results show that TSEE outperforms the state-of-the-art methods, which demonstrates its superiority.

09:45-10:00 (Central 3 Ballroom)
**Reading Order Matters: Information Extraction from Visually-rich Documents by Token Path Prediction**
*Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang and Tao Gui*
Recent advances in multimodal pre-trained models have significantly improved information extraction from visually-rich documents (VrDs), in which named entity recognition (NER) is treated as a sequence-labeling task of predicting the BIO entity tags for tokens, following the typical setting of NLP. However, BIO-tagging scheme relies on the correct order of model inputs, which is not guaranteed in real-world NER on scanned VrDs where text are recognized and arranged by OCR systems. Such reading order issue hinders the accurate marking of entities by BIO-tagging scheme, making it impossible for sequence-labeling methods to predict correct named entities. To address the reading order issue, we introduce Token Path Prediction (TPP), a simple prediction head to predict entity mentions as token sequences within documents. Alternative to token classification, TPP models the document layout as a complete directed graph of tokens, and predicts token paths within the graph as entities. For better evaluation of VrD-NER systems, we also propose two revised benchmark datasets of NER on scanned documents which can reflect real-world scenarios. Experiment results demonstrate the effectiveness of our method, and suggest its potential to be a universal solution to various information extraction tasks on documents.

10:00-10:15 (Central 3 Ballroom)
**MultiTurnCleanup: A Benchmark for Multi-Turn Spoken Conversational Transcript Cleanup**
*Hua Shen, Vicky Zayats, Johann C Rocholl, Daniel David Walker and Dirk Padfield*
Current disfluency detection models focus on individual utterances each from a single speaker. However, numerous discontinuity phenomena in spoken conversational transcripts occur across multiple turns, which can not be identified by disfluency detection models. This study addresses these phenomena by proposing an innovative Multi-Turn Cleanup task for spoken conversational transcripts and collecting a new dataset, MultiTurnCleanup. We design a data labeling schema to collect the high-quality dataset and provide extensive data analysis. Furthermore, we leverage two modeling approaches for experimental evaluation as benchmarks for future research.

10:15-10:30 (Central 3 Ballroom)

**Large Language Models and Multimodal Retrieval for Visual Word Sense Disambiguation**
*Anastasia Kritharoula, Maria Lymperaiou and Giorgos Stamou*
Visual Word Sense Disambiguation (VWSD) is a novel challenging task with the goal of retrieving an image among a set of candidates, which better represents the meaning of an ambiguous word within a given context. In this paper, we make a substantial step towards unveiling this interesting task by applying a varying set of approaches. Since VWSD is primarily a text-image retrieval task, we explore the latest transformer-based methods for multimodal retrieval. Additionally, we utilize Large Language Models (LLMs) as knowledge bases to enhance the given phrases and resolve ambiguity related to the target word. We also study VWSD as a unimodal problem by converting to text-to-text and image-to-image retrieval, as well as question-answering (QA), to fully explore the capabilities of relevant models. To tap into the implicit knowledge of LLMs, we experiment with Chain-of-Thought (CoT) prompting to guide explainable answer generation. On top of all, we train a learn to rank (LTR) model in order to combine our different modules, achieving competitive ranking results. Extensive experiments on VWSD demonstrate valuable insights to effectively drive future directions.

## Summarization

09:00-10:30 (West 1 Ballroom)

---

09:00-09:15 (West 1 Ballroom)
**Instructive Dialogue Summarization with Query Aggregations**
*Bin Wang, Zhengyuan Liu and Nancy F. Chen*
Conventional dialogue summarization methods directly generate summaries and do not consider user's specific interests. This poses challenges in cases where the users are more focused on particular topics or aspects. With the advancement of instruction-finetuned language models, we introduce instruction-tuning to dialogues to expand the capability set of dialogue summarization models. To overcome the scarcity of instructive dialogue summarization data, we propose a three-step approach to synthesize high-quality query-based summarization triples. This process involves summary-anchored query generation, query filtering and query-based summary generation. By training a unified model called InstructDS (Instructive Dialogue Summarization) on three summarization datasets with multi-purpose instructive triples, we expand the capability of dialogue summarization models. We evaluate our method on four datasets, including dialogue summarization and dialogue reading comprehension. Experimental results show that our approach outperforms the state-of-the-art models and even models with larger sizes. Additionally, our model exhibits higher generalizability and faithfulness, as confirmed by human subjective evaluations.

09:15-09:30 (West 1 Ballroom)
**Investigating Efficiently Extending Transformers for Long Input Summarization**
*Jason Phang, Yao Zhao and Peter J Liu*
While large pretrained Transformer models have proven highly capable at tackling natural language tasks, handling long sequence inputs still poses a significant challenge. One such task is long input summarization, where inputs are longer than the maximum input context of most models. Through an extensive set of experiments, we investigate what model architectural changes and pretraining paradigms most efficiently adapt a pretrained Transformer for long input summarization. We find that a staggered, block-local Transformer with global encoder tokens strikes a good balance of performance and efficiency, and that an additional pretraining phase on long sequences meaningfully improves downstream summarization performance. Based on our findings, we introduce PEGASUS-X, an extension of the PEGASUS model with additional long input pretraining to handle inputs of up to 16K tokens, which achieves strong performance on long input summarization tasks comparable with much larger models.

09:30-09:45 (West 1 Ballroom)
**Zero-shot Faithfulness Evaluation for Text Summarization with Foundation Language Model**
*Qi Jia, Siyu Ren, Yizhu Liu and Kenny Q. Zhu*
Despite tremendous improvements in natural language generation, summarization models still suffer from the unfaithfulness issue. Previous work evaluates faithfulness either using models trained on the other tasks or in-domain synthetic data, or prompting a large model such as ChatGPT. This paper proposes to do zero-shot faithfulness evaluation simply with a moderately-sized foundation language model. We introduce a new metric FFLM, which is a combination of probability changes based on the intuition that prefixing a piece of text that is consistent with the output will increase the probability of predicting the output. Experiments show that FFLM performs competitively with or even outperforms ChatGPT on both inconsistency detection and faithfulness rating with 24x fewer parameters. FFLM also achieves improvements over other strong baselines.

09:45-10:00 (West 1 Ballroom)
**Indicative Summarization of Long Discussions**
*Shahbaz Syed, Dominik Schwabe, Khalid Al Khatib and Martin Potthast*
Online forums encourage the exchange and discussion of different stances on many topics. Not only do they provide an opportunity to present one's own arguments, but may also gather a broad cross-section of others' arguments. However, the resulting long discussions are difficult to overview. This paper presents a novel unsupervised approach using large language models (LLMs) to generating indicative summaries for long discussions that basically serve as tables of contents. Our approach first clusters argument sentences, generates cluster labels as abstractive summaries, and classifies the generated cluster labels into argumentation frames resulting in a two-level summary. Based on an extensively optimized prompt engineering approach, we evaluate 19 LLMs for generative cluster labeling and frame classification. To evaluate the usefulness of our indicative summaries, we conduct a purpose-driven user study via a new visual interface called **Discussion Explorer**: It shows that our proposed indicative summaries serve as a convenient navigation tool to explore long discussions.

10:00-10:15 (West 1 Ballroom)
**Promoting Topic Coherence and Inter-Document Consorts in Multi-Document Summarization via Simplicial Complex and Sheaf Graph**
*Yash Kumar Atri, Arun Iyer, Tanmoy Chakraborty and Vikram Goyal*
Multi-document Summarization (MDS) characterizes compressing information from multiple source documents to its succinct summary. An ideal summary should encompass all topics and accurately model cross-document relations expounded upon in the source documents. However, existing systems either impose constraints on the length of tokens during the encoding or falter in capturing the intricate cross-document relationships. These limitations impel the systems to produce summaries that are non-factual and unfaithful, thereby imparting an unfair comprehension of the topic to the readers. To counter these limitations and promote the information equivalence between the source document and generated summary, we propose FIBER, a novel encoder-decoder model that uses pre-trained BART to comprehensively analyze linguistic nuances, simplicial complex layer to apprehend inherent properties that transcend pairwise associations and sheaf graph attention to effectively capture the heterophilic properties. We benchmark FIBER with eleven baselines over four widely-used MDS datasets – Multi-news, CQASumm, DUC and Opinosis, and show that FIBER achieves consistent performance improvement across all the evaluation metrics

(syntactical, semantical and faithfulness). We corroborate these improvements further through qualitative human evaluation.

10:15-10:30 (West 1 Ballroom)
**Length Does Matter: Summary Length can Bias Summarization Metrics**
*Xiaobo Guo and Soroush Vosoughi*
Establishing the characteristics of an effective summary is a complicated and often subjective endeavor. Consequently, the development of metrics for the summarization task has become a dynamic area of research within natural language processing. In this paper, we reveal that existing summarization metrics exhibit a bias toward the length of generated summaries. Our thorough experiments, conducted on a variety of datasets, metrics, and models, substantiate these findings. The results indicate that most metrics tend to favor longer summaries, even after accounting for other factors. To address this issue, we introduce a Bayesian normalization technique that effectively diminishes this bias. We demonstrate that our approach significantly improves the concordance between human annotators and the majority of metrics in terms of summary coherence.

## Machine Learning for NLP

09:00-10:30 (West 2 Ballroom)

---

09:00-09:15 (West 2 Ballroom)
**Explicit Planning Helps Language Models in Logical Reasoning**
*Hongyu Zhao, Kangrui Wang, Mo Yu and Hongyuan Mei*
Language models have been shown to perform remarkably well on a wide range of natural language processing tasks. In this paper, we propose LEAP, a novel system that uses language models to perform multi-step logical reasoning and incorporates explicit planning into the inference procedure. Explicit planning enables the system to make more informed reasoning decisions at each step by looking ahead into their future effects. Moreover, we propose a training strategy that safeguards the planning process from being led astray by spurious features. Our full system significantly outperforms other competing methods on multiple standard datasets. When using small T5 models as its core selection and deduction components, our system performs competitively compared to GPT-3 despite having only about 1B parameters (i.e., 175 times smaller than GPT-3). When using GPT-3.5, it significantly outperforms chain-of-thought prompting on the challenging PrOntoQA dataset. We have conducted extensive empirical studies to demonstrate that explicit planning plays a crucial role in the system's performance.

09:15-09:30 (West 2 Ballroom)
**Where to start? Analyzing the potential value of intermediate models**
*Leshem Choshen, Elad Venezian, Shachar Don-Yehiya, Noam Slonim and Yoav Katz*
Previous studies observed that finetuned models may be better base models than the vanilla pretrained model. Such a model, finetuned on some source dataset, may provide a better starting point for a new finetuning process on a desired target dataset. Here, we perform a systematic analysis of this *intertraining* scheme, over a wide range of English classification tasks. Surprisingly, our analysis suggests that the potential intertraining gain can be analyzed *independently* for the target dataset under consideration, and for a base model being considered as a starting point. Hence, a performant model is generally strong, even if its training data was not aligned with the target dataset. Furthermore, we leverage our analysis to propose a practical and efficient approach to determine if and how to select a base model in real-world settings. Last, we release an updating ranking of best models in the HuggingFace hub per architecture.

09:30-09:45 (West 2 Ballroom)
**Fair Text Classification with Wasserstein Independence**
*Thibaud Leteno, Antoine Gourru, Charlotte Laclau, Rémi Emonet and Christophe Gravier*
Group fairness is a central research topic in text classification, where reaching fair treatment between sensitive groups (e.g. women vs. men) remains an open challenge. This paper presents a novel method for mitigating biases in neural text classification, agnostic to the model architecture. Considering the difficulty to distinguish fair from unfair information in a text encoder, we take inspiration from adversarial training to induce Wasserstein independence between representations learned to predict our target label and the ones learned to predict some sensitive attribute. Our approach provides two significant advantages. Firstly, it does not require annotations of sensitive attributes in both testing and training data. This is more suitable for real-life scenarios compared to existing methods that require annotations of sensitive attributes at train time. Secondly, our approach exhibits a comparable or better fairness-accuracy trade-off compared to existing methods.

09:45-10:00 (West 2 Ballroom)
**Improving Bias Mitigation through Bias Experts in Natural Language Understanding**
*Eojin Jeon, Mingyu Lee, Juhyeong Park, Yeachan Kim, Wing-Lam Mok and SangKeun Lee*
Biases in the dataset often enable the model to achieve high performance on in-distribution data, while poorly performing on out-of-distribution data. To mitigate the detrimental effect of the bias on the networks, previous works have proposed debiasing methods that down-weight the biased examples identified by an auxiliary model, which is trained with explicit bias labels. However, finding a type of bias in datasets is a costly process. Therefore, recent studies have attempted to make the auxiliary model biased without the guidance (or annotation) of bias labels, by constraining the model's training environment or the capability of the model itself. Despite the promising debiasing results of recent works, the multi-class learning objective, which has been naively used to train the auxiliary model, may harm the bias mitigation effect due to its regularization effect and competitive nature across classes. As an alternative, we propose a new debiasing framework that introduces binary classifiers between the auxiliary model and the main model, coined bias experts. Specifically, each bias expert is trained on a binary classification task derived from the multi-class classification task via the One-vs-Rest approach. Experimental results demonstrate that our proposed strategy improves the bias identification ability of the auxiliary model. Consequently, our debiased model consistently outperforms the state-of-the-art on various challenge datasets.

10:00-10:15 (West 2 Ballroom)
**DSI++: Updating Transformer Memory with New Documents**
*Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Jinfeng Rao, Marc Najork, Emma Strubell and Donald Metzler*
Differentiable Search Indices (DSIs) encode a corpus of documents in the parameters of a model and use the same model to map queries directly to relevant document identifiers. Despite the solid performance of DSI models, successfully deploying them in scenarios where document corpora change with time is an open problem. In this work, we introduce DSI++, a continual learning challenge for DSI with the goal of continuously indexing new documents while being able to answer queries related to both previously and newly indexed documents. Across different model scales and document identifier representations, we show that continual indexing of new documents leads to considerable forgetting of previously indexed documents. We also hypothesize and verify that the model experiences forgetting events during training, leading to unstable learning. To mitigate these issues, we investigate two approaches. The first focuses on modifying the training dynamics. Flatter minima implicitly alleviates forgetting, so we explicitly optimize for flatter loss basins and show that the model stably memorizes

more documents (+12%). Next, we introduce a parametric memory to generate pseudo-queries for documents and supplement them during incremental indexing to prevent forgetting for the retrieval task. Extensive experiments on a novel continual indexing benchmark based on Natural Questions demonstrate that our proposed solution mitigates the forgetting in DSI++ by a significant margin and improves the average Hits@10 by +21.1% over competitive baselines.

10:15-10:30 (West 2 Ballroom)
**Translate-and-Test Transfer Learning for Cross-Lingual Text Classification**
*Inigo Jauregi Unanue, Gholamreza Haffari and Massimo Piccardi*
Cross-lingual text classification leverages text classifiers trained in a high-resource language to perform text classification in other languages with no or minimal finetuning (zero/few-shots cross-lingual transfer). Nowadays, cross-lingual text classifiers are typically built on large-scale, multilingual language models (LMs) pretrained on a variety of languages of interest. However, the performance of these models vary significantly across languages and classification tasks, suggesting that the superposition of the language modelling and classification tasks is not always effective. For this reason, in this paper we propose revisiting the classic translate-and-test pipeline to neatly separate the translation and classification stages. The proposed approach couples 1) a neural machine translator translating from the targeted language to a high-resource language, with 2) a text classifier trained in the high-resource language, but the neural machine translator generates soft translations to permit end-to-end backpropagation during fine-tuning of the pipeline. Extensive experiments have been carried out over three cross-lingual text classification datasets (XNLI, MLDoc and MultiEURLEX), with the results showing that the proposed approach has significantly improved performance over a competitive baseline.

## Syntax, Parsing and their Applications

09:00-10:30 (West 3 Ballroom)

---

09:00-09:15 (West 3 Ballroom)
**Linear-Time Modeling of Linguistic Structure: An Order-Theoretic Perspective**
*Tianyu Liu, Afra Amini, Mrinmaya Sachan and Ryan Cotterell*
Tasks that model the relation between pairs of tokens in a string are a vital part of understanding natural language. Such tasks, in general, require exhaustive pair-wise comparisons of tokens, thus having a quadratic runtime complexity in the length of the string. We show that these exhaustive comparisons can be avoided, and, moreover, the complexity of such tasks can be reduced to linear by casting the relation between tokens as a partial order over the string. Our method predicts real numbers for each token in a string in parallel and sorts the tokens accordingly, resulting in total orders of the tokens in the string. Each total order implies a set of arcs oriented from smaller to greater tokens, sorted by their predicted numbers. The intersection of total orders results in a partial order over the set of tokens in the string, which is then decoded into a directed graph representing the desired linguistic structure. Our experiments on dependency parsing and coreference resolution show that our method achieves state-of-the-art or comparable performance. Moreover, the linear complexity and parallelism of our method double the speed of graph-based coreference resolution models, and bring a 10-times speed-up over graph-based dependency parsers.

09:15-09:30 (West 3 Ballroom)
**4 and 7-bit Labeling for Projective and Non-Projective Dependency Trees**
*Carlos Gómez-Rodríguez, Diego Roca and David Vilares*
We introduce an encoding for parsing as sequence labeling that can represent any projective dependency tree as a sequence of 4-bit labels, one per word. The bits in each word's label represent (1) whether it is a right or left dependent, (2) whether it is the outermost (left/right) dependent of its parent, (3) whether it has any left children and (4) whether it has any right children. We show that this provides an injective mapping from trees to labels that can be encoded and decoded in linear time. We then define a 7-bit extension that represents an extra plane of arcs, extending the coverage to almost full non-projectivity (over 99.9% empirical arc coverage). Results on a set of diverse treebanks show that our 7-bit encoding obtains substantial accuracy gains over the previously best-performing sequence labeling encodings.

09:30-09:45 (West 3 Ballroom)
**Syntactic Substitutability as Unsupervised Dependency Syntax**
*Jasper Jian and Siva Reddy*
Syntax is a latent hierarchical structure which underpins the robust and compositional nature of human language. In this work, we explore the hypothesis that syntactic dependencies can be represented in language model attention distributions and propose a new method to induce these structures theory-agnostically. Instead of modeling syntactic relations as defined by annotation schemata, we model a more general property implicit in the definition of dependency relations, syntactic substitutability. This property captures the fact that words at either end of a dependency can be substituted with words from the same category. Substitutions can be used to generate a set of syntactically invariant sentences whose representations are then used for parsing. We show that increasing the number of substitutions used improves parsing accuracy on natural data. On long-distance subject-verb agreement constructions, our method achieves 79.5% recall compared to 8.9% using a previous method. Our method also provides improvements when transferred to a different parsing setup, demonstrating that it generalizes.

09:45-10:00 (West 3 Ballroom)
**Structural generalization in COGS: Supertagging is (almost) all you need**
*Alban Petit, Caio Filippo Corro and François Yvon*
In many Natural Language Processing applications, neural networks have been found to fail to generalize on out-of-distribution examples. In particular, several recent semantic parsing datasets have put forward important limitations of neural networks in cases where compositional generalization is required. In this work, we extend a neural graph-based parsing framework in several ways to alleviate this issue, notably: (1) the introduction of a supertagging step with valency constraints, expressed as an integer linear program; (2) the reduction of the graph prediction problem to the maximum matching problem; (3) the design of an incremental early-stopping training strategy to prevent overfitting. Experimentally, our approach significantly improves results on examples that require structural generalization in the COGS dataset, a known challenging benchmark for compositional generalization. Overall, these results confirm that structural constraints are important for generalization in semantic parsing.

10:00-10:15 (West 3 Ballroom)
**CoRec: An Easy Approach for Coordination Recognition**
*Qing Wang, Haojie Jia, Wenfei Song and Qi Li*
In this paper, we observe and address the challenges of the coordination recognition task. Most existing methods rely on syntactic parsers to identify the coordinators in a sentence and detect the coordination boundaries. However, state-of-the-art syntactic parsers are slow and suffer from errors, especially for long and complicated sentences. To better solve the problems, we propose a pipeline model COordination RECognizer (CoRec). It consists of two components: coordinator identifier and conjunct boundary detector. The experimental results on

datasets from various domains demonstrate the effectiveness and efficiency of the proposed method. Further experiments show that CoRec positively impacts downstream tasks, improving the yield of state-of-the-art Open IE models.

10:15-10:30 (West 3 Ballroom)
**LLM-enhanced Self-training for Cross-domain Constituency Parsing**
*Jianling Li, Meishan Zhang, Peiming Guo, Min Zhang and Yue Zhang*
Self-training has proven to be an effective approach for cross-domain tasks, and in this study, we explore its application to cross-domain constituency parsing. Traditional self-training methods rely on limited and potentially low-quality raw corpora. To overcome this limitation, we propose enhancing self-training with the large language model (LLM) to generate domain-specific raw corpora iteratively. For the constituency parsing, we introduce grammar rules that guide the LLM in generating raw corpora and establish criteria for selecting pseudo instances. Our experimental results demonstrate that self-training for constituency parsing, equipped with an LLM, outperforms traditional methods regardless of the LLM's performance. Moreover, the combination of grammar rules and confidence criteria for pseudo-data selection yields the highest performance in the cross-domain constituency parsing.

## Demo session 6

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs**
*Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch and Alan Akbik*
Most NLP tasks are modeled as supervised learning and thus require labeled training data to train effective models. However, manually producing such data at sufficient quality and quantity is known to be costly and time-intensive. Current research addresses this bottleneck by exploring a novel paradigm called zero-shot learning via dataset generation. Here, a powerful LLM is prompted with a task description to generate labeled data that can be used to train a downstream NLP model. For instance, an LLM might be prompted to "generate 500 movie reviews with positive overall sentiment, and another 500 with negative sentiment." The generated data could then be used to train a binary sentiment classifier, effectively leveraging an LLM as a teacher to a smaller student model. With this demo, we introduce Fabricator, an open-source Python toolkit for dataset generation. Fabricator implements common dataset generation workflows, supports a wide range of downstream NLP tasks (such as text classification, question answering, and entity recognition), and is integrated with well-known libraries to facilitate quick experimentation. With Fabricator, we aim to support researchers in conducting reproducible dataset generation experiments using LLMs and help practitioners apply this approach to train models for downstream tasks.

09:00-10:30 (East Foyer / Virtual)
**End-to-End Evaluation for Low-Latency Simultaneous Speech Translation**
*Christian Huber, Tu Anh Dinh, Carlos Mullov, Ngoc-Quan Pham, Thai Binh Nguyen, Fabian Retkowski, Stefan Constantin, Enes Ugan, Danni Liu, Zhaolin Li, Sai Koneru, Jan Niehues and Alexander Waibel*
The challenge of low-latency speech translation has recently draw significant interest in the research community as shown by several publications and shared tasks. Therefore, it is essential to evaluate these different approaches in realistic scenarios. However, currently only specific aspects of the systems are evaluated and often it is not possible to compare different approaches. In this work, we propose the first framework to perform and evaluate the various aspects of low-latency speech translation under realistic conditions. The evaluation is carried out in an end-to-end fashion. This includes the segmentation of the audio as well as the run-time of the different components. Secondly, we compare different approaches to low-latency speech translation using this framework. We evaluate models with the option to revise the output as well as methods with fixed output. Furthermore, we directly compare state-of-the-art cascaded as well as end-to-end systems. Finally, the framework allows to automatically evaluate the translation quality as well as latency and also provides a web interface to show the low-latency model outputs to the user.

09:00-10:30 (East Foyer / Virtual)
**Gentopia.AI: A Collaborative Platform for Tool-Augmented LLMs**
*Binfeng Xu, Xukun Liu, Hua Shen, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao and Dongkuan Xu*
Augmented Language Models (ALMs) empower large language models with the ability to use tools, transforming them into intelligent agents for real-world interactions. However, most existing frameworks for ALMs, to varying degrees, are deficient in the following critical features: flexible customization, collaborative democratization, and holistic evaluation. This paper proposes Gentopia, a lightweight and extensible framework for ALMs. Gentopia allows the flexible customization of agents through simple configurations, seamlessly integrating various language models, task formats, prompting modules, and plugins into a unified paradigm. Furthermore, we establish Gentpool, a public platform enabling the registration and sharing of user-customized agents. Agents registered in Gentpool are composable such that they can be assembled together for agent collaboration, advancing the democratization of artificial intelligence. To ensure high-quality agents, Gentbench, an integral component of Gentpool, is designed to thoroughly evaluate user-customized agents across diverse aspects such as safety, robustness, efficiency, etc. We release Gentopia on Github and will continuously move forward.

09:00-10:30 (East Foyer / Virtual)
**SentAlign: Accurate and Scalable Sentence Alignment**
*Steinthor Steingrimsson, Hrafn Loftsson and Andy Way*
We present SentAlign, an accurate sentence alignment tool designed to handle very large parallel document pairs. Given user-defined parameters, the alignment algorithm evaluates all possible alignment paths in fairly large documents of thousands of sentences and uses a divide-and-conquer approach to align documents containing tens of thousands of sentences. The scoring function is based on LaBSE bilingual sentence representations. SentAlign outperforms five other sentence alignment tools when evaluated on two different evaluation sets, German-French and English-Icelandic, and on a downstream machine translation task.

09:00-10:30 (East Foyer / Virtual)
**QACheck: A Demonstration System for Question-Guided Multi-Hop Fact-Checking**
*Liangming Pan, Xinyuan Lu, Min-Yen Kan and Preslav Nakov*
Fact-checking real-world claims often requires intricate, multi-step reasoning due to the absence of direct evidence to support or refute them. However, existing fact-checking systems often lack transparency in their decision-making, making it challenging for users to comprehend their reasoning process. To address this, we propose the Question-guided Multi-hop Fact-Checking (QACheck) system, which guides the model's reasoning process by asking a series of questions critical for verifying a claim. QACheck has five key modules: a claim verifier, a question generator, a question-answering module, a QA validator, and a reasoner. Users can input a claim into QACheck, which then predicts its veracity and provides a comprehensive report detailing its reasoning process, guided by a sequence of (question, answer) pairs. QACheck

also provides the source of evidence supporting each question, fostering a transparent, explainable, and user-friendly fact-checking process.

09:00-10:30 (East Foyer / Virtual)
**Kandinsky: An Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion**
*Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov and Denis Dimitrov*
Text-to-image generation is a significant domain in modern computer vision and achieved substantial improvements through the evolution of generative architectures. Among these, diffusion-based models demonstrated essential quality enhancements. These models generally split into two categories: pixel-level and latent-level approaches. We present Kandinsky – a novel exploration of latent diffusion architecture, combining the principles of image prior models with latent diffusion techniques. The image prior model, is trained separately to map CLIP text and image embeddings. Another distinct feature of the proposed model is the modified MoVQ implementation, which serves as the image autoencoder component. Overall the designed model contains 3.3B parameters. We also deployed a user-friendly demo system that supports diverse generative modes such as text-to-image generation, image fusion, text and image fusion, image variations generation and text-guided inpainting/outpainting. Additionally we released the source code and checkpoints for Kandinsky models. Experimental evaluations demonstrate FID score of 8.03 on the COCO-30K dataset, marking our model as the top open source performer in terms of measurable image generation quality.

09:00-10:30 (East Foyer / Virtual)
**NewsRecLib: A PyTorch-Lightning Library for Neural News Recommendation**
*Andreea Iana, Goran Glavaš and Heiko Paulheim*
NewsRecLib is an open-source library based on Pytorch-Lightning and Hydra developed for training and evaluating neural news recommendation models. The foremost goals of NewsRecLib are to promote reproducible research and rigorous experimental evaluation by (i) providing a unified and highly configurable framework for exhaustive experimental studies and (ii) enabling a thorough analysis of the performance contribution of different model architecture components and training regimes. NewsRecLib is highly modular, allows specifying experiments in a single configuration file, and includes extensive logging facilities. Moreover, NewsRecLib provides out-of-the-box implementations of several prominent neural models, training methods, standard evaluation benchmarks, and evaluation metrics for news recommendation.

09:00-10:30 (East Foyer / Virtual)
**MiniChain: A Small Library for Coding with Large Language Models**
*Alexander Rush*
Programming augmented by large language models (LLMs) opens up many new application areas, but also requires care. LLMs are accurate enough, on average, to replace core functionality, yet make basic mistakes that demonstrate a lack of robustness. An ecosystem of prompting tools, from intelligent agents to new programming languages, have emerged with different solutions for patching LLMs with other tools. In this work, we introduce MiniChain, an opinionated tool for LLM augmented programming, with the design goals of ease-of-use of prototyping, transparency through automatic visualization, and a minimalistic approach to advanced features. The MiniChain library provides core primitives for coding LLM calls, separating out prompt templates, and capturing program structure. The library includes demo implementations of the main applications papers in the area, including chat-bots, code generation, retrieval-based question answering, and complex information extraction. The library is open-source and available at https://github.com/srush/MiniChain, with code demos available at https://srush-minichain.hf.space/, and video demo at https://www.youtube.com/watch?v=VszZ1VnO7sk.

## Poster session 6

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**#1 Rethinking Model Selection and Decoding for Keyphrase Generation with Pre-trained Sequence-to-Sequence Models**
*Di Wu, Wasi Uddin Ahmad and Kai-Wei Chang*
Keyphrase Generation (KPG) is a longstanding task in NLP with widespread applications. The advent of sequence-to-sequence (seq2seq) pre-trained language models (PLMs) has ushered in a transformative era for KPG, yielding promising performance improvements. However, many design decisions remain unexplored and are often made arbitrarily. This paper undertakes a systematic analysis of the influence of model selection and decoding strategies on PLM-based KPG. We begin by elucidating why seq2seq PLMs are apt for KPG, anchored by an attention-driven hypothesis. We then establish that conventional wisdom for selecting seq2seq PLMs lacks depth: (1) merely increasing model size or performing task-specific adaptation is not parameter-efficient; (2) although combining in-domain pre-training with task adaptation benefits KPG, it does partially hinder generalization. Regarding decoding, we demonstrate that while greedy search achieves strong F1 scores, it lags in recall compared with sampling-based methods. Based on these insights, we propose DeSel, a likelihood-based decode-select algorithm for seq2seq PLMs. DeSel improves greedy search by an average of 4.7% semantic F1 across five datasets. Our collective findings pave the way for deeper future investigations into PLM-based KPG.

09:00-10:30 (East Foyer / Virtual)
**#2 RepoCoder: Repository-Level Code Completion Through Iterative Retrieval and Generation**
*Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou and Weizhu Chen*
The task of repository-level code completion is to continue writing the unfinished code based on a broader context of the repository. While for automated code completion tools, it is difficult to utilize the useful information scattered in different files. We propose RepoCoder, a simple, generic, and effective framework to address the challenge. It streamlines the repository-level code completion process by incorporating a similarity-based retriever and a pre-trained code language model in an iterative retrieval-generation pipeline. RepoCoder makes effective utilization of repository-level information for code completion and has the ability to generate code at various levels of granularity. Moreover, we propose a new benchmark RepoBench, which consists of the latest and high-quality real-world repositories covering line, API invocation, and function body completion scenarios. Experimental results indicate that RepoCoder significantly improves the In-File completion baseline by over 10% in all settings and consistently outperforms the vanilla retrieval-augmented code completion approach. Furthermore, we validate the effectiveness of RepoCoder through comprehensive analysis, providing valuable insights for future research. Our source code and benchmark will be publicly available after the paper review.

09:00-10:30 (East Foyer / Virtual)
**#3 NameGuess: Column Name Expansion for Tabular Data**
*Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Shen Wang, Huzefa Rangwala and George Karypis*
Recent advances in large language models have revolutionized many sectors, including the database industry. One common challenge when dealing with large volumes of tabular data is the pervasive use of abbreviated column names, which can negatively impact performance on various data search, access, and understanding tasks. To address this issue, we introduce a new task, called NameGuess, to expand column

names (used in database schema) as a natural language generation problem. We create a training dataset of 384K abbreviated-expanded column pairs using a new data fabrication method and a human-annotated evaluation benchmark that includes 9.2K examples from real-world tables. To tackle the complexities associated with polysemy and ambiguity in NameGuess, we enhance auto-regressive language models by conditioning on table content and column header names – yielding a fine-tuned model (with 2.7B parameters) that matches human performance. Furthermore, we conduct a comprehensive analysis (on multiple LLMs) to validate the effectiveness of table content in NameGuess and identify promising future opportunities. Code has been made available at https://github.com/amazon-science/nameguess.

09:00-10:30 (East Foyer / Virtual)
**#4 FLatS: Principled Out-of-Distribution Detection with Feature-Based Likelihood Ratio Score**
*Haowei Lin and Yuntian Gu*
Detecting out-of-distribution (OOD) instances is crucial for NLP models in practical applications. Although numerous OOD detection methods exist, most of them are empirical. Backed by theoretical analysis, this paper advocates for the measurement of the "OOD-ness" of a test case x through the *likelihood ratio* between out-distribution $\mathcal{P}_{out}$ and in-distribution $\mathcal{P}_{in}$. We argue that the state-of-the-art (SOTA) feature-based OOD detection methods, such as Maha and KNN, are suboptimal since they only estimate in-distribution density p_in x. To address this issue, we propose **FLATS**, a principled solution for OOD detection based on likelihood ratio. Moreover, we demonstrate that FLATS can serve as a general framework capable of enhancing other OOD detection methods by incorporating out-distribution density p_out x estimation. Experiments show that FLATS establishes a new SOTA on popular benchmarks.

09:00-10:30 (East Foyer / Virtual)
**#5 Query-as-context Pre-training for Dense Passage Retrieval**
*Xing W, Guangyuan Ma, Wanhui Qian, Zijia Lin and Songlin Hu*
Recently, methods have been developed to improve the performance of dense passage retrieval by using context-supervised pre-training. These methods simply consider two passages from the same document to be relevant, without taking into account the potential negative impacts of weakly correlated pairs. Thus, this paper proposes query-as-context pre-training, a simple yet effective pre-training technique to alleviate the issue. Query-as-context pre-training assumes that the query derived from a passage is more likely to be relevant to that passage and forms a passage-query pair. These passage-query pairs are then used in contrastive or generative context-supervised pre-training. The pre-trained models are evaluated on large-scale passage retrieval benchmarks and out-of-domain zero-shot benchmarks. Experimental results show that query-as-context pre-training brings considerable gains for retrieval performances, demonstrating its effectiveness and efficiency.

09:00-10:30 (East Foyer / Virtual)
**#6 Make Every Example Count: On the Stability and Utility of Self-Influence for Learning from Noisy NLP Datasets**
*Irina Bejan, Artem Sokolov and Katja Filippova*
Increasingly larger datasets have become a standard ingredient to advancing the state-of-the-art in NLP. However, data quality might have already become the bottleneck to unlock further gains. Given the diversity and the sizes of modern datasets, standard data filtering is not straight-forward to apply, because of the multifacetedness of the harmful data and elusiveness of filtering rules that would generalize across multiple tasks. We study the fitness of task-agnostic self-influence scores of training examples for data cleaning, analyze their efficacy in capturing naturally occurring outliers, and investigate to what extent self-influence based data cleaning can improve downstream performance in machine translation, question answering and text classification, building up on recent approaches to self-influence calculation and automated curriculum learning.

09:00-10:30 (East Foyer / Virtual)
**#7 Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive**
*Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M Homan and Ashiqur R. KhudaBukhsh*
Offensive speech detection is a key component of content moderation. However, what is offensive can be highly subjective. This paper investigates how machine and human moderators disagree on what is offensive when it comes to real-world social web political discourse. We show that (1) there is extensive disagreement among the moderators (humans and machines); and (2) human and large-language-model classifiers are unable to predict how other human raters will respond, based on their political leanings. For (1), we conduct a ***noise audit*** at an unprecedented scale that combines both machine and human responses. For (2), we introduce a first-of-its-kind dataset of ***vicarious offense***. Our noise audit reveals that moderation outcomes vary wildly across different machine moderators. Our experiments with human moderators suggest that political leanings combined with sensitive issues affect both first-person and vicarious offense. The dataset is available through https://github.com/Homan-Lab/voiced.

09:00-10:30 (East Foyer / Virtual)
**#8 CaseEncoder: A Knowledge-enhanced Pre-trained Model for Legal Case Encoding**
*Yixiao Ma, Yueyue WU, Weihang Su, Qingyao Ai and Yiqun Liu*
Legal case retrieval is a critical process for modern legal information systems. While recent studies have utilized pre-trained language models (PLMs) based on the general domain self-supervised pre-training paradigm to build models for legal case retrieval, there are limitations in using general domain PLMs as backbones. Specifically, these models may not fully capture the underlying legal features in legal case documents. To address this issue, we propose CaseEncoder, a legal document encoder that leverages fine-grained legal knowledge in both the data sampling and pre-training phases. In the data sampling phase, we enhance the quality of the training data by utilizing fine-grained law article information to guide the selection of positive and negative examples. In the pre-training phase, we design legal-specific pre-training tasks that align with the judging criteria of relevant legal cases. Based on these tasks, we introduce an innovative loss function called Biased Circle Loss to enhance the model's ability to recognize case relevance in fine grains. Experimental results on multiple benchmarks demonstrate that CaseEncoder significantly outperforms both existing general pre-training models and legal-specific pre-training models in zero-shot legal case retrieval. The source code of CaseEncoder can be found at https://github.com/Anonymous-EMNLP2023/CaseEncoder.

09:00-10:30 (East Foyer / Virtual)
**#9 PTP: Boosting Stability and Performance of Prompt Tuning with Perturbation-Based Regularizer**
*Lichang Chen, Jiuhai Chen, Heng Huang and Minhao Cheng*
Recent studies show that prompt tuning can better leverage the power of large language models than fine-tuning on downstream natural language understanding tasks. However, the existing prompt tuning methods have training instability issues, as the variance of scores under different random seeds is quite large. To address this critical problem, we first investigate and find that the loss landscape of vanilla prompt tuning is precipitous when it is visualized, where a slight change of input data can cause a big fluctuation in the loss landscape. This is an essential factor that leads to the instability of prompt tuning. Based on this observation, we introduce perturbation-based regularizers, which can smooth the loss landscape, into prompt tuning. We propose a new algorithm, called Prompt Tuning with Perturbation-based regularizer (PTP), which can not only alleviate training instability dramatically but also boost the performance of prompt tuning. We design two kinds of perturbation-based regularizers, including random-noise-based and adversarial-based. In particular, our proposed perturbations are flexible on both text space and embedding space. Extensive experiments show the effectiveness of our proposed methods in stabilizing the training.

Our new algorithms improve the state-of-the-art prompt tuning methods by 1.94% and 2.34% on SuperGLUE and FewGLUE benchmarks, respectively.

09:00-10:30 (East Foyer / Virtual)
**#10 UDAPDR: Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers**
*Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Arafat Sultan and Christopher Potts*
Many information retrieval tasks require large labeled datasets for fine-tuning. However, such datasets are often unavailable, and their utility for real-world applications can diminish quickly due to domain shifts. To address this challenge, we develop and motivate a method for using large language models (LLMs) to generate large numbers of synthetic queries cheaply. The method begins by generating a small number of synthetic queries using an expensive LLM. After that, a much less expensive one is used to create large numbers of synthetic queries, which are used to fine-tune a family of reranker models. These rerankers are then distilled into a single efficient retriever for use in the target domain. We show that this technique boosts zero-shot accuracy in long-tail domains and achieves substantially lower latency than standard reranking methods.

09:00-10:30 (East Foyer / Virtual)
**#11 Modeling Conceptual Attribute Likeness and Domain Inconsistency for Metaphor Detection**
*Yuan Tian, Nan Xu, Wenji Mao and Daniel Dajun Zeng*
Metaphor detection is an important and challenging task in natural language processing, which aims to distinguish between metaphorical and literal expressions in text. Previous studies mainly leverage the incongruity of source and target domains and contextual clues for detection, neglecting similar attributes shared between source and target concepts in metaphorical expressions. Based on conceptual metaphor theory, these similar attributes are essential to infer implicit meanings conveyed by the metaphor. Under the guidance of conceptual metaphor theory, in this paper, we model the likeness of attribute for the first time and propose a novel Attribute Likeness and Domain Inconsistency Learning framework (AIDIL) for word-pair metaphor detection. Specifically, we propose an attribute siamese network to mine similar attributes between source and target concepts. We then devise a domain contrastive learning strategy to learn the semantic inconsistency of concepts in source and target domains. Extensive experiments on four datasets verify that our method significantly outperforms the previous state-of-the-art methods, and demonstrate the generalization ability of our method.

09:00-10:30 (East Foyer / Virtual)
**#12 Empower Nested Boolean Logic via Self-Supervised Curriculum Learning**
*Hongqiu Wu, Linfeng Liu, Hai Zhao and Min Zhang*
Beyond the great cognitive powers showcased by language models, it is crucial to scrutinize whether their reasoning capabilities stem from strong generalization or merely exposure to relevant data. As opposed to constructing increasingly complex logic, this paper probes into the boolean logic, the root capability of a logical reasoner. We find that any pre-trained language models even including large language models only behave like a random selector in the face of multi-nested boolean logic, a task that humans can handle with ease. To empower language models with this fundamental capability, this paper proposes a new self-supervised learning method Curriculum Logical Reasoning (Clr), where we augment the training data with nested boolean logic chain step-by-step, and program the training from simpler logical patterns gradually to harder ones. This new training paradigm allows language models to effectively generalize to much harder and longer-hop logic, which can hardly be learned through naive training. Furthermore, we show that boolean logic is a great foundation for improving the subsequent general logical tasks.

09:00-10:30 (East Foyer / Virtual)
**#13 Longtriever: a Pre-trained Long Text Encoder for Dense Document Retrieval**
*Junhan Yang, Zheng Liu, Chaozhuo Li, Guangzhong Sun and Xing Xie*
Pre-trained language models (PLMs) have achieved the preeminent position in dense retrieval due to their powerful capacity in modeling intrinsic semantics. However, most existing PLM-based retrieval models encounter substantial computational costs and are infeasible for processing long documents. In this paper, a novel retrieval model Longtriever is proposed to embrace three core challenges of long document retrieval: substantial computational cost, incomprehensive document understanding, and scarce annotations. Longtriever splits long documents into short blocks and then efficiently models the local semantics within a block and the global context semantics across blocks in a tightly-coupled manner. A pre-training phase is further proposed to empower Longtriever to achieve a better understanding of underlying semantic correlations. Experimental results on two popular benchmark datasets demonstrate the superiority of our proposal.

09:00-10:30 (East Foyer / Virtual)
**#14 Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4**
*Kent K. Chang, Mackenzie Hanh Cramer, Sandeep Soni and David Bamman*
In this work, we carry out a data archaeology to infer books that are known to ChatGPT and GPT-4 using a name cloze membership inference query. We find that OpenAI models have memorized a wide collection of copyrighted materials, and that the degree of memorization is tied to the frequency with which passages of those books appear on the web. The ability of these models to memorize an unknown set of books complicates assessments of measurement validity for cultural analytics by contaminating test data; we show that models perform much better on memorized books than on non-memorized books for downstream tasks. We argue that this supports a case for open models whose training data is known.

09:00-10:30 (East Foyer / Virtual)
**#15 DREAM: Deployment of Recombination and Ensembles in Argument Mining**
*Florian Ruosch, Cristina Sarasua and Abraham Bernstein*
Current approaches to Argument Mining (AM) tend to take a holistic or black-box view of the overall pipeline. This paper, in contrast, aims to provide a solution to achieve increased performance based on current components instead of independent all-new solutions. To that end, it presents the Deployment of Recombination and Ensemble methods for Argument Miners (DREAM) framework that allows for the (automated) combination of AM components. Using ensemble methods, DREAM combines sets of AM systems to improve accuracy for the four tasks in the AM pipeline. Furthermore, it leverages recombination by using different argument miners elements throughout the pipeline. Experiments with five systems previously included in a benchmark show that the systems combined with DREAM can outperform the previous best single systems in terms of accuracy measured by an AM benchmark.

09:00-10:30 (East Foyer / Virtual)
**#16 Instructed Language Models with Retrievers Are Powerful Entity Linkers**
*Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou and Daxin Jiang*
Generative approaches powered by large language models (LLMs) have demonstrated emergent abilities in tasks that require complex reasoning abilities. Yet the generative nature still makes the generated content suffer from hallucinations, thus unsuitable for entity-centric tasks like entity linking (EL) requiring precise entity predictions over a large knowledge base. We present Instructed Generative Entity Linker

(INSGENEL), the first approach that enables casual language models to perform entity linking over knowledge bases. Several methods of equipping language models with EL ability were proposed in this work, including (i) a sequence-to-sequence training EL objective with instruction-tuning, (ii) a novel generative EL framework based on a light-weight potential mention retriever that frees the model from heavy and non-parallelizable decoding, achieving 4× speedup without compromise on linking metrics. INSGENEL outperforms previous generative alternatives with +6.8 F1 points gain on average, also with a huge advantage in training data efficiency and training compute consumption. In addition, our skillfully-engineered in-context learning (ICL) framework for EL still lags behind INSGENEL significantly, reaffirming that the EL task remains a persistent hurdle for general LLMs.

09:00-10:30 (East Foyer / Virtual)
**#17 CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network**
*Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar and Dinesh Manocha*
The tremendous growth of social media users interacting in online conversations has led to significant growth in hate speech affecting people from various demographics. Most of the prior works focus on detecting explicit hate speech, which is overt and leverages hateful phrases, with very little work focusing on detecting hate speech that is implicit or denotes hatred through indirect or coded language. In this paper, we present CoSyn, a context synergized neural network that explicitly incorporates user- and conversational-context for detecting implicit hate speech in online conversations. CoSyn introduces novel ways to encode these external contexts and employs a novel context interaction mechanism that clearly captures the interplay between them, making independent assessments of the amounts of information to be retrieved from these noisy contexts. Additionally, it carries out all these operations in the hyperbolic space to account for the scale-free dynamics of social media. We demonstrate the effectiveness of CoSyn on 6 hate speech datasets and show that CoSyn outperforms all our baselines in detecting implicit hate speech with absolute improvements in the range of 1.24% - 57.8%. We make our code available.

09:00-10:30 (East Foyer / Virtual)
**#18 Transductive Learning for Textual Few-Shot Classification in API-based Embedding Models**
*Pierre Colombo, Victor Pellegrain, Malik Boudiaf, Myriam Tami, Victor Storchan, Ismail Ben Ayed and Pablo Piantanida*
Proprietary and closed APIs are becoming increasingly common to process natural language, and are impacting the practical applications of natural language processing, including few-shot classification. Few-shot classification involves training a model to perform a new classification task with a handful of labeled data. This paper presents three contributions. First, we introduce a scenario where the embedding of a pre-trained model is served through a gated API with compute-cost and data-privacy constraints. Second, we propose a transductive inference, a learning paradigm that has been overlooked by the NLP community. Transductive inference, unlike traditional inductive learning, leverages the statistics of unlabelled data. We also introduce a new parameter-free transductive regularizer based on the Fisher-Rao loss, which can be used on top of the gated API embeddings. This method fully utilizes unlabelled data, does not share any label with the third-party API provider and could serve as a baseline for future research. Third, we propose an improved experimental setting and compile a benchmark of eight datasets involving multiclass classification in four different languages, with up to 151 classes. We evaluate our methods using eight backbone models, along with an episodic evaluation over 1,000 episodes, which demonstrate the superiority of transductive inference over the standard inductive setting.

09:00-10:30 (East Foyer / Virtual)
**#19 STINMatch: Semi-Supervised Semantic-Topological Iteration Network for Financial Risk Detection via News Label Diffusion**
*Xurui Li, Yue Qin, Rui Zhu, Tianqianjin Lin, Yongming Fan, Yangyang Kang, Kaisong Song, Fubang Zhao, Changlong Sun, Haixu Tang and Xiaozhong Liu*
Commercial news provide rich semantics and timely information for automated financial risk detection. However, unaffordable large-scale annotation as well as training data sparseness barrier the full exploitation of commercial news in risk detection. To address this problem, we propose a semi-supervised Semantic-Topological Iteration Network, STINMatch, along with a news-enterprise knowledge graph (NEKG) to endorse the risk detection enhancement. The proposed model incorporates a label correlation matrix and interactive consistency regularization techniques into the iterative joint learning framework of text and graph modules. The carefully designed framework takes full advantage of the labeled and unlabeled data as well as their interrelations, enabling deep label diffusion coordination between article-level semantics and label correlations following the topological structure. Extensive experiments demonstrate the superior effectiveness and generalization ability of STINMatch.

09:00-10:30 (East Foyer / Virtual)
**#20 Clinical Contradiction Detection**
*Dave Makhervaks, Plia Gillis and Kira Radinsky*
Detecting contradictions in text is essential in determining the validity of the literature and sources that we consume. Medical corpora are riddled with conflicting statements. This is due to the large throughput of new studies and the difficulty in replicating experiments, such as clinical trials. Detecting contradictions in this domain is hard since it requires clinical expertise. We present a distant supervision approach that leverages a medical ontology to build a seed of potential clinical contradictions over 22 million medical abstracts. We automatically build a labeled training dataset consisting of paired clinical sentences that are grounded in an ontology and represent potential medical contradiction. The dataset is used to weakly-supervise state-of-the-art deep learning models showing significant empirical improvements across multiple medical contradiction datasets.

09:00-10:30 (East Foyer / Virtual)
**#21 Toward a Critical Toponymy Framework for Named Entity Recognition: A Case Study of Airbnb in New York City**
*Mikael Brunila, Jack LaViolette, Sky CH-Wang, Priyanka Verma, Clara Féré and Grant McKenzie*
Critical toponymy examines the dynamics of power, capital, and resistance through place names and the sites to which they refer. Studies here have traditionally focused on the semantic content of toponyms and the top-down institutional processes that produce them. However, they have generally ignored the ways in which toponyms are used by ordinary people in everyday discourse, as well as the other strategies of geospatial description that accompany and contextualize toponymic reference. Here, we develop computational methods to measure how cultural and economic capital shape the ways in which people refer to places, through a novel annotated dataset of 47,440 New York City Airbnb listings from the 2010s. Building on this dataset, we introduce a new named entity recognition (NER) model able to identify important discourse categories integral to the characterization of place. Our findings point toward new directions for critical toponymy and to a range of previously understudied linguistic signals relevant to research on neighborhood status, housing and tourism markets, and gentrification.

09:00-10:30 (East Foyer / Virtual)
**#22 Large-scale similarity search with Optimal Transport**
*Cléa Mehnia Laouar, Yuki Takezawa and Makoto Yamada*
Wasserstein distance is a powerful tool for comparing probability distributions and is widely used for document classification and retrieval tasks in NLP. In particular, it is known as the word mover's distance (WMD) in the NLP community. WMD exhibits excellent performance for various NLP tasks; however, one of its limitations is its computational cost and thus is not useful for large-scale distribution comparisons. In this study, we propose a simple and effective nearest neighbor search based on the Wasserstein distance. Specifically, we employ the L1

embedding method based on the tree-based Wasserstein approximation and subsequently used the nearest neighbor search to efficiently find the $k$-nearest neighbors. Through benchmark experiments, we demonstrate that the proposed approximation has comparable performance to the vanilla Wasserstein distance and can be computed three orders of magnitude faster than the vanilla Wasserstein distance.

09:00-10:30 (East Foyer / Virtual)
**#23 A linear time approximation of Wasserstein distance with word embedding selection**
*Sho Otao and Makoto Yamada*
Wasserstein distance, which can be computed by solving the optimal transport problem, is a powerful method for measuring the dissimilarity between documents. In the NLP community, it is referred to as word mover's distance (WMD). One of the key challenges of Wasserstein distance is its computational cost since it needs cubic time. Although the Sinkhorn algorithm is a powerful tool to speed up to compute the Wasserstein distance, it still requires square time. Recently, a linear time approximation of the Wasserstein distance including the sliced Wasserstein and the tree-Wasserstein distance (TWD) has been proposed. However, a linear time approximation method suffers when the dimensionality of word vectors is high. In this study, we propose a method to combine feature selection and tree approximation of Wasserstein distance to handle high-dimensional problems. More specifically, we use multiple word embeddings and automatically select useful word embeddings in a tree approximation of Wasserstein distance. To this end, we approximate Wasserstein distance for each word vector by tree approximation technique, and select the discriminative (i.e., large Wasserstein distance) word embeddings by solving an entropic regularized maximization problem. Through our experiments on document classification, our proposed method achieved high performance.

09:00-10:30 (East Foyer / Virtual)
**#24 Rethinking Negative Pairs in Code Search**
*Haochen Li, Xin Zhou, Anh Tuan Luu and Chunyan Miao*
Recently, contrastive learning has become a key component in fine-tuning code search models for software development efficiency and effectiveness. It pulls together positive code snippets while pushing negative samples away given search queries. Among contrastive learning, InfoNCE is the most widely used loss function due to its better performance. However, the following problems in negative samples of InfoNCE may deteriorate its representation learning: 1) The existence of false negative samples in large code corpora due to duplications. 2). The failure to explicitly differentiate between the potential relevance of negative samples. As an example, a bubble sorting algorithm example is less "negative" than a file saving function for the quick sorting algorithm query. In this paper, we tackle the above problems by proposing a simple yet effective Soft-InfoNCE loss that inserts weight terms into InfoNCE. In our proposed loss function, we apply three methods to estimate the weights of negative pairs and show that the vanilla InfoNCE loss is a special case of Soft-InfoNCE. Theoretically, we analyze the effects of Soft-InfoNCE on controlling the distribution of learnt code representations and on deducing a more precise mutual information estimation. We furthermore discuss the superiority of proposed loss functions with other design alternatives. Extensive experiments demonstrate the effectiveness of Soft-InfoNCE and weights estimation methods under state-of-the-art code search models on a large-scale public dataset consisting of six programming languages.

09:00-10:30 (East Foyer / Virtual)
**#25 Decoding the Silent Majority: Inducing Belief Augmented Social Graph with Large Language Model for Response Forecasting**
*Chenkai Sun, Jinning Li, Yi Fung, Hou Pong Chan, Tarek Abdelzaher, ChengXiang Zhai and Heng Ji*
Automatic response forecasting for news media plays a crucial role in enabling content producers to efficiently predict the impact of news releases and prevent unexpected negative outcomes such as social conflict and moral injury. To effectively forecast responses, it is essential to develop measures that leverage the social dynamics and contextual information surrounding individuals, especially in cases where explicit profiles or historical actions of the users are limited (referred to as lurkers). As shown in a previous study, 97% of all tweets are produced by only the most active 25% of users. However, existing approaches have limited exploration of how to best process and utilize these important features. To address this gap, we propose a novel framework, named SocialSense, that leverages a large language model to induce a belief-centered graph on top of an existent social network, along with graph-based propagation to capture social dynamics. We hypothesize that the induced graph that bridges the gap between distant users who share similar beliefs allows the model to effectively capture the response patterns. Our method surpasses existing state-of-the-art in experimental evaluations for both zero-shot and supervised settings, demonstrating its effectiveness in response forecasting. Moreover, the analysis reveals the framework's capability to effectively handle unseen user and lurker scenarios, further highlighting its robustness and practical applicability.

09:00-10:30 (East Foyer / Virtual)
**#26 CAPSTONE: Curriculum Sampling for Dense Retrieval with Document Expansion**
*Xingwei He, Yeyun Gong, A-Long Jin, Hang Zhang, Anlei Dong, Jian Jiao, Siu Ming Yiu and Nan Duan*
The dual-encoder has become the de facto architecture for dense retrieval. Typically, it computes the latent representations of the query and document independently, thus failing to fully capture the interactions between the query and document. To alleviate this, recent research has focused on obtaining query-informed document representations. During training, it expands the document with a real query, but during inference, it replaces the real query with a generated one. This inconsistency between training and inference causes the dense retrieval model to prioritize query information while disregarding the document when computing the document representation. Consequently, it performs even worse than the vanilla dense retrieval model because its performance heavily relies on the relevance between the generated queries and the real query. In this paper, we propose a curriculum sampling strategy that utilizes pseudo queries during training and progressively enhances the relevance between the generated query and the real query. By doing so, the retrieval model learns to extend its attention from the document alone to both the document and query, resulting in high-quality query-informed document representations. Experimental results on both in-domain and out-of-domain datasets demonstrate that our approach outperforms previous dense retrieval models.

09:00-10:30 (East Foyer / Virtual)
**#27 The Distributional Hypothesis Does Not Fully Explain the Benefits of Masked Language Model Pretraining**
*Ting-Rui Chiang and Dani Yogatama*
We analyze the masked language modeling pretraining objective function from the perspective of the Distributional Hypothesis. We investigate whether the better sample efficiency and the better generalization capability of models pretrained with masked language modeling can be attributed to the semantic similarity encoded in the pretraining data's distributional property. Via a synthetic dataset, our analysis suggests that distributional property indeed leads to the better sample efficiency of pretrained masked language models, but does not fully explain the generalization capability. We also conduct an analysis over two real-world datasets and demonstrate that the distributional property does not explain the generalization ability of pretrained natural language models either. Our results illustrate our limited understanding of model pretraining and provide future research directions.

09:00-10:30 (East Foyer / Virtual)
**#28 Query2doc: Query Expansion with Large Language Models**
*Liang Wang, Nan Yang and Furu Wei*
This paper introduces a simple yet effective query expansion approach, denoted as query2doc, to improve both sparse and dense retrieval systems. The proposed method first generates pseudo-documents by few-shot prompting large language models (LLMs), and then expands

the query with generated pseudo documents. LLMs are trained on web-scale text corpora and are adept at knowledge memorization. The pseudo-documents from LLMs often contain highly relevant information that can aid in query disambiguation and guide the retrievers. Experimental results demonstrate that query2doc boosts the performance of BM25 by 3% to 15% on ad-hoc IR datasets, such as MS-MARCO and TREC DL, without any model fine-tuning. Furthermore, our method also benefits state-of-the-art dense retrievers in terms of both in-domain and out-of-domain results.

09:00-10:30 (East Foyer / Virtual)
**#29 From Values to Opinions: Predicting Human Behaviors and Stances Using Value-Injected Large Language Models**
*Dongjun Kang, Joonsuk Park, Yohan Jo and JinYeong Bak*
Being able to predict people's opinions on issues and behaviors in realistic scenarios can be helpful in various domains, such as politics and marketing. However, conducting large-scale surveys like the European Social Survey to solicit people's opinions on individual issues can incur prohibitive costs. Leveraging prior research showing influence of core human values on individual decisions and actions, we propose to use value-injected large language models (LLM) to predict opinions and behaviors. To this end, we present Value Injection Method (VIM), a collection of two methods—argument generation and question answering—designed to inject targeted value distributions into LLMs via fine-tuning. We then conduct a series of experiments on four tasks to test the effectiveness of VIM and the possibility of using value-injected LLMs to predict opinions and behaviors of people. We find that LLMs value-injected with variations of VIM substantially outperform the baselines. Also, the results suggest that opinions and behaviors can be better predicted using value-injected LLMs than the baseline approaches.

09:00-10:30 (East Foyer / Virtual)
**#30 Out-of-Distribution Generalization in Natural Language Processing: Past, Present, and Future**
*Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster and Yue Zhang*
Machine learning (ML) systems in natural language processing (NLP) face significant challenges in generalizing to out-of-distribution (OOD) data, where the test distribution differs from the training data distribution. This poses important questions about the robustness of NLP models and their high accuracy, which may be artificially inflated due to their underlying sensitivity to systematic biases. Despite these challenges, there is a lack of comprehensive surveys on the generalization challenge from an OOD perspective in natural language understanding. Therefore, this paper aims to fill this gap by presenting the first comprehensive review of recent progress, methods, and evaluations on this topic. We further discuss the challenges involved and potential future research directions. By providing convenient access to existing work, we hope this survey will encourage future research in this area.

09:00-10:30 (East Foyer / Virtual)
**#31 Enhancing Generative Retrieval with Reinforcement Learning from Relevance Feedback**
*Yujia Zhou, Zhicheng Dou and Ji-Rong Wen*
The recent advent of end-to-end generative retrieval marks a significant shift in document retrieval methods, leveraging differentiable search indexes to directly produce relevant document identifiers (docids) in response to a specific query. Nevertheless, this approach faces two fundamental challenges: (i) a discrepancy between the token-level probabilistic optimization and the broader document-level relevance estimation; (ii) an overemphasis on top-1 results at the expense of overall ranking quality. To tackle these challenges, we propose a generative retrieval model with reinforcement learning from relevance feedback, which aims to align token-level docid generation with document-level relevance estimation. The training process incorporates three stages: supervised fine-tuning, relevance reward model training, and reinforced learning-to-rank from relevance feedback. To train a high-quality reward model, we define "relevance" under three progressive scenarios, which collectively offer a comprehensive evaluation of the document relevance. Experiments conducted on two benchmark datasets demonstrate the effectiveness of our proposed approach.

09:00-10:30 (East Foyer / Virtual)
**#32 PHD: Pixel-Based Language Modeling of Historical Documents**
*Nadav Borenstein, Phillip Rust, Desmond Elliott and Isabelle Augenstein*
The digitisation of historical documents has provided historians with unprecedented research opportunities. Yet, the conventional approach to analysing historical documents involves converting them from images to text using OCR, a process that overlooks the potential benefits of treating them as images and introduces high levels of noise. To bridge this gap, we take advantage of recent advancements in pixel-based language models trained to reconstruct masked patches of pixels instead of predicting token distributions. Due to the scarcity of real historical scans, we propose a novel method for generating synthetic scans to resemble real historical documents. We then pre-train our model, PHD, on a combination of synthetic scans and real historical newspapers from the 1700-1900 period. Through our experiments, we demonstrate that PHD exhibits high proficiency in reconstructing masked image patches and provide evidence of our model's noteworthy language understanding capabilities. Notably, we successfully apply our model to a historical QA task, highlighting its usefulness in this domain.

09:00-10:30 (East Foyer / Virtual)
**#33 Multi-view Contrastive Learning for Entity Typing over Knowledge Graphs**
*Zhiwei Hu, Victor Gutierrez Basulto, Zhiliang Xiang, Ru Li and Jeff Z. Pan*
Knowledge graph entity typing (KGET) aims at inferring plausible types of entities in knowledge graphs. Existing approaches to KGET focus on how to better encode the knowledge provided by the neighbors and types of an entity into its representation. However, they ignore the semantic knowledge provided by the way in which types can be clustered together. In this paper, we propose a novel method called Multi-view Contrastive Learning for knowledge graph Entity Typing MCLET, which effectively encodes the coarse-grained knowledge provided by clusters into entity and type embeddings. MCLET is composed of three modules: i) Multi-view Generation and Encoder module, which encodes structured information from entity-type, entity-cluster and cluster-type views; ii) Cross-view Contrastive Learning module, which encourages different views to collaboratively improve view-specific representations of entities and types; iii) Entity Typing Prediction module, which integrates multi-head attention and a Mixture-of-Experts strategy to infer missing entity types. Extensive experiments show the strong performance of MCLET compared to the state-of-the-art

09:00-10:30 (East Foyer / Virtual)
**#34 Sociocultural Norm Similarities and Differences via Situational Alignment and Explainable Textual Entailment**
*Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu and Smaranda Muresan*
Designing systems that can reason across cultures requires that they are grounded in the norms of the contexts in which they operate. However, current research on developing computational models of social norms has primarily focused on American society. Here, we propose a novel approach to discover and compare descriptive social norms across Chinese and American cultures. We demonstrate our approach by leveraging discussions on a Chinese Q&A platform—Zhihu—and the existing SocialChemistry dataset as proxies for contrasting cultural axes, align social situations cross-culturally, and extract social norms from texts using in-context learning. Embedding Chain-of-Thought prompting in a human-AI collaborative framework, we build a high-quality dataset of 3,069 social norms aligned with social situations across Chinese and American cultures alongside corresponding free-text explanations. To test the ability of models to reason about social norms across cultures, we introduce the task of explainable social norm entailment, showing that existing models under 3B parameters have significant room for im-

provement in both automatic and human evaluation. Further analysis of cross-cultural norm differences based on our dataset shows empirical alignment with the social orientations framework, revealing several situational and descriptive nuances in norms across these cultures.

09:00-10:30 (East Foyer / Virtual)
**#35 TaskWeb: Selecting Better Source Tasks for Multi-task NLP**
*Joongwon Kim, Akari Asai, Gabriel Ilharco and Hannaneh Hajishirzi*
Recent work in NLP has shown promising results in training models on large amounts of tasks to achieve better generalization. However, it is not well-understood how tasks are related, and how helpful training tasks can be chosen for a new task. In this work, we investigate whether knowing task relationships via pairwise task transfer improves choosing one or more source tasks that help to learn a new target task. We provide TaskWeb, a large-scale benchmark of pairwise task transfers for 22 NLP tasks using three different model types, sizes, and adaptation methods, spanning about 25,000 experiments. Then, we design a new method TaskShop based on our analysis of TaskWeb. TaskShop uses TaskWeb to estimate the benefit of using a source task for learning a new target task, and to choose a subset of helpful training tasks for multi-task training. Our method improves overall rankings and top-k precision of source tasks by 10% and 38%, respectively. We also use TaskShop to build much smaller multi-task training sets that improve zero-shot performances across 11 different target tasks by at least 4.3%.

09:00-10:30 (East Foyer / Virtual)
**#36 Mitigating Backdoor Poisoning Attacks through the Lens of Spurious Correlation**
*Xuanli He, Qiongkai Xu, Jun Wang, Benjamin I. P. Rubinstein and Trevor Cohn*
Modern NLP models are often trained over large untrusted datasets, raising the potential for a malicious adversary to compromise model behaviour. For instance, backdoors can be implanted through crafting training instances with a specific textual trigger and a target label. This paper posits that backdoor poisoning attacks exhibit a spurious correlation between simple text features and classification labels, and accordingly, proposes methods for mitigating spurious correlation as means of defence. Our empirical study reveals that the malicious triggers are highly correlated to their target labels; therefore such correlations are extremely distinguishable compared to those scores of benign features, and can be used to filter out potentially problematic instances. Compared with several existing defences, our defence method significantly reduces attack success rates across backdoor attacks, and in the case of insertion-based attacks, our method provides a near-perfect defence.

09:00-10:30 (East Foyer / Virtual)
**#37 How Does Generative Retrieval Scale to Millions of Passages?**
*Ronak Pradeep, Kai Hui, Jai Gupta, Adam D Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler and Vinh Q. Tran*
The emerging paradigm of generative retrieval re-frames the classic information retrieval problem into a sequence-to-sequence modeling task, forgoing external indices and encoding an entire document corpus within a single Transformer. Although many different approaches have been proposed to improve the effectiveness of generative retrieval, they have only been evaluated on document corpora on the order of 100K in size. We conduct the first empirical study of generative retrieval techniques across various corpus scales, ultimately scaling up to the entire MS MARCO passage ranking task with a corpus of 8.8M passages and evaluating model sizes up to 11B parameters. We uncover several findings about scaling generative retrieval to millions of passages; notably, the central importance of using synthetic queries as document representations during indexing, the ineffectiveness of existing proposed architecture modifications when accounting for compute cost, and the limits of naively scaling model parameters with respect to retrieval performance. While we find that generative retrieval is competitive with state-of-the-art dual encoders on small corpora, scaling to millions of passages remains an important and unsolved challenge. We believe these findings will be valuable for the community to clarify the current state of generative retrieval, highlight the unique challenges, and inspire new research directions.

09:00-10:30 (East Foyer / Virtual)
**#38 Tree Prompting: Efficient Task Adaptation without Fine-Tuning**
*Chandan Singh, John Xavier Morris, Alexander M Rush, Jianfeng Gao and Yuntian Deng*
Prompting language models (LMs) is the main interface for applying them to new tasks. However, for smaller LMs, prompting provides low accuracy compared to gradient-based fine-tuning. Tree Prompting is an approach to prompting which builds a decision tree of prompts, linking multiple prompt-LM calls together to solve a task. At inference time, each call to the LM is determined by efficiently routing the outcome of the previous call using the tree. Experiments on classification datasets show that Tree Prompting improves accuracy over competing methods and is competitive with fine-tuning. We also show that variants of Tree Prompting allow inspection of a model's decision-making process.

09:00-10:30 (East Foyer / Virtual)
**#39 Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking**
*Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas and Suzan Verberne*
Generating synthetic training data based on large language models (LLMs) for ranking models has gained attention recently. Prior studies use LLMs to build pseudo query-document pairs by generating synthetic queries from documents in a corpus. In this paper, we propose a new perspective of data augmentation: generating synthetic documents from queries. To achieve this, we propose DocGen, that consists of a three-step pipeline that utilizes the few-shot capabilities of LLMs. DocGen pipeline performs synthetic document generation by (i) expanding, (ii) highlighting the original query, and then (iii) generating a synthetic document that is likely to be relevant to the query. To further improve the relevance between generated synthetic documents and their corresponding queries, we propose DocGen-RL, which regards the estimated relevance of the document as a reward and leverages reinforcement learning (RL) to optimize DocGen pipeline. Extensive experiments demonstrate that DocGen pipeline and DocGen-RL significantly outperform existing state-of-theart data augmentation methods, such as InPars, indicating that our new perspective of generating documents leverages the capacity of LLMs in generating synthetic data more effectively. We release the code, generated data, and model checkpoints to foster research in this area.

09:00-10:30 (East Foyer / Virtual)
**#40 mAggretriever: A Simple yet Effective Approach to Zero-Shot Multilingual Dense Retrieval**
*Sheng-Chieh Lin, Amin Ahmad and Jimmy Lin*
Multilingual information retrieval (MLIR) is a crucial yet challenging task due to the need for human annotations in multiple languages, making training data creation labor-intensive. In this paper, we introduce mAggretriever, which effectively leverages semantic and lexical features from pre-trained multilingual transformers (e.g., mBERT and XLM-R) for dense retrieval. To enhance training and inference efficiency, we employ approximate masked-language modeling prediction for computing lexical features, reducing 70–85% GPU memory requirement for mAggretriever fine-tuning. Empirical results demonstrate that mAggretriever, fine-tuned solely on English training data, surpasses existing state-of-the-art multilingual dense retrieval models that undergo further training on large-scale MLIR training data. Our code is available at url.

09:00-10:30 (East Foyer / Virtual)
**#41 Augmenting Zero-Shot Dense Retrievers with Plug-in Mixture-of-Memories**
*Suyu Ge, Chenyan Xiong, Corby Rosset, Arnold Overwijk, Jiawei Han and Paul N. Bennett*
In this paper we improve the zero-shot generalization ability of language models via Mixture-Of-Memory Augmentation (MoMA), a mechanism that retrieves augmentation documents from multiple information corpora (external memories), with the option to "plug in" unseen

memory at inference time. We develop a joint learning mechanism that trains the augmentation component with latent labels derived from the end retrieval task, paired with hard negatives from the memory mixture. We instantiate the model in a zero-shot dense retrieval setting by augmenting strong T5-based retrievers with MoMA. With only T5-base, our model obtains strong zero-shot retrieval accuracy on the eighteen tasks included in the standard BEIR benchmark, outperforming some systems with larger model sizes. As a plug-in-play model, our model can efficiently generalize to any unseen corpus, meanwhile achieving comparable or even better performance than methods relying on target-specific pretraining. Our analysis further illustrates the necessity of augmenting with mixture-of-memory for robust generalization, the benefits of augmentation learning, and how MoMA utilizes the plug-in memory at inference time without changing its parameters. Our code can be found at https://github.com/gesy17/MoMA.

09:00-10:30 (East Foyer / Virtual)
**#42 Quantifying Character Similarity with Vision Transformers**
*Xinmei Yang, Abhishek Arora, Shao-Yu Jheng and Melissa Dell*
Record linkage is a bedrock of quantitative social science, as analyses often require linking data from multiple, noisy sources. Off-the-shelf string matching methods are widely used, as they are straightforward and cheap to implement and scale. Not all character substitutions are equally probable, and for some settings there are widely used handcrafted lists denoting which string substitutions are more likely, that improve the accuracy of string matching. However, such lists do not exist for many settings, skewing research with linked datasets towards a few high-resource contexts that are not representative of the diversity of human societies. This study develops an extensible way to measure character substitution costs for OCR'ed documents, by employing large-scale self-supervised training of vision transformers (ViT) with augmented digital fonts. For each language written with the CJK script, we contrastively learn a metric space where different augmentations of the same character are represented nearby. In this space, homoglyphic characters - those with similar appearance such as "O" and "0" - have similar vector representations. Using the cosine distance between characters' representations as the substitution cost in an edit distance matching algorithm significantly improves record linkage compared to other widely used string matching methods, as OCR errors tend to be homoglyphic in nature. Homoglyphs can plausibly capture character visual similarity across any script, including low-resource settings. We illustrate this by creating homoglyph sets for 3,000 year old ancient Chinese characters, which are highly pictorial. Fascinatingly, a ViT is able to capture relationships in how different abstract concepts were conceptualized by ancient societies, that have been noted in the archaeological literature.

09:00-10:30 (East Foyer / Virtual)
**#43 SPT: Learning to Selectively Insert Prompts for Better Prompt Tuning**
*Wei Zhu and Ming Tan*
Prompt tuning prepends a soft prompt to the input embeddings or hidden states and only optimizes the prompt to adapt pretrained models (PTMs) to downstream tasks. The previous work manually selects prompt layers which are far from optimal and failed to exploit the potential of prompt tuning. In this work, we propose a novel framework, Selective Prompt Tuning (SPT), that learns to select the proper prompt layers by inserting a prompt controlled by a learnable probabilistic gate at each intermediate layer. We further propose a novel bi-level optimization framework, SPT-DARTS, that can better optimize the learnable gates and improve the final prompt tuning performances of the learned prompt layer settings. We conduct extensive experiments with ten benchmark datasets under the full-data and few-shot scenarios. The results demonstrate that our SPT framework can perform better than the previous state-of-the-art PETuning baselines with comparable or fewer tunable parameters.

09:00-10:30 (East Foyer / Virtual)
**#44 MingOfficial: A Ming Official Career Dataset and a Historical Context-Aware Representation Learning Framework**
*You-Jun Chen, Hsin-Yi Hsieh, Yu Tung Lin, Yingtao Tian, Bert Chan, Yu-Sin Liu, Yi-Hsuan Lin and Richard Tzong-Han Tsai*
In Chinese studies, understanding the nuanced traits of historical figures, often not explicitly evident in biographical data, has been a key interest. However, identifying these traits can be challenging due to the need for domain expertise, specialist knowledge, and context-specific insights, making the process time-consuming and difficult to scale. Our focus on studying officials from China's Ming Dynasty is no exception. To tackle this challenge, we propose MingOfficial, a large-scale multi-modal dataset consisting of both structured (career records, annotated personnel types) and text (historical texts) data for $9,376$ officials. We further couple the dataset with a a graph neural network (GNN) to combine both modalities in order to allow investigation of social structures and provide features to boost down-stream tasks. Experiments show that our proposed MingOfficial could enable exploratory analysis of official identities, and also significantly boost performance in tasks such as identifying nuance identities (e.g. civil officials holding military power) from $24.6\%$ to $98.2\%$ $F_1$ score in hold-out test set. By making MingOfficial publicly available (see main text for the URL) as both a dataset and an interactive tool, we aim to stimulate further research into the role of social context and representation learning in identifying individual characteristics, and hope to provide inspiration for computational approaches in other fields beyond Chinese studies.

09:00-10:30 (East Foyer / Virtual)
**#45 Learning Knowledge-Enhanced Contextual Language Representations for Domain Natural Language Understanding**
*Taolin Zhang, Ruyao Xu, Chengyu Wang, Zhongjie Duan, Cen Chen, Minghui Qiu, Dawei Cheng, Xiaofeng He and Weining Qian*
Knowledge-Enhanced Pre-trained Language Models (KEPLMs) improve the performance of various downstream NLP tasks by injecting knowledge facts from large-scale Knowledge Graphs (KGs). However, existing methods for pre-training KEPLMs with relational triples are difficult to be adapted to close domains due to the lack of sufficient domain graph semantics. In this paper, we propose a Knowledge-enhanced language representation learning framework for various closed domains (KANGAROO) via capturing the implicit graph structure among the entities. Specifically, since the entity coverage rates of closed-domain KGs can be relatively low and may exhibit the global sparsity phenomenon for knowledge injection, we consider not only the shallow relational representations of triples but also the hyperbolic embeddings of deep hierarchical entity-class structures for effective knowledge fusion. Moreover, as two closed-domain entities under the same entity-class often havel locally dense neighbor subgraphs counted by max point biconnected component, we further propose a data augmentation strategy based on contrastive learning over subgraphs to construct hard negative samples of higher quality. It makes the underlying KELPMs better distinguish the semantics of these neighboring entities to further complement the global semantic sparsity. In the experiments, we evaluate KANGAROO over various knowledge-aware and general NLP tasks in both full and few-shot learning settings, outperforming various KE-PLM training paradigms performance in closed-domains significantly.

09:00-10:30 (East Foyer / Virtual)
**#46 FedID: Federated Interactive Distillation for Large-Scale Pretraining Language Models**
*Xinge Ma, Jiangming Liu, Jin Wang and Xuejie Zhang*
The growing concerns and regulations surrounding the protection of user data privacy have necessitated decentralized training paradigms. To this end, federated learning (FL) is widely studied in user-related natural language processing (NLP). However, it suffers from several critical limitations including extensive communication overhead, inability to handle heterogeneity, and vulnerability to white-box inference attacks. Federated distillation (FD) is proposed to alleviate these limitations, but its performance is faded by confirmation bias. To tackle this issue, we propose Federated Interactive Distillation (FedID), which utilizes a small amount of labeled data retained by the server to further rectify the local models during knowledge transfer. Additionally, based on the GLUE benchmark, we develop a benchmarking framework across

multiple tasks with diverse data distributions to contribute to the research of FD in NLP community. Experiments show that our proposed FedID framework achieves the best results in homogeneous and heterogeneous federated scenarios. The code for this paper is available at: https://github.com/maxinge8698/FedID.

09:00-10:30 (East Foyer / Virtual)
**#47 VIBE: Topic-Driven Temporal Adaptation for Twitter Classification**
*Yuji Zhang, Jing Li and Wenjie Li*
Language features are evolving in real-world social media, resulting in the deteriorating performance of text classification in dynamics. To address this challenge, we study temporal adaptation, where models trained on past data are tested in the future. Most prior work focused on continued pretraining or knowledge updating, which may compromise their performance on noisy social media data. To tackle this issue, we reflect feature change via modeling latent topic evolution and propose a novel model, VIBE: Variational Information Bottleneck for Evolutions. Concretely, we first employ two Information Bottleneck (IB) regularizers to distinguish past and future topics. Then, the distinguished topics work as adaptive features via multi-task training with timestamp and class label prediction. In adaptive learning, VIBE utilizes retrieved unlabeled data from online streams created posterior to training data time. Substantial Twitter experiments on three classification tasks show that our model, with only 3% of data, significantly outperforms previous state-of-the-art continued-pretraining methods.

09:00-10:30 (East Foyer / Virtual)
**#48 Chain-of-Thought Tuning: Masked Language Models can also Think Step By Step in Natural Language Understanding**
*Caoyun Fan, Jidong Tian, Yitian Li, Wenqing Chen, Hao He and Yaohui Jin*
Chain-of-Thought (CoT) is a technique that guides Large Language Models (LLMs) to decompose complex tasks into multi-step reasoning through intermediate steps in natural language form. Briefly, CoT enables LLMs to think step by step. However, although many Natural Language Understanding (NLU) tasks also require thinking step by step, LLMs perform less well than small-scale Masked Language Models (MLMs). To migrate CoT from LLMs to MLMs, we propose Chain-of-Thought Tuning (CoTT), a two-step reasoning framework based on prompt tuning, to implement step-by-step thinking for MLMs on NLU tasks. From the perspective of CoT, CoTT's two-step framework enables MLMs to implement task decomposition; CoTT's prompt tuning allows intermediate steps to be used in natural language form. Thereby, the success of CoT can be extended to NLU tasks through MLMs. To verify the effectiveness of CoTT, we conduct experiments on two NLU tasks: hierarchical classification and relation extraction, and the results show that CoTT outperforms baselines and achieves state-of-the-art performance.

09:00-10:30 (East Foyer / Virtual)
**#49 A Fine-Grained Taxonomy of Replies to Hate Speech**
*Xinchen Yu, Ashley Zhao, Eduardo Blanco and Lingzi Hong*
Countering rather than censoring hate speech has emerged as a promising strategy to address hatred. There are many types of counterspeech in user-generated content: addressing the hateful content or its author, generic requests, well-reasoned counter arguments, insults, etc. The effectiveness of counterspeech, which we define as subsequent incivility, depends on these types. In this paper, we present a theoretically grounded taxonomy of replies to hate speech and a new corpus. We work with real, user-generated hate speech and all the replies it elicits rather than replies generated by a third party. Our analyses provide insights into the content real users reply with as well as which replies are empirically most effective. We also experiment with models to characterize the replies to hate speech, thereby opening the door to estimating whether a reply to hate speech will result in further incivility.

09:00-10:30 (East Foyer / Virtual)
**#50 Text Representation Distillation via Information Bottleneck Principle**
*Yanzhao Zhang, Dingkun Long, Zehan Li and Pengjun Xie*
Pre-trained language models (PLMs) have recently shown great success in text representation field. However, the high computational cost and high-dimensional representation of PLMs pose significant challenges for practical applications. To make models more accessible, an effective method is to distill large models into smaller representation models. In order to relieve the issue of performance degradation after distillation, we propose a novel Knowledge Distillation method called **IBKD**. This approach is motivated by the Information Bottleneck principle and aims to maximize the mutual information between the final representation of the teacher and student model, while simultaneously reducing the mutual information between the student model's representation and the input data. This enables the student model to preserve important learned information while avoiding unnecessary information, thus reducing the risk of over-fitting. Empirical studies on two main downstream applications of text representation (Semantic Textual Similarity and Dense Retrieval tasks) demonstrate the effectiveness of our proposed approach.

09:00-10:30 (East Foyer / Virtual)
**#51 BioFEG: Generate Latent Features for Biomedical Entity Linking**
*Xuhui Sui, Ying Zhang, Xiangrui Cai, Kehui Song, Baohang Zhou, Xiaojie Yuan and Wensheng Zhang*
Biomedical entity linking is an essential task in biomedical text processing, which aims to map entity mentions in biomedical text, such as clinical notes, to standard terms in a given knowledge base. However, this task is challenging due to the rarity of many biomedical entities in real-world scenarios, which often leads to a lack of annotated data for them. Limited by understanding these unseen entities, traditional biomedical entity linking models suffer from multiple types of linking errors. In this paper, we propose a novel latent feature generation framework BioFEG to address these challenges. Specifically, our BioFEG leverages domain knowledge to train a generative adversarial network, which generates latent semantic features of corresponding mentions for unseen entities. Utilizing these features, we fine-tune our entity encoder to capture fine-grained coherence information of unseen entities and better understand them. This allows models to make linking decisions more accurately, particularly for ambiguous mentions involving rare entities. Extensive experiments on the two benchmark datasets demonstrate the superiority of our proposed framework.

09:00-10:30 (East Foyer / Virtual)
**#52 NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly**
*Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan and Heng Ji*
Knowledge of norms is needed to understand and reason about acceptable behavior in human communication and interactions across sociocultural scenarios. Most computational research on norms has focused on a single culture, and manually built datasets, from non-conversational settings. We address these limitations by proposing a new framework, NormSage, to automatically extract culture-specific norms from multi-lingual conversations. NormSage uses GPT-3 prompting to 1) extract candidate norms directly from conversations and 2) provide explainable self-verification to ensure correctness and relevance. Comprehensive empirical results show the promise of our approach to extract high-quality culture-aware norms from multi-lingual conversations (English and Chinese), across several quality metrics. Further, our relevance verification can be extended to assess the adherence and violation of any norm with respect to a conversation on-the-fly, along with textual explanation. NormSage achieves an AUC of 94.6% in this grounding setup, with generated explanations matching human-written quality.

09:00-10:30 (East Foyer / Virtual)
**#53 JointMatch: A Unified Approach for Diverse and Collaborative Pseudo-Labeling to Semi-Supervised Text Classification**

*Henry Peng Zou and Cornelia Caragea*
Semi-supervised text classification (SSTC) has gained increasing attention due to its ability to leverage unlabeled data. However, existing approaches based on pseudo-labeling suffer from the issues of pseudo-label bias and error accumulation. In this paper, we propose JointMatch, a holistic approach for SSTC that addresses these challenges by unifying ideas from recent semi-supervised learning and the task of learning with noise. JointMatch adaptively adjusts classwise thresholds based on the learning status of different classes to mitigate model bias towards current easy classes. Additionally, JointMatch alleviates error accumulation by utilizing two differently initialized networks to teach each other in a cross-labeling manner. To maintain divergence between the two networks for mutual learning, we introduce a strategy that weighs more disagreement data while also allowing the utilization of high-quality agreement data for training. Experimental results on benchmark datasets demonstrate the superior performance of JointMatch, achieving a significant 5.13% improvement on average. Notably, JointMatch delivers impressive results even in the extremely-scarce-label setting, obtaining 86% accuracy on AG News with only 5 labels per class. We make our code available at https://github.com/HenryPengZou/JointMatch.

09:00-10:30 (East Foyer / Virtual)
**#54 People Make Better Edits: Measuring the Efficacy of LLM-Generated Counterfactually Augmented Data for Harmful Language Detection**
*Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst and Claudia Wagner*
NLP models are used in a variety of critical social computing tasks, such as detecting sexist, racist, or otherwise hateful content. Therefore, it is imperative that these models are robust to spurious features. Past work has attempted to tackle such spurious features using training data augmentation, including Counterfactually Augmented Data (CADs). CADs introduce minimal changes to existing training data points and flip their labels; training on them may reduce model dependency on spurious features. However, manually generating CADs can be time-consuming and expensive. Hence in this work, we assess if this task can be automated using generative NLP models. We automatically generate CADs using Polyjuice, ChatGPT, and Flan-T5, and evaluate their usefulness in improving model robustness compared to manually-generated CADs. By testing both model performance on multiple out-of-domain test sets and individual data point efficacy, our results show that while manual CADs are still the most effective, CADs generated by ChatGPT come a close second. One key reason for the lower performance of automated methods is that the changes they introduce are often insufficient to flip the original label.

09:00-10:30 (East Foyer / Virtual)
**#55 All Things Considered: Detecting Partisan Events from News Media with Cross-Article Comparison**
*Yujian Liu, Xinliang Frederick Zhang, Kaijian Zou, Ruihong Huang, Nicholas Beauchamp and Lu Wang*
Public opinion is shaped by the information news media provide, and that information in turn may be shaped by the ideological preferences of media outlets. But while much attention has been devoted to media bias via overt ideological language or topic selection, a more unobtrusive way in which the media shape opinion is via the strategic inclusion or omission of *partisan events* that may *support* one side or the other. We develop a latent variable-based framework to predict the ideology of news articles by comparing multiple articles on the same story and identifying partisan events whose inclusion or omission reveals ideology. Our experiments first validate the existence of partisan event selection, and then show that article alignment and cross-document comparison detect partisan events and article ideology better than competitive baselines. Our results reveal the high-level form of media bias, which is present even among mainstream media with strong norms of objectivity and nonpartisanship. Our codebase and dataset are available at https://github.com/launchnlp/ATC.

09:00-10:30 (East Foyer / Virtual)
**#56 Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification**
*Liam Cripwell, Joël Legrand and Claire Gardent*
Automatic evaluation for sentence simplification remains a challenging problem. Most popular evaluation metrics require multiple high-quality references – something not readily available for simplification – which makes it difficult to test performance on unseen domains. Furthermore, most existing metrics conflate simplicity with correlated attributes such as fluency or meaning preservation. We propose a new learned evaluation metric — SLE — which focuses on simplicity, outperforming almost all existing metrics in terms of correlation with human judgements.

09:00-10:30 (East Foyer / Virtual)
**#57 Efficient Grammatical Error Correction Via Multi-Task Training and Optimized Training Schedule**
*Andrey Bout, Alexander Podolskiy, Sergey Nikolenko and Irina Piontkovskaya*
Progress in neural grammatical error correction (GEC) is hindered by the lack of annotated training data. Sufficient amounts of high-quality manually annotated data are not available, so recent research has relied on generating synthetic data, pretraining on it, and then fine-tuning on real datasets; performance gains have been achieved either by ensembling or by using huge pretrained models such as XXL-T5 as the backbone. In this work, we explore an orthogonal direction: how to use available data more efficiently. First, we propose auxiliary tasks that exploit the alignment between the original and corrected sentences, such as predicting a sequence of corrections. We formulate each task as a sequence-to-sequence problem and perform multi-task training. Second, we discover that the order of datasets used for training and even individual instances within a dataset may have important effects on the final performance, so we set out to find the best training schedule. Together, these two ideas lead to significant improvements, producing results that improve state of the art with much smaller models; in particular, we outperform the best models based on T5-XXL (11B parameters) with a BART-based model (400M parameters).

09:00-10:30 (East Foyer / Virtual)
**#58 GLEN: Generative Retrieval via Lexical Index Learning**
*Sunkyung Lee, Minjin Choi and Jongwuk Lee*
Generative retrieval shed light on a new paradigm of document retrieval, aiming to directly generate the identifier of a relevant document for a query. While it takes advantage of bypassing the construction of auxiliary index structures, existing studies face two significant challenges: (i) the discrepancy between the knowledge of pre-trained language models and identifiers and (ii) the gap between training and inference that poses difficulty in learning to rank. To overcome these challenges, we propose a novel generative retrieval method, namely Generative retrieval via LExical iNdex learning (GLEN). For training, GLEN effectively exploits a dynamic lexical identifier using a two-phase index learning strategy, enabling it to learn meaningful lexical identifiers and relevance signals between queries and documents. For inference, GLEN utilizes collision-free inference, using identifier weights to rank documents without additional overhead. Experimental results prove that GLEN achieves state-of-the-art or competitive performance against existing generative retrieval methods on various benchmark datasets, e.g., NQ320k, MS MARCO, and BEIR. The code is available at https://github.com/skleee/GLEN.

09:00-10:30 (East Foyer / Virtual)
**#59 NAIL: Lexical Retrieval Indices with Efficient Non-Autoregressive Decoders**
*Livio Baldini Soares, Daniel Gillick, Jeremy R. Cole and Tom Kwiatkowski*
Neural document rerankers are extremely effective in terms of accuracy. However, the best models require dedicated hardware for serving, which is costly and often not feasible. To avoid this servingtime requirement, we present a method of capturing up to 86% of the gains of a Transformer cross-attention model with a lexicalized scoring function that only requires $10^{-6}$% of the Transformer's FLOPs per document

and can be served using commodity CPUs. When combined with a BM25 retriever, this approach matches the quality of a state-of-the art dual encoder retriever, that still requires an accelerator for query encoding. We introduce nail (Non-Autoregressive Indexing with Language models) as a model architecture that is compatible with recent encoder-decoder and decoder-only large language models, such as T5, GPT-3 and PaLM. This model architecture can leverage existing pre-trained checkpoints and can be fine-tuned for efficiently constructing document representations that do not require neural processing of queries.

09:00-10:30 (East Foyer / Virtual)
**#60 Once is Enough: A Light-Weight Cross-Attention for Fast Sentence Pair Modeling**
*Yuanhang Yang, Shiyi Qi, Chuanyi Liu, Qifan Wang, Cuiyun Gao and Zenglin Xu*
Transformer-based models have achieved great success on sentence pair modeling tasks, such as answer selection and natural language inference (NLI). These models generally perform cross-attention over input pairs, leading to prohibitive computational cost. Recent studies propose dual-encoder and late interaction architectures for faster computation. However, the balance between the expressive of cross-attention and computation speedup still needs better coordinated. To this end, this paper introduces a novel paradigm TopicAns for efficient sentence pair modeling. TopicAns involves a lightweight cross-attention mechanism. It conducts query encoding only once while modeling the query-candidate interaction in parallel. Extensive experiments conducted on four tasks demonstrate that our TopicAnscan speed up sentence pairing by over 113x while achieving comparable performance as the more expensive cross-attention models.

09:00-10:30 (East Foyer / Virtual)
**#61 GradSim: Gradient-Based Language Grouping for Effective Multilingual Training**
*Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen and Hinrich Schuetze*
Most languages of the world pose low-resource challenges to natural language processing models. With multilingual training, knowledge can be shared among languages. However, not all languages positively influence each other and it is an open research question how to select the most suitable set of languages for multilingual training and avoid negative interference among languages whose characteristics or data distributions are not compatible. In this paper, we propose GradSim, a language grouping method based on gradient similarity. Our experiments on three diverse multilingual benchmark datasets show that it leads to the largest performance gains compared to other similarity measures and it is better correlated with cross-lingual model performance. As a result, we set the new state of the art on AfriSenti, a benchmark dataset for sentiment analysis on low-resource African languages. In our extensive analysis, we further reveal that besides linguistic features, the topics of the datasets play an important role for language grouping and that lower layers of transformer models encode language-specific features while higher layers capture task-specific information.

09:00-10:30 (East Foyer / Virtual)
**#62 A Quality-based Syntactic Template Retriever for Syntactically-Controlled Paraphrase Generation**
*Xue Zhang, Songming Zhang, Yunlong Liang, Yufeng Chen, Jian Liu, Wenjuan Han and Jinan Xu*
Existing syntactically-controlled paraphrase generation (SPG) models perform promisingly with human-annotated or well-chosen syntactic templates. However, the difficulty of obtaining such templates actually hinders the practical application of SPG models. For one thing, the prohibitive cost makes it unfeasible to manually design decent templates for every source sentence. For another, the templates automatically retrieved by current heuristic methods are usually unreliable for SPG models to generate qualified paraphrases. To escape this dilemma, we propose a novel Quality-based Syntactic Template Retriever (QSTR) to retrieve templates based on the quality of the to-be-generated paraphrases. Furthermore, for situations requiring multiple paraphrases for each source sentence, we design a Diverse Templates Search (DTS) algorithm, which can enhance the diversity between paraphrases without sacrificing quality. Experiments demonstrate that QSTR can significantly surpass existing retrieval methods in generating high-quality paraphrases and even perform comparably with human-annotated templates in terms of reference-free metrics. Additionally, human evaluation and the performance on downstream tasks using our generated paraphrases for data augmentation showcase the potential of our QSTR and DTS algorithm in practical scenarios.

09:00-10:30 (East Foyer / Virtual)
**#63 Critic-Driven Decoding for Mitigating Hallucinations in Data-to-text Generation**
*Mateusz Lango and Ondrej Dusek*
Hallucination of text ungrounded in the input is a well-known problem in neural data-to-text generation. Many methods have been proposed to mitigate it, but they typically require altering model architecture or collecting additional data, and thus cannot be easily applied to an existing model. In this paper, we explore a new way to mitigate hallucinations by combining the probabilistic output of a generator language model (LM) with the output of a special "text critic" classifier, which guides the generation by assessing the match between the input data and the text generated so far. Our method does not need any changes to the underlying LM's architecture or training procedure and can thus be combined with any model and decoding operating on word probabilities. The critic does not need any additional training data, using the base LM's training data and synthetic negative examples. Our experimental results show that our method improves over the baseline on the WebNLG and OpenDialKG benchmarks.

09:00-10:30 (East Foyer / Virtual)
**#64 PromptMix: A Class Boundary Augmentation Method for Large Language Model Distillation**
*Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau and Issam H. Laradji*
Data augmentation is a widely used technique to address the problem of text classification when there is a limited amount of training data. Recent work often tackles this problem using large language models (LLMs) like GPT3 that can generate new examples given already available ones. In this work, we propose a method to generate more helpful augmented data by utilizing the LLM's abilities to follow instructions and perform few-shot classifications. Our specific PromptMix method consists of two steps: 1) generate challenging text augmentations near class boundaries; however, generating borderline examples increases the risk of false positives in the dataset, so we 2) relabel the text augmentations using a prompting-based LLM classifier to enhance the correctness of labels in the generated data. We evaluate the proposed method in challenging 2-shot and zero-shot settings on four text classification datasets: Banking77, TREC6, Subjectivity (SUBJ), and Twitter Complaints. Our experiments show that generating and, crucially, relabeling borderline examples facilitates the transfer of knowledge of a massive LLM like GPT3.5-turbo into smaller and cheaper classifiers like DistilBERT-base and BERT-base. Furthermore, 2-shot PromptMix outperforms multiple 5-shot data augmentation methods on the four datasets. Our code is available at https://github.com/ServiceNow/PromptMix-EMNLP-2023.

09:00-10:30 (East Foyer / Virtual)
**#65 Cross-Lingual Cross-Target Stance Detection with Dual Knowledge Distillation Framework**
*Ruike Zhang, Hanxuan Yang and Wenji Mao*
Stance detection aims to identify the user's attitude toward specific *targets* from text, which is an important research area in text mining and benefits a variety of application domains. Existing studies on stance detection were conducted mainly in English. Due to the low-resource problem in most non-English languages, cross-lingual stance detection was proposed to transfer knowledge from high-resource (source) language to low-resource (target) language. However, previous research has ignored the practical issue of no labeled training data available in target language. Moreover, target inconsistency in cross-lingual stance detection brings about the additional issue of unseen targets in target

language, which in essence requires the transfer of both language and target-oriented knowledge from source to target language. To tackle these challenging issues, in this paper, we propose the new task of cross-lingual cross-target stance detection and develop the first computational work with dual knowledge distillation. Our proposed framework designs a cross-lingual teacher and a cross-target teacher using the source language data and a dual distillation process that transfers the two types of knowledge to target language. To bridge the target discrepancy between languages, cross-target teacher mines target category information and generalizes it to the unseen targets in target language via category-oriented learning. Experimental results on multilingual stance datasets demonstrate the effectiveness of our method compared to the competitive baselines.

09:00-10:30 (East Foyer / Virtual)
**#66 Zero-shot Sharpness-Aware Quantization for Pre-trained Language Models**
*Miaoxi Zhu, Qihuang Zhong, Li Shen, Liang Ding, Juhua Liu, Bo Du and Dacheng Tao*
Quantization is a promising approach for reducing memory overhead and accelerating inference, especially in large pre-trained language model (PLM) scenarios. While having no access to original training data due to security and privacy concerns has emerged the demand for zero-shot quantization. Most of the cutting-edge zero-shot quantization methods primarily 1) apply to computer vision tasks, and 2) neglect of overfitting problem in the generative adversarial learning process, leading to sub-optimal performance. Motivated by this, we propose a novel zero-shot sharpness-aware quantization (ZSAQ) framework for the zero-shot quantization of various PLMs. The key algorithm in solving ZSAQ is the SAM-SGA optimization, which aims to improve the quantization accuracy and model generalization via optimizing a minimax problem. We theoretically prove the convergence rate for the minimax optimization problem and this result can be applied to other nonconvex-PL minimax optimization frameworks. Extensive experiments on 11 tasks demonstrate that our method brings consistent and significant performance gains on both discriminative and generative PLMs, i.e., up to +6.98 average score. Furthermore, we empirically validate that our method can effectively improve the model generalization.

09:00-10:30 (East Foyer / Virtual)
**#67 DNA: Denoised Neighborhood Aggregation for Fine-grained Category Discovery**
*Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Qinghua Zheng, QianYing Wang and Ping Chen*
Discovering fine-grained categories from coarsely labeled data is a practical and challenging task, which can bridge the gap between the demand for fine-grained analysis and the high annotation cost. Previous works mainly focus on instance-level discrimination to learn low-level features, but ignore semantic similarities between data, which may prevent these models learning compact cluster representations. In this paper, we propose *Denoised Neighborhood Aggregation* (DNA), a self-supervised framework that encodes semantic structures of data into the embedding space. Specifically, we retrieve $k$-nearest neighbors of a query as its positive keys to capture semantic similarities between data and then aggregate information from the neighbors to learn compact cluster representations, which can make fine-grained categories more separatable. However, the retrieved neighbors can be noisy and contain many false-positive keys, which can degrade the quality of learned embeddings. To cope with this challenge, we propose three principles to filter out these false neighbors for better representation learning. Furthermore, we theoretically justify that the learning objective of our framework is equivalent to a clustering loss, which can capture semantic similarities between data to form compact fine-grained clusters. Extensive experiments on three benchmark datasets show that our method can retrieve more accurate neighbors (21.31% accuracy improvement) and outperform state-of-the-art models by a large margin (average 9.96% improvement on three metrics). Our code and data are available at https://github.com/Lackel/DNA.

09:00-10:30 (East Foyer / Virtual)
**#68 Analysing State-Backed Propaganda Websites: a New Dataset and Linguistic Study**
*Freddy Heppell, Kalina Bontcheva and Carolina Scarton*
This paper analyses two hitherto unstudied sites sharing state-backed disinformation, Reliable Recent News (rrn.world) and WarOnFakes (waronfakes.com), which publish content in Arabic, Chinese, English, French, German, and Spanish. We describe our content acquisition methodology and perform cross-site unsupervised topic clustering on the resulting multilingual dataset. We also perform linguistic and temporal analysis of the web page translations and topics over time, and investigate articles with false publication dates. We make publicly available this new dataset of 14,053 articles, annotated with each language version, and additional metadata such as links and images. The main contribution of this paper for the NLP community is in the novel dataset which enables studies of disinformation networks, and the training of NLP tools for disinformation detection.

09:00-10:30 (East Foyer / Virtual)
**#69 CP-BCS: Binary Code Summarization Guided by Control Flow Graph and Pseudo Code**
*Tong Ye, Lingfei Wu, Tengfei Ma, Xuhong Zhang, Yangkai Du, Peiyu Liu, Shouling Ji and Wenhai Wang*
Automatically generating function summaries for binaries is an extremely valuable but challenging task, since it involves translating the execution behavior and semantics of the low-level language (assembly code) into human-readable natural language. However, most current works on understanding assembly code are oriented towards generating function names, which involve numerous abbreviations that make them still confusing. To bridge this gap, we focus on generating complete summaries for binary functions, especially for stripped binary (no symbol table and debug information in reality). To fully exploit the semantics of assembly code, we present a control flow graph and pseudo code guided binary code summarization framework called CP-BCS. CP-BCS utilizes a bidirectional instruction-level control flow graph and pseudo code that incorporates expert knowledge to learn the comprehensive binary function execution behavior and logic semantics. We evaluate CP-BCS on 3 different binary optimization levels (O1, O2, and O3) for 3 different computer architectures (X86, X64, and ARM). The evaluation results demonstrate CP-BCS is superior and significantly improves the efficiency of reverse engineering.

09:00-10:30 (East Foyer / Virtual)
**#70 Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning**
*Ximing Lu, Faeze Brahman, Peter West, Jaehun Jung, Khyathi Chandu, Abhilasha Ravichander, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Yuchen Lin, Skyler Hallinan, Lianhui Qin, Xiang Ren, Sean Welleck and Yejin Choi*
While extreme-scale language models have demonstrated exceptional performance on a variety of language tasks, the degree of control over these language models through pure prompting can often be limited. Directly fine-tuning such language models can be effective for tailoring them, but it can be either extremely costly (e.g., GPT-3) or not even feasible for the broader community (e.g., GPT-4). We propose Inference-time Policy Adapters (IPA), which efficiently tailors a language model such as GPT-3 without fine-tuning it. IPA guides a large base model during decoding time through a lightweight policy adapter trained to optimize an arbitrary user objective with reinforcement learning. On five challenging text generation tasks, such as toxicity reduction and lexically constrained generation, IPA consistently brings significant improvements over off-the-shelf language models. It outperforms competitive baseline methods, sometimes even including expensive fine-tuning. In particular, tailoring GPT-2 with IPA can outperform GPT-3, while tailoring GPT-3 with IPA brings a major performance boost over GPT-3 (and sometimes even over GPT-4). Our promising results highlight the potential of IPA as a lightweight alternative to tailoring extreme-scale language models.

09:00-10:30 (East Foyer / Virtual)
**#71 Practical Computational Power of Linear Transformers and Their Recurrent and Self-Referential Extensions**

*Kazuki Irie, Róbert Csordás and Jürgen Schmidhuber*
Recent studies of the computational power of recurrent neural networks (RNNs) reveal a hierarchy of RNN architectures, given real-time and finite-precision assumptions. Here we study auto-regressive Transformers with linearised attention, a.k.a. linear Transformers (LTs) or Fast Weight Programmers (FWPs). LTs are special in the sense that they are equivalent to RNN-like sequence processors with a fixed-size state, while they can also be expressed as the now-popular self-attention networks. We show that many well-known results for the standard Transformer directly transfer to LTs/FWPs. Our formal language recognition experiments demonstrate how recently proposed FWP extensions such as recurrent FWPs and self-referential weight matrices successfully overcome certain limitations of the LT, e.g., allowing for generalisation on the parity problem. Our code is public.

09:00-10:30 (East Foyer / Virtual)
**#72 HistAlign: Improving Context Dependency in Language Generation by Aligning with History**
*David Wan, Shiyue Zhang and Mohit Bansal*
Language models (LMs) can generate hallucinations and incoherent outputs, which highlights their weak context dependency. Cache-LMs, which augment LMs with a memory of recent history, can increase context dependency and have shown remarkable performance in diverse language generation tasks. However, we find that even with training, the performance gain stemming from the cache component of current cache-LMs is suboptimal due to the misalignment between the current hidden states and those stored in the memory. In this work, we present HistAlign, a new training approach to ensure good cache alignment such that the model receives useful signals from the history. We first prove our concept on a simple and synthetic task where the memory is essential for correct predictions, and we show that the cache component of HistAlign is better aligned and improves overall performance. Next, we evaluate HistAlign on diverse downstream language generation tasks, including prompt continuation, abstractive summarization, and data-to-text. We demonstrate that HistAlign improves text coherence and faithfulness in open-ended and conditional generation settings respectively. HistAlign is also generalizable across different model families, showcasing its strength in improving context dependency of LMs in diverse scenarios.

09:00-10:30 (East Foyer / Virtual)
**#73 Contrastive Learning of Sentence Embeddings from Scratch**
*Junlei Zhang, Zhenzhong Lan and Junxian He*
Contrastive learning has been the dominant approach to train state-of-the-art sentence embeddings. Previous studies have typically learned sentence embeddings either through the use of human-annotated natural language inference (NLI) data or via large-scale unlabeled sentences in an unsupervised manner. However, even in the case of unlabeled data, their acquisition presents challenges in certain domains due to various reasons. due to copyright restrictions, data distribution issues, and messy formats, among other factors. To address these issues, we present SynCSE, a contrastive learning framework that trains sentence embeddings with synthetic data. Specifically, we explore utilizing large language models to synthesize the required data samples for contrastive learning, including (1) producing positive and negative annotations given unlabeled sentences SynCSE-partial, and (2) generating sentences along with their corresponding annotations from scratch SynCSE-scratch. Notably, SynCSE-scratch constitutes the first contrastive learning method to learn sentence embeddings from scratch without manually collecting any data sample. Experimental results on sentence similarity and reranking tasks indicate that both SynCSE-partial and SynCSE-scratch greatly outperform unsupervised baselines, and SynCSE-partial even achieves comparable performance to the supervised models in most settings.

09:00-10:30 (East Foyer / Virtual)
**#74 Uncertainty Guided Global Memory Improves Multi-Hop Question Answering**
*Alsu Sagirova and Mikhail Burtsev*
Transformers have become the gold standard for many natural language processing tasks and, in particular, for multi-hop question answering (MHQA). This task includes processing a long document and reasoning over the multiple parts of it. The landscape of MHQA approaches can be classified into two primary categories. The first group focuses on extracting supporting evidence, thereby constraining the QA model's context to predicted facts. Conversely, the second group relies on the attention mechanism of the long input encoding model to facilitate multi-hop reasoning. However, attention-based token representations lack explicit global contextual information to connect reasoning steps. To address these issues, we propose GEMFormer, a two-stage method that first collects relevant information over the entire document to the memory and then combines it with local context to solve the task. Our experimental results show that fine-tuning a pre-trained model with memory-augmented input, including the most certain global elements, improves the model's performance on three MHQA datasets compared to the baseline. We also found that the global explicit memory contains information from supporting facts required for the correct answer.

09:00-10:30 (East Foyer / Virtual)
**#75 ClusterLLM: Large Language Models as a Guide for Text Clustering**
*Yuwei Zhang, Zihan Wang and Jingbo Shang*
We introduce ClusterLLM, a novel text clustering framework that leverages feedback from an instruction-tuned large language model, such as ChatGPT. Compared with traditional unsupervised methods that builds upon "small" embedders, ClusterLLM exhibits two intriguing advantages: (1) it enjoys the emergent capability of LLM even if its embeddings are inaccessible; and (2) it understands the user's preference on clustering through textual instruction and/or a few annotated data. First, we prompt ChatGPT for insights on clustering perspective by constructing hard triplet questions <does A better correspond to B than C>, where A, B and C are similar data points that belong to different clusters according to small embedder. We empirically show that this strategy is both effective for fine-tuning small embedder and cost-efficient to query ChatGPT. Second, we prompt ChatGPT for helps on clustering granularity by carefully designed pairwise questions <do A and B belong to the same category>, and tune the granularity from cluster hierarchies that is the most consistent with the ChatGPT answers. Extensive experiments on 14 datasets show that ClusterLLM consistently improves clustering quality, at an average cost of ~$0.6 per dataset.

09:00-10:30 (East Foyer / Virtual)
**#76 Multilingual estimation of political-party positioning: From label aggregation to long-input Transformers**
*Dmitry Nikolaev, Tanise Ceron and Sebastian Padó*
Scaling analysis is a technique in computational political science that assigns a political actor (e.g. politician or party) a score on a predefined scale based on a (typically long) body of text (e.g. a parliamentary speech or an election manifesto). For example, political scientists have often used the left–right scale to systematically analyse political landscapes of different countries. NLP methods for automatic scaling analysis can find broad application provided they (i) are able to deal with long texts and (ii) work robustly across domains and languages. In this work, we implement and compare two approaches to automatic scaling analysis of political-party manifestos: label aggregation, a pipeline strategy relying on annotations of individual statements from the manifestos, and long-input-Transformer-based models, which compute scaling values directly from raw text. We carry out the analysis of the Comparative Manifestos Project dataset across 41 countries and 27 languages and find that the task can be efficiently solved by state-of-the-art models, with label aggregation producing the best results.

09:00-10:30 (East Foyer / Virtual)
**#77 Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning**
*Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter and Jay Pujara*

Temporal knowledge graph (TKG) forecasting benchmarks challenge models to predict future facts using knowledge of past facts. In this paper, we develop an approach to use in-context learning (ICL) with large language models (LLMs) for TKG forecasting. Our extensive evaluation compares diverse baselines, including both simple heuristics and state-of-the-art (SOTA) supervised models, against pre-trained LLMs across several popular benchmarks and experimental settings. We observe that naive LLMs perform on par with SOTA models, which employ carefully designed architectures and supervised training for the forecasting task, falling within the (-3.6%, +1.5%) Hits@1 margin relative to the median performance. To better understand the strengths of LLMs for forecasting, we explore different approaches for selecting historical facts, constructing prompts, controlling information propagation, and parsing outputs into a probability distribution. A surprising finding from our experiments is that LLM performance endures ($\pm$0.4% Hit@1) even when semantic information is removed by mapping entities/relations to arbitrary numbers, suggesting that prior semantic knowledge is unnecessary; rather, LLMs can leverage the symbolic patterns in the context to achieve such a strong performance. Our analysis also reveals that ICL enables LLMs to learn irregular patterns from the historical context, going beyond frequency and recency biases

09:00-10:30 (East Foyer / Virtual)
**#78 Hierarchical Pretraining on Multimodal Electronic Health Records**
*Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang and Fenglong Ma*
Pretraining has proven to be a powerful technique in natural language processing (NLP), exhibiting remarkable success in various NLP downstream tasks. However, in the medical domain, existing pretrained models on electronic health records (EHR) fail to capture the hierarchical nature of EHR data, limiting their generalization capability across diverse downstream tasks using a single pretrained model. To tackle this challenge, this paper introduces a novel, general, and unified pretraining framework called MedHMP, specifically designed for hierarchically multimodal EHR data. The effectiveness of the proposed MedHMP is demonstrated through experimental results on eight downstream tasks spanning three levels. Comparisons against eighteen baselines further highlight the efficacy of our approach.

09:00-10:30 (East Foyer / Virtual)
**#79 Multi-Task Knowledge Distillation with Embedding Constraints for Scholarly Keyphrase Boundary Classification**
*Seo Yeon Park and Cornelia Caragea*
The task of scholarly keyphrase boundary classification aims at identifying keyphrases from scientific papers and classifying them with their types from a set of predefined classes (e.g., task, process, or material). Despite the importance of keyphrases and their types in many downstream applications including indexing, searching, and question answering over scientific documents, scholarly keyphrase boundary classification is still an under-explored task. In this work, we propose a novel embedding constraint on multi-task knowledge distillation which enforces the teachers (single-task models) and the student (multi-task model) similarity in the embedding space. Specifically, we enforce that the student model is trained not only to imitate the teachers' output distribution over classes, but also to produce language representations that are similar to those produced by the teachers. Our results show that the proposed approach outperforms previous works and strong baselines on three datasets of scientific documents.

09:00-10:30 (East Foyer / Virtual)
**#80 What to Read in a Contract? Party-Specific Summarization of Legal Obligations, Entitlements, and Prohibitions**
*Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan and Rachel Rudinger*
Reviewing and comprehending key obligations, entitlements, and prohibitions in legal contracts can be a tedious task due to their length and domain-specificity. Furthermore, the key rights and duties requiring review vary for each contracting party. In this work, we propose a new task of *party-specific* extractive summarization for legal contracts to facilitate faster reviewing and improved comprehension of rights and duties. To facilitate this, we curate a dataset comprising of party-specific pairwise importance comparisons annotated by legal experts, covering $\sim$293K sentence pairs that include obligations, entitlements, and prohibitions extracted from lease agreements. Using this dataset, we train a pairwise importance ranker and propose a pipeline-based extractive summarization system that generates a party-specific contract summary. We establish the need for incorporating domain-specific notions of importance during summarization by comparing our system against various baselines using both automatic and human evaluation methods.

09:00-10:30 (East Foyer / Virtual)
**#81 CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation**
*Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu and Diyi Yang*
Annotated data plays a critical role in Natural Language Processing (NLP) in training models and evaluating their performance. Given recent developments in Large Language Models (LLMs), models such as ChatGPT demonstrate zero-shot capability on many text-annotation tasks, comparable with or even exceeding human annotators. Such LLMs can serve as alternatives for manual annotation, due to lower costs and higher scalability. However, limited work has leveraged LLMs as complementary annotators, nor explored how annotation work is best allocated among humans and LLMs to achieve both quality and cost objectives. We propose CoAnnotating, a novel paradigm for Human-LLM co-annotation of unstructured texts at scale. Under this framework, we utilize uncertainty to estimate LLMs' annotation capability. Our empirical study shows CoAnnotating to be an effective means to allocate work from results on different datasets, with up to 21% performance improvement over random baseline. For code implementation, see https://github.com/SALT-NLP/CoAnnotating.

09:00-10:30 (East Foyer / Virtual)
**#82 Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning**
*Saibo Geng, Martin Josifoski, Maxime Peyrard and Robert West*
Despite their impressive performance, large language models (LMs) still struggle with reliably generating complex output structures when not finetuned to follow the required output format exactly. To address this issue, grammar-constrained decoding (GCD) can be used to control the generation of LMs, guaranteeing that the output follows a given structure. Most existing GCD methods are, however, limited to specific tasks, such as parsing or code generation. In this work, we demonstrate that formal grammars can describe the output space for a much wider range of tasks and argue that GCD can serve as a unified framework for structured NLP tasks in general. For increased flexibility, we introduce input-dependent grammars, which allow the grammar to depend on the input and thus enable the generation of different output structures for different inputs. We then empirically demonstrate the power and flexibility of GCD-enhanced LMs on (1) information extraction, (2) entity disambiguation, and (3) constituency parsing. Our results indicate that grammar-constrained LMs substantially outperform unconstrained LMs or even beat task-specific finetuned models. Grammar constraints thus hold great promise for harnessing off-the-shelf LMs for a wide range of structured NLP tasks, especially where training data is scarce or finetuning is expensive. Code and data: https://github.com/epfl-dlab/GCD.

09:00-10:30 (East Foyer / Virtual)
**#83 Reward-Augmented Decoding: Efficient Controlled Text Generation With a Unidirectional Reward Model**
*Haikang Deng and Colin Raffel*
While large language models have proven effective in a huge range of downstream applications, they often generate text that is problematic or lacks a desired attribute. In this paper, we introduce Reward-Augmented Decoding (RAD), a text generation procedure that uses a small unidirectional reward model to encourage a language model to generate text that has certain properties. Specifically, RAD uses the reward

model to score generations as they are produced and rescales sampling probabilities to favor high-reward tokens. By using a unidirectional reward model, RAD can cache activations from prior generation steps to decrease computational overhead. Through experiments on generating non-toxic and sentiment-controlled text, we demonstrate that RAD performs best among methods that change only the generation procedure and matches the performance of state-of-the-art methods that involve re-training the language model. We further validate that RAD is effective on very large language models while incurring a minimal computational overhead.

09:00-10:30 (East Foyer / Virtual)
**#84 Harnessing Black-Box Control to Boost Commonsense in LM's Generation**
*Yufei Tian, Felix Zhang and Nanyun Peng*
Large language models (LLMs) such as GPT-3 have demonstrated a strong capability to generate coherent and contextually relevant text. However, amidst their successes, a crucial issue persists: their generated outputs still lack commonsense at times. Moreover, fine-tuning the entire LLM towards more commonsensical outputs is computationally expensive if not infeasible. In this paper, we present a computation-efficient framework that steers a frozen Pre-Trained Language Model (PTLM) towards more commonsensical generation (i.e., producing a plausible output that incorporates a list of concepts in a meaningful way). Specifically, we first construct a reference-free evaluator that assigns a sentence with a commonsensical score by grounding the sentence to a dynamic commonsense knowledge base from four different relational aspects. We then use the scorer as the oracle for commonsense knowledge, and extend the controllable generation method called NADO to train an auxiliary head that guides a fixed PTLM to better satisfy the oracle. We test our framework on a series of GPT-2-, Flan-T5-, and Alpaca-based language models (LMs) on two constrained concept-to-sentence benchmarks. Human evaluation results demonstrate that our method consistently leads to the most commonsensical outputs.

09:00-10:30 (East Foyer / Virtual)
**#85 A Scalable Framework for Table of Contents Extraction from Complex ESG Annual Reports**
*Xinyu Wang, Lin Gui and Yulan He*
Table of contents (ToC) extraction centres on structuring documents in a hierarchical manner. In this paper, we propose a new dataset, ESG-Doc, comprising 1,093 ESG annual reports from 563 companies spanning from 2001 to 2022. These reports pose significant challenges due to their diverse structures and extensive length. To address these challenges, we propose a new framework for Toc extraction, consisting of three steps: (1) Constructing an initial tree of text blocks based on reading order and font sizes; (2) Modelling each tree node (or text block) independently by considering its contextual information captured in node-centric subtree; (3) Modifying the original tree by taking appropriate action on each tree node (Keep, Delete, or Move). This construction-modelling-modification (CMM) process offers several benefits. It eliminates the need for pairwise modelling of section headings as in previous approaches, making document segmentation practically feasible. By incorporating structured information, each section heading can leverage both local and long-distance context relevant to itself. Experimental results show that our approach outperforms the previous state-of-the-art baseline with a fraction of running time. Our framework proves its scalability by effectively handling documents of any length.

09:00-10:30 (East Foyer / Virtual)
**#86 Ling-CL: Understanding NLP Models through Linguistic Curricula**
*Mohamed Elgaar and Hadi Amiri*
We employ a characterization of linguistic complexity from psycholinguistic and language acquisition research to develop data-driven curricula to understand the underlying linguistic knowledge that models learn to address NLP tasks. The novelty of our approach is in the development of linguistic curricula derived from data, existing knowledge about linguistic complexity, and model behavior during training. Through the evaluation of several benchmark NLP datasets, our curriculum learning approaches identify sets of linguistic metrics (indices) that inform the challenges and reasoning required to address each task. Our work will inform future research in all NLP areas, allowing linguistic complexity to be considered early in the research and development process. In addition, our work prompts an examination of gold standards and fair evaluation in NLP.

09:00-10:30 (East Foyer / Virtual)
**#87 Distance-Based Propagation for Efficient Knowledge Graph Reasoning**
*Harry Shomer, Yao Ma, Juanhui Li, Bo Wu, Charu C. Aggarwal and Jiliang Tang*
Knowledge graph completion (KGC) aims to predict unseen edges in knowledge graphs (KGs), resulting in the discovery of new facts. A new class of methods have been proposed to tackle this problem by aggregating path information. These methods have shown tremendous ability in the task of KGC. However they are plagued by efficiency issues. Though there are a few recent attempts to address this through learnable path pruning, they often sacrifice the performance to gain efficiency. In this work, we identify two intrinsic limitations of these methods that affect the efficiency and representation quality. To address the limitations, we introduce a new method, TAGNet, which is able to efficiently propagate information. This is achieved by only aggregating paths in a fixed window for each source-target pair. We demonstrate that the complexity of TAGNet is independent of the number of layers. Extensive experiments demonstrate that TAGNet can cut down on the number of propagated messages by as much as 90% while achieving competitive performance on multiple KG datasets.

09:00-10:30 (East Foyer / Virtual)
**#88 Debiasing Made State-of-the-art: Revisiting the Simple Seed-based Weak Supervision for Text Classification**
*Chengyu Dong, Zihan Wang and Jingbo Shang*
Recent advances in weakly supervised text classification mostly focus on designing sophisticated methods to turn high-level human heuristics into quality pseudo-labels. In this paper, we revisit the seed matching-based method, which is arguably the simplest way to generate pseudo-labels, and show that its power was greatly underestimated. We show that the limited performance of seed matching is largely due to the label bias injected by the simple seed-match rule, which prevents the classifier from learning reliable confidence for selecting high-quality pseudo-labels. Interestingly, simply deleting the seed words present in the matched input texts can mitigate the label bias and help learn better confidence. Subsequently, the performance achieved by seed matching can be improved significantly, making it on par with or even better than the state-of-the-art. Furthermore, to handle the case when the seed words are not made known, we propose to simply delete the word tokens in the input text randomly with a high deletion ratio. Remarkably, seed matching equipped with this random deletion method can often achieve even better performance than that with seed deletion.

09:00-10:30 (East Foyer / Virtual)
**#89 Unlearn What You Want to Forget: Efficient Unlearning for LLMs**
*Jiaao Chen and Diyi Yang*
Large language models (LLMs) have achieved significant progress from pre-training on and memorizing a wide range of textual data, however, this process might suffer from privacy issues and violations of data protection regulations. As a result, the ability to easily remove data related to individual users from such models while not deteriorating their predictive quality after the removal becomes increasingly important. To address these issues, in this work, we propose an efficient unlearning framework that could efficiently update LLMs without having to retrain the whole model after data removals, by introducing lightweight unlearning layers learned with a selective teacher-student objective into the transformers. In addition, we introduce a fusion mechanism to effectively combine different unlearning layers that learns to forget different

sets of data to handle a sequence of forgetting operations. Experiments on classification and generation tasks demonstrate the effectiveness of our proposed methods compared to the state-of-the-art baselines.

09:00-10:30 (East Foyer / Virtual)
**#90 A Cheaper and Better Diffusion Language Model with Soft-Masked Noise**
*Jiaao Chen, Aston Zhang, Mu Li, Alex Smola and Diyi Yang*
Diffusion models that are based on iterative denoising have been recently proposed and leveraged in various generation tasks like image generation. Whereas, as a way inherently built for continuous data, existing diffusion models still have some limitations in modeling discrete data, e.g., languages. For example, the generally used Gaussian noise can not handle the discrete corruption well, and the objectives in continuous spaces fail to be stable for textual data in the diffusion process especially when the dimension is high. To alleviate these issues, we introduce a novel diffusion model for language modeling, Masked-Diffuse LM, with lower training cost and better performances, inspired by linguistic features in languages. Specifically, we design a linguistic-informed forward process which adds corruptions to the text through strategically soft-masking to better noise the textual data. Also, we directly predict the categorical distribution with cross-entropy loss function in every diffusion step to connect the continuous space and discrete space in a more efficient and straightforward way. Through experiments on 5 controlled generation tasks, we demonstrate that our Masked-Diffuse LM can achieve better generation quality than the state-of-the-art diffusion models with better efficiency.

09:00-10:30 (East Foyer / Virtual)
**#91 Natural Language Decompositions of Implicit Content Enable Better Text Representations**
*Alexander Hoyle, Rupak Sarkar, Pranav Goel and Philip Resnik*
When people interpret text, they rely on inferences that go beyond the observed language itself. Inspired by this observation, we introduce a method for the analysis of text that takes implicitly communicated content explicitly into account. We use a large language model to produce sets of propositions that are inferentially related to the text that has been observed, then validate the plausibility of the generated content via human judgments. Incorporating these explicit representations of implicit content proves useful in multiple problem settings that involve the human interpretation of utterances: assessing the similarity of arguments, making sense of a body of opinion data, and modeling legislative behavior. Our results suggest that modeling the meanings behind observed language, rather than the literal text alone, is a valuable direction for NLP and particularly its applications to social science.

09:00-10:30 (East Foyer / Virtual)
**#92 GeoLM: Empowering Language Models for Geospatially Grounded Language Understanding**
*Zekun Li, Wenxuan Zhou, Yao-Yi Chiang and Muhao Chen*
Humans subconsciously engage in geospatial reasoning when reading articles. We recognize place names and their spatial relations in text and mentally associate them with their physical locations on Earth. Although pretrained language models can mimic this cognitive process using linguistic context, they do not utilize valuable geospatial information in large, widely available geographical databases, e.g., OpenStreetMap. This paper introduces GeoLM, a geospatially grounded language model that enhances the understanding of geo-entities in natural language. GeoLM leverages geo-entity mentions as anchors to connect linguistic information in text corpora with geospatial information extracted from geographical databases. GeoLM connects the two types of context through contrastive learning and masked language modeling. It also incorporates a spatial coordinate embedding mechanism to encode distance and direction relations to capture geospatial context. In the experiment, we demonstrate that GeoLM exhibits promising capabilities in supporting toponym recognition, toponym linking, relation extraction, and geo-entity typing, which bridge the gap between natural language processing and geospatial sciences. The code is publicly available at https://github.com/knowledge-computing/geolm.

09:00-10:30 (East Foyer / Virtual)
**#93 JASMINE: Arabic GPT Models for Few-Shot Learning**
*El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, Alcides Alcoba Inciarte and Md Tawkat Islam Khondaker*
Scholarship on generative pretraining (GPT) remains acutely Anglocentric, leaving serious gaps in our understanding of the whole class of autoregressive models. For example, we have little knowledge about the potential of these models and their societal impacts in diverse linguistic and cultural settings. We alleviate this issue for Arabic, a wide collection of languages and dialectal varieties with more than 400 million population, by introducing JASMINE. JASMINE is a suite of powerful Arabic autoregressive Transformer language models ranging in size between 300 million-6.7 billion parameters pretrained on a large and diverse dataset ( 235 GB of text). We also carefully design and release a comprehensive benchmark for both automated and human evaluation of Arabic autoregressive models, with coverage of potential social biases, harms, and toxicity. Using our novel benchmark, we evaluate JASMINE extensively showing powerful performance intrinsically as well as in few-shot learning on a wide range of NLP tasks. We aim to responsibly release our models and evaluation benchmark with interested researchers, along with code for experimenting with them.

09:00-10:30 (East Foyer / Virtual)
**#94 Semantic matching for text classification with complex class descriptions**
*Brian M De Silva, Kuan-Wen Huang, Gwang Gook Lee, Karen Hovsepian, Yan Xu and Mingwei Shen*
Text classifiers are an indispensable tool for machine learning practitioners, but adapting them to new classes is expensive. To reduce the cost of new classes, previous work exploits class descriptions and/or labels from existing classes. However, these approaches leave a gap in the model development cycle as they support either zero- or few-shot learning, but not both. Existing classifiers either do not work on zero-shot problems, or fail to improve much with few-shot labels. Further, prior work is aimed at concise class descriptions, which may be insufficient for complex classes. We overcome these shortcomings by casting text classification as a matching problem, where a model matches examples with relevant class descriptions. This formulation lets us leverage labels and complex class descriptions to perform zero- and few-shot learning on new classes. We compare this approach with numerous baselines on text classification tasks with complex class descriptions and find that it achieves strong zero-shot performance and scales well with few-shot samples, beating strong baselines by 22.48% (average precision) in the 10-shot setting. Furthermore, we extend the popular Model-Agnostic Meta-Learning algorithm to the zero-shot matching setting and show it improves zero-shot performance by 4.29%. Our results show that expressing text classification as a matching problem is a cost-effective way to address new classes. This strategy enables zero-shot learning for cold-start scenarios and few-shot learning so the model can improve until it is capable enough to deploy.

09:00-10:30 (East Foyer / Virtual)
**#95 Hallucination Mitigation in Natural Language Generation from Large-Scale Open-Domain Knowledge Graphs**
*Xiao Shi, Zhengyuan Zhu, Zeyu Zhang and Chengkai Li*
In generating natural language descriptions for knowledge graph triples, prior works used either small-scale, human-annotated datasets or datasets with limited variety of graph shapes, e.g., those having mostly star graphs. Graph-to-text models trained and evaluated on such datasets are largely not assessed for more realistic large-scale, open-domain settings. We introduce a new dataset, GraphNarrative, to fill this gap. Fine-tuning transformer-based pre-trained language models has achieved state-of-the-art performance among graph-to-text models. However, this method suffers from information hallucination—the generated text may contain fabricated facts not present in input graphs. We

propose a novel approach that, given a graph-sentence pair in GraphNarrative, trims the sentence to eliminate portions that are not present in the corresponding graph, by utilizing the sentence's dependency parse tree. Our experiment results verify this approach using models trained on GraphNarrative and existing datasets. The dataset, source code, and trained models are released at https://github.com/idirlab/graphnarrator.

09:00-10:30 (East Foyer / Virtual)
**#96 Evaluating Cross-Domain Text-to-SQL Models and Benchmarks**
*Mohammadreza Pourreza and Davood Rafiei*
Text-to-SQL benchmarks play a crucial role in evaluating the progress made in the field and the ranking of different models. However, accurately matching a model-generated SQL query to a reference SQL query in a benchmark fails for various reasons, such as underspecified natural language queries, inherent assumptions in both model-generated and reference queries, and the non-deterministic nature of SQL output under certain conditions. In this paper, we conduct an extensive study of several prominent cross-domain text-to-SQL benchmarks and re-evaluate some of the top-performing models within these benchmarks, by both manually evaluating the SQL queries and rewriting them in equivalent expressions. Our evaluation reveals that attaining a perfect performance on these benchmarks is unfeasible due to the multiple interpretations that can be derived from the provided samples. Furthermore, we find that the true performance of the models is underestimated and their relative performance changes after a re-evaluation. Most notably, our evaluation reveals a surprising discovery: a recent GPT4-based model surpasses the gold standard reference queries in the Spider benchmark in our human evaluation. This finding highlights the importance of interpreting benchmark evaluations cautiously, while also acknowledging the critical role of additional independent evaluations in driving advancements in the field.

09:00-10:30 (East Foyer / Virtual)
**#97 PAC-tuning: Fine-tuning Pre-trained Language Models with PAC-driven Perturbed Gradient Descent**
*Guangliang Liu, Zhiyu Xue, Xitong Zhang, Kristen Johnson and Rongrong Wang*
Fine-tuning pretrained language models (PLMs) for downstream tasks is a large-scale optimization problem, in which the choice of the training algorithm critically determines how well the trained model can generalize to unseen test data, especially in the context of few-shot learning. To achieve good generalization performance and avoid overfitting, techniques such as data augmentation and pruning are often applied. However, adding these regularizations necessitates heavy tuning of the hyperparameters of optimization algorithms, such as the popular Adam optimizer. In this paper, we propose a two-stage fine-tuning method, PAC-tuning, to address this optimization challenge. First, based on PAC-Bayes training, PAC-tuning directly minimizes the PAC-Bayes generalization bound to learn proper parameter distribution. Second, PAC-tuning modifies the gradient by injecting noise with the variance learned in the first stage into the model parameters during training, resulting in a variant of perturbed gradient descent (PGD). In the past, the few-shot scenario posed difficulties for PAC-Bayes training because the PAC-Bayes bound, when applied to large models with limited training data, might not be stringent. Our experimental results across 5 GLUE benchmark tasks demonstrate that PAC-tuning successfully handles the challenges of fine-tuning tasks and outperforms strong baseline methods by a visible margin, further confirming the potential to apply PAC training for any other settings where the Adam optimizer is currently used for training.

09:00-10:30 (East Foyer / Virtual)
**#98 Meta-Learning Online Adaptation of Language Models**
*Nathan Zixia Hu, Eric Mitchell, Christopher D Manning and Chelsea Finn*
Large language models encode impressively broad world knowledge in their parameters. However, the knowledge in static language models falls out of date, limiting the model's effective "shelf life." While online fine-tuning can reduce this degradation, we find that naively fine-tuning on a stream of documents leads to a low level of information uptake. We hypothesize that online fine-tuning does not sufficiently attend to important information. That is, the gradient signal from important tokens representing factual information is drowned out by the gradient from inherently noisy tokens, suggesting that a dynamic, context-aware learning rate may be beneficial. We therefore propose learning which tokens to upweight. We meta-train a small, autoregressive model to reweight the language modeling loss for each token during online fine-tuning, with the objective of maximizing the out-of-date base question-answering model's ability to answer questions about a document after a single weighted gradient step. We call this approach Context-aware Meta-learned Loss Scaling (CaMeLS). Across three different distributions of documents, our experiments find that CaMeLS provides substantially improved information uptake on streams of thousands of documents compared with standard fine-tuning and baseline heuristics for reweighting token losses.

09:00-10:30 (East Foyer / Virtual)
**#99 CodeBERTScore: Evaluating Code Generation with Pretrained Models of Code**
*Shuyan Zhou, Uri Alon, Sumit Agarwal and Graham Neubig*
Since the rise of neural natural-language-to-code models (NL→Code) that can generate long expressions and statements rather than a single next-token, one of the major problems has been reliably evaluating their generated output. In this paper, we propose CodeBERTScore: an evaluation metric for code generation, which builds on BERTScore (Zhang et al., 2020). Instead of encoding only the generated tokens as in BERTScore, CodeBERTScore also encodes the natural language input preceding the generated code, thus modeling the consistency between the generated code and its given natural language context as well. We perform an extensive evaluation of CodeBERTScore across four programming languages. We find that CodeBERTScore achieves a higher correlation with human preference and with functional correctness than all existing metrics. That is, generated code that receives a higher score by CodeBERTScore is more likely to be preferred by humans, as well as to function correctly when executed. We release five language-specific pretrained models to use with our publicly available code. Our language-specific models have been downloaded more than **1,000,000** times from the Huggingface Hub. Our code and data are available at https://github.com/neulab/code-bert-score

09:00-10:30 (East Foyer / Virtual)
**#100 Identifying Informational Sources in News Articles**
*Alexander Spangher, Nanyun Peng, Emilio Ferrara and Jonathan May*
News articles are driven by the informational sources journalists use in reporting. Modeling when, how and why sources get used together in stories can help us better understand the information we consume and even help journalists with the task of producing it. In this work, we take steps toward this goal by constructing the largest and widest-ranging annotated dataset, to date, of informational sources used in news writing. We first show that our dataset can be used to train high-performing models for information detection and source attribution. Then, we introduce a novel task, source prediction, to study the compositionality of sources in news articles – i.e. how they are chosen to complement each other. We show good modeling performance on this task, indicating that there is a pattern to the way different sources are used *together* in news storytelling. This insight opens the door for a focus on sources in narrative science (i.e. planning-based language generation) and computational journalism (i.e. a source-recommendation system to aid journalists writing stories). All data and model code can be found at https://github.com/alex2awesome/source-exploration.

09:00-10:30 (East Foyer / Virtual)
**#101 Model-tuning Via Prompts Makes NLP Models Adversarially Robust**
*Mrigank Raman, Pratyush Maini, J Zico Kolter, Zachary Chase Lipton and Danish Pruthi*

In recent years, NLP practitioners have converged on the following practice: (i) import an off-the-shelf pretrained (masked) language model; (ii) append a multilayer perceptron atop the CLS token's hidden representation (with randomly initialized weights); and (iii) fine-tune the entire model on a downstream task (MLP-FT). This procedure has produced massive gains on standard NLP benchmarks, but these models remain brittle, even to mild adversarial perturbations. In this work, we demonstrate surprising gains in adversarial robustness enjoyed by Model-tuning Via Prompts (MVP), an alternative method of adapting to downstream tasks. Rather than appending an MLP head to make output prediction, MVP appends a prompt template to the input, and makes prediction via text infilling/completion. Across 5 NLP datasets, 4 adversarial attacks, and 3 different models, MVP improves performance against adversarial substitutions by an average of 8% over standard methods and even outperforms adversarial training-based state-of-art defenses by 3.5%. By combining MVP with adversarial training, we achieve further improvements in adversarial robustness while maintaining performance on unperturbed examples. Finally, we conduct ablations to investigate the mechanism underlying these gains. Notably, we find that the main causes of vulnerability of MLP-FT can be attributed to the misalignment between pre-training and fine-tuning tasks, and the randomly initialized MLP parameters.

09:00-10:30 (East Foyer / Virtual)
**#102 Poisoning Retrieval Corpora by Injecting Adversarial Passages**
*Zexuan Zhong, Ziqing Huang, Alexander Wettig and Danqi Chen*
Dense retrievers have achieved state-of-the-art performance in various information retrieval tasks, but to what extent can they be safely deployed in real-world applications? In this work, we propose a novel attack for dense retrieval systems in which a malicious user generates a small number of adversarial passages by perturbing discrete tokens to maximize similarity with a provided set of training queries. When these adversarial passages are inserted into a large retrieval corpus, we show that this attack is highly effective in fooling these systems to retrieve them for queries that were not seen by the attacker. More surprisingly, these adversarial passages can directly generalize to out-of-domain queries and corpora with a high success attack rate — for instance, we find that 50 generated passages optimized on Natural Questions can mislead >94% of questions posed in financial documents or online forums. We also benchmark and compare a range of state-of-the-art dense retrievers, both unsupervised and supervised. Although different systems exhibit varying levels of vulnerability, we show they can all be successfully attacked by injecting up to 500 passages, a small fraction compared to a retrieval corpus of millions of passages.

09:00-10:30 (East Foyer / Virtual)
**#103 PIEClass: Weakly-Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training**
*Yunyi Zhang, Minhao Jiang, Yu Meng, Yu Zhang and Jiawei Han*
Weakly-supervised text classification trains a classifier using the label name of each target class as the only supervision, which largely reduces human annotation efforts. Most existing methods first use the label names as static keyword-based features to generate pseudo labels, which are then used for final classifier training. While reasonable, such a commonly adopted framework suffers from two limitations: (1) keywords can have different meanings in different contexts and some text may not have any keyword, so keyword matching can induce noisy and inadequate pseudo labels; (2) the errors made in the pseudo label generation stage will directly propagate to the classifier training stage without a chance of being corrected. In this paper, we propose a new method, PIEClass, consisting of two modules: (1) a pseudo label acquisition module that uses zero-shot prompting of pre-trained language models (PLM) to get pseudo labels based on contextualized text understanding beyond static keyword matching, and (2) a noise-robust iterative ensemble training module that iteratively trains classifiers and updates pseudo labels by utilizing two PLM fine-tuning methods that regularize each other. Extensive experiments show that PIEClass achieves overall better performance than existing strong baselines on seven benchmark datasets and even achieves similar performance to fully-supervised classifiers on sentiment classification tasks.

09:00-10:30 (East Foyer / Virtual)
**#104 On the Automatic Generation and Simplification of Children's Stories**
*Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga and Katharina von der Wense*
With recent advances in large language models (LLMs), the concept of automatically generating children's educational materials has become increasingly realistic. Working toward the goal of age-appropriate simplicity in generated educational texts, we first examine the ability of several popular LLMs to generate stories with properly adjusted lexical and readability levels. We find that, in spite of the growing capabilities of LLMs, they do not yet possess the ability to limit their vocabulary to levels appropriate for younger age groups. As a second experiment, we explore the ability of state-of-the-art lexical simplification models to generalize to the domain of children's stories and, thus, create an efficient pipeline for their automatic generation. In order to test these models, we develop a dataset of child-directed lexical simplification instances, with examples taken from the LLM-generated stories in our first experiment. We find that, while the strongest-performing current lexical simplification models do not perform as well on material designed for children due to their reliance on large language models behind the scenes, some models that still achieve fairly strong results on general data can mimic or even improve their performance on children-directed data with proper fine-tuning, which we conduct using our newly created child-directed simplification dataset.

09:00-10:30 (East Foyer / Virtual)
**#105 Unifying Discrete and Continuous Representations for Unsupervised Paraphrase Generation**
*Mingfeng Xue, Dayiheng Liu, Wenqiang Lei, Jie Fu, Jian Lan, Mei Li, Baosong Yang, Jun Xie, Yidan Zhang, Dezhong Peng and Jiancheng Lv*
Unsupervised paraphrase generation is a challenging task that benefits a variety of downstream NLP applications. Current unsupervised methods for paraphrase generation typically employ round-trip translation or denoising, which require translation corpus and result in paraphrases overly similar to the original sentences in surface structure. Most of these methods lack explicit control over the similarity between the original and generated sentences, and the entities are also less correctly kept. To obviate the reliance on translation data and prompt greater variations in surface structure, we propose a self-supervised pseudo-data construction method that generates diverse pseudo-paraphrases in distinct surface structures for a given sentence. To control the similarity and generate accurate entities, we propose an unsupervised paraphrasing model that encodes the sentence meaning and the entities with discrete and continuous variables, respectively. The similarity can be controlled by sampling discrete variables and the entities are kept substantially accurate due to the specific modeling of entities using continuous variables. Experimental results on two benchmark datasets demonstrate the advantages of our pseudo-data construction method compared to round-trip translation, and the superiority of our paraphrasing model over the state-of-the-art unsupervised methods.

09:00-10:30 (East Foyer / Virtual)
**#106 ULF: Unsupervised Labeling Function Correction using Cross-Validation for Weak Supervision**
*Anastasiia Sedova and Benjamin Roth*
A cost-effective alternative to manual data labeling is weak supervision (WS), where data samples are automatically annotated using a predefined set of labeling functions (LFs), rule-based mechanisms that generate artificial labels for the associated classes. In this work, we investigate noise reduction techniques for WS based on the principle of k-fold cross-validation. We introduce a new algorithm ULF for Unsupervised Labeling Function correction, which denoises WS data by leveraging models trained on all but some LFs to identify and correct biases specific to the held-out LFs. Specifically, ULF refines the allocation of LFs to classes by re-estimating this assignment on highly reliable cross-validated samples. Evaluation on multiple datasets confirms ULF's effectiveness in enhancing WS learning without the need for manual labeling.

09:00-10:30 (East Foyer / Virtual)
**#107 Enhancing Structured Evidence Extraction for Fact Verification**
*Zirui Wu, Nan Hu and Yansong Feng*
Open-domain fact verification is the task of verifying claims in natural language texts against extracted evidence. FEVEROUS is a benchmark that requires extracting and integrating both unstructured and structured evidence to verify a given claim. Previous models suffer from low recall of structured evidence extraction, i.e., table extraction and cell selection. In this paper, we propose a simple but effective method to enhance the extraction of structured evidence by leveraging the row and column semantics of tables. Our method comprises two components: (i) a coarse-grained table extraction module that selects tables based on rows and columns relevant to the claim and (ii) a fine-grained cell selection graph that combines both formats of evidence and enables multi-hop and numerical reasoning. We evaluate our method on FEVEROUS and achieve an evidence recall of $60.01\%$ on the test set, which is $6.14\%$ higher than the previous state-of-the-art performance. Our results demonstrate that our method can extract tables and select cells effectively, and provide better evidence sets for verdict prediction. Our code is released at https://github.com/WilliamZR/see-st

09:00-10:30 (East Foyer / Virtual)
**#108 Enhancing the Ranking Context of Dense Retrieval through Reciprocal Nearest Neighbors**
*George Zerveas, Navid Rekabsaz and Carsten Eickhoff*
Sparse annotation poses persistent challenges to training dense retrieval models; for example, it distorts the training signal when unlabeled relevant documents are used spuriously as negatives in contrastive learning. To alleviate this problem, we introduce evidence-based label smoothing, a novel, computationally efficient method that prevents penalizing the model for assigning high relevance to false negatives. To compute the target relevance distribution over candidate documents within the ranking context of a given query, we assign a non-zero relevance probability to those candidates most similar to the ground truth based on the degree of their similarity to the ground-truth document(s). To estimate relevance we leverage an improved similarity metric based on reciprocal nearest neighbors, which can also be used independently to rerank candidates in post-processing. Through extensive experiments on two large-scale ad hoc text retrieval datasets, we demonstrate that reciprocal nearest neighbors can improve the ranking effectiveness of dense retrieval models, both when used for label smoothing, as well as for reranking. This indicates that by considering relationships between documents and queries beyond simple geometric distance we can effectively enhance the ranking context.

09:00-10:30 (East Foyer / Virtual)
**#109 System Combination via Quality Estimation for Grammatical Error Correction**
*Muhammad Reza Qorib and Hwee Tou Ng*
Quality estimation models have been developed to assess the corrections made by grammatical error correction (GEC) models when the reference or gold-standard corrections are not available. An ideal quality estimator can be utilized to combine the outputs of multiple GEC systems by choosing the best subset of edits from the union of all edits proposed by the GEC base systems. However, we found that existing GEC quality estimation models are not good enough in differentiating good corrections from bad ones, resulting in a low F0.5 score when used for system combination. In this paper, we propose GRECO, a new state-of-the-art quality estimation model that gives a better estimate of the quality of a corrected sentence, as indicated by having a higher correlation to the F0.5 score of a corrected sentence. It results in a combined GEC system with a higher F0.5 score. We also propose three methods for utilizing GEC quality estimation models for system combination with varying generality: model-agnostic, model-agnostic with voting bias, and model-dependent method. The combined GEC system outperforms the state of the art on the CoNLL-2014 test set and the BEA-2019 test set, achieving the highest F0.5 scores published to date.

09:00-10:30 (East Foyer / Virtual)
**#110 Using Interpretation Methods for Model Enhancement**
*Zhuo Chen, Chengyue Jiang and Kewei Tu*
In the age of neural natural language processing, there are plenty of works trying to derive interpretations of neural models. Intuitively, when gold rationales exist during training, one can additionally train the model to match its interpretation with the rationales. However, this intuitive idea has not been fully explored. In this paper, we propose a framework of utilizing interpretation methods and gold rationales to enhance models. Our framework is very general in the sense that it can incorporate various interpretation methods. Previously proposed gradient-based methods can be shown as an instance of our framework. We also propose two novel instances utilizing two other types of interpretation methods, erasure/replace-based and extractor-based methods, for model enhancement. We conduct comprehensive experiments on a variety of tasks. Experimental results show that our framework is effective especially in low-resource settings in enhancing models with various interpretation methods, and our two newly-proposed methods outperform gradient-based methods in most settings. Code is available at https://github.com/Chord-Chen-30/UIMER.

09:00-10:30 (East Foyer / Virtual)
**#111 Joint Geometrical and Statistical Domain Adaptation for Cross-domain Code Vulnerability Detection**
*Qianjin Du, Shiji Zhou, Xiaohui Kuang, Gang Zhao and Jidong Zhai*
In code vulnerability detection tasks, a detector trained on a label-rich source domain fails to provide accurate prediction on new or unseen target domains due to the lack of labeled training data on target domains. Previous studies mainly utilize domain adaptation to perform cross-domain vulnerability detection. But they ignore the negative effect of private semantic characteristics of the target domain for domain alignment, which easily causes the problem of negative transfer. In addition, these methods forcibly reduce the distribution discrepancy between domains and do not take into account the interference of irrelevant target instances for distributional domain alignment, which leads to the problem of excessive alignment. To address the above issues, we propose a novel cross-domain code vulnerability detection framework named MNCRI. Specifically, we introduce mutual nearest neighbor contrastive learning to align the source domain and target domain geometrically, which could align the common semantic characteristics of two domains and separate out the private semantic characteristics of each domain. Furthermore, we introduce an instance re-weighting scheme to alleviate the problem of excessive alignment. This scheme dynamically assign different weights to instances, reducing the contribution of irrelevant instances so as to achieve better domain alignment. Finally, extensive experiments demonstrate that MNCRI significantly outperforms state-of-the-art cross-domain code vulnerability detection methods by a large margin.

09:00-10:30 (East Foyer / Virtual)
**#112 Reducing Sequence Length by Predicting Edit Spans with Large Language Models**
*Masahiro Kaneko and Naoaki Okazaki*
Large Language Models (LLMs) have demonstrated remarkable performance in various tasks and gained significant attention. LLMs are also used for local sequence transduction tasks, including grammatical error correction (GEC) and formality style transfer, where most tokens in a source text are kept unchanged. However, the models that generate all target tokens in such tasks have a tendency to simply copy the input text as is, without making needed changes, because the difference between input and output texts is minimal in the training data. This is also inefficient because the computational cost grows quadratically with the target sequence length with Transformer. This paper proposes predicting edit spans for the source text for local sequence transduction tasks. Representing an edit span with a position of the source text and

corrected tokens, we can reduce the length of the target sequence and the computational cost for inference. We apply instruction tuning for LLMs on the supervision data of edit spans. Experiments show that the proposed method achieves comparable performance to the baseline in four tasks, paraphrasing, formality style transfer, GEC, and text simplification, despite reducing the length of the target text by as small as 21%. Furthermore, we report that the task-specific fine-tuning with the proposed method achieved state-of-the-art performance in the four tasks.

09:00-10:30 (East Foyer / Virtual)
**#113 Beware of Model Collapse! Fast and Stable Test-time Adaptation for Robust Question Answering**
*Yi Su, Yixin Ji, Juntao Li, Hai Ye and Min Zhang*
Although pre-trained language models (PLM) have achieved great success in question answering (QA), their robustness is still insufficient to support their practical applications, especially in the face of distribution shifts. Recently, test-time adaptation (TTA) has shown great potential for solving this problem, which adapts the model to fit the test samples at test time. However, TTA sometimes causes model collapse, making almost all the model outputs incorrect, which has raised concerns about its stability and reliability. In this paper, we delve into why TTA causes model collapse and find that the imbalanced label distribution inherent in QA is the reason for it. To address this problem, we propose Anti-Collapse Fast test-time adaptation (Anti-CF), which utilizes the source model's output to regularize the update of the adapted model during test time. We further design an efficient side block to reduce its inference time. Extensive experiments on various distribution shift scenarios and pre-trained language models (e.g., XLM-RoBERTa, BLOOM) demonstrate that our method can achieve comparable or better results than previous TTA methods at a speed close to vanilla forward propagation, which is $1.8\times$ to $4.4\times$ speedup compared to previous TTA methods.

09:00-10:30 (East Foyer / Virtual)
**#114 DisCo: Distilled Student Models Co-training for Semi-supervised Text Mining**
*Weifeng Jiang, Qianren Mao, Chenghua Lin, Jianxin Li, Ting Deng, Weiyi Yang and Zheng Wang*
Many text mining models are constructed by fine-tuning a large deep pre-trained language model (PLM) in downstream tasks. However, a significant challenge that arises nowadays is how to maintain performance when we use a lightweight model with limited labeled samples. We present DisCo, a semi-supervised learning (SSL) framework for fine-tuning a cohort of small student models generated from a large PLM using knowledge distillation. Our key insight is to share complementary knowledge among distilled student cohorts to promote their SSL effectiveness. DisCo employs a novel co-training technique to optimize a cohort of multiple small student models by promoting knowledge sharing among students under diversified views: model views produced by different distillation strategies and data views produced by various input augmentations. We evaluate DisCo on both semi-supervised text classification and extractive summarization tasks. Experimental results show that DisCo can produce student models that are $7.6\times$ smaller and $4.8\times$ faster in inference than the baseline PLMs while maintaining comparable performance. We also show that DisCo-generated student models outperform the similar-sized models elaborately tuned in distinct tasks.

09:00-10:30 (East Foyer / Virtual)
**#115 Learning to Rank Generation with Pairwise Partial Rewards**
*Youngwon Lee, Jinu Lee and Seung-won Hwang*
This paper studies the use of reinforcement learning for conditional text generation, which overcomes the limitation of the prevalent supervised maximum likelihood estimation approach. However, it still suffers from challenges including the large action space and the delayed reward, as the reward can be computed only after an entire sequence is generated. To address these challenges, we propose a method that provides partial rewards for intermediate actions taken on partial sequences. This enables the model to promptly prioritize actions that lead to the generation of more desirable sequences. Our method's key contribution lies in its focus on distinguishing relatively more desirable actions rather than striving to precisely estimate pointwise values for arbitrary partial sequences. Instead, our model learns to discern the relative desirability between pairs of actions, or rank actions in a pairwise manner, only when necessary and feasible. This is materialized in an efficient way by leveraging the prefix tree constructed from the sampled sequences. Experimental results on paraphrase generation and constrained machine translation tasks showcase the effectiveness of our method.

09:00-10:30 (East Foyer / Virtual)
**#116 Revisiting the Knowledge Injection Frameworks**
*Peng Fu, Yiming Zhang, Haobo Wang, Weikang Qiu and Junbo Zhao*
In recent years, large language models (LLMs), such as GPTs, have attained great impact worldwide. However, how to adapt these LLMs to better suit the vertical domain-specific tasks by utilizing external knowledge remains not completely solved. Indeed, there have emerged a few works on this line where most of them rely on an alignment heuristic that is built to inject the corresponding knowledge tuple into the associated text sample. However, despite the promise, we identify a pivotal problem in this work ubiquitously. Simply put, we find that injecting unaligned (i.e., random) knowledge tuple into the LLMs achieves comparable (and sometimes better) results than the aligned knowledge being injected. We therefore take a thorough investigation of this frustrating finding on a variety of related prior work and further provide a chain of potential interpretations for the phenomenon. Based on all that, we offer a simple remediated technique. Briefly, the core of this technique roots in an ideological emphasis on the pruning and purification of the external knowledge base to be injected into LLMs. At last, we show that by integrating this technique into most (if not all) knowledge injection frameworks and recent LLMs, it manages to overcome the aforementioned sanity problem and further pushes the boundary of the performance of the domain-adaptive LLMs.

09:00-10:30 (East Foyer / Virtual)
**#117 Non-autoregressive Text Editing with Copy-aware Latent Alignments**
*Yu Zhang, Yue Zhang, Leyang Cui and Guohong Fu*
Recent work has witnessed a paradigm shift from Seq2Seq to Seq2Edit in the field of text editing, with the aim of addressing the slow autoregressive inference problem posed by the former. Despite promising results, Seq2Edit approaches still face several challenges such as inflexibility in generation and difficulty in generalizing to other languages. In this work, we propose a novel non-autoregressive text editing method to circumvent the above issues, by modeling the edit process with latent CTC alignments. We make a crucial extension to CTC by introducing the copy operation into the edit space, thus enabling more efficient management of textual overlap in editing. We conduct extensive experiments on GEC and sentence fusion tasks, showing that our proposed method significantly outperforms existing Seq2Edit models and achieves similar or even better results than Seq2Seq with over $4\times$ speedup. Moreover, it demonstrates good generalizability on German and Russian. In-depth analyses reveal the strengths of our method in terms of the robustness under various scenarios and generating fluent and flexible outputs.

09:00-10:30 (East Foyer / Virtual)
**#118 Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales**
*Winston Wu, Lu Wang and Rada Mihalcea*
Folk tales are strong cultural and social influences in children's lives, and they are known to teach morals and values. However, existing studies on folk tales are largely limited to European tales. In our study, we compile a large corpus of over 1,900 tales originating from 27 diverse

cultures across six continents. Using a range of lexicons and correlation analyses, we examine how human values, morals, and gender biases are expressed in folk tales across cultures. We discover differences between cultures in prevalent values and morals, as well as cross-cultural trends in problematic gender biases. Furthermore, we find trends of reduced value expression when examining public-domain fiction stories, extrinsically validate our analyses against the multicultural Schwartz Survey of Cultural Values and the Global Gender Gap Report, and find traditional gender biases associated with values, morals, and agency. This large-scale cross-cultural study of folk tales paves the way towards future studies on how literature influences and reflects cultural norms.

09:00-10:30 (East Foyer / Virtual)
**#119 Unsupervised Grammatical Error Correction Rivaling Supervised Methods**
*Hannan Cao, Liping Yuan, Yuchen Zhang and Hwee Tou Ng*
State-of-the-art grammatical error correction (GEC) systems rely on parallel training data (ungrammatical sentences and their manually corrected counterparts), which are expensive to construct. In this paper, we employ the Break-It-Fix-It (BIFI) method to build an unsupervised GEC system. The BIFI framework generates parallel data from unlabeled text using a fixer to transform ungrammatical sentences into grammatical ones, and a critic to predict sentence grammaticality. We present an unsupervised approach to build the fixer and the critic, and an algorithm that allows them to iteratively improve each other. We evaluate our unsupervised GEC system on English and Chinese GEC. Empirical results show that our GEC system outperforms previous unsupervised GEC systems, and achieves performance comparable to supervised GEC systems without ensemble. Furthermore, when combined with labeled training data, our system achieves new state-of-the-art results on the CoNLL-2014 and NLPCC-2018 test sets.

09:00-10:30 (East Foyer / Virtual)
**#120 ATHENA: Mathematical Reasoning with Thought Expansion**
*JB. Kim, Hazel Kim, Joonghyuk Hahn and Yo-Sub Han*
Solving math word problems depends on how to articulate the problems, the lens through which models view human linguistic expressions. Real-world settings count on such a method even more due to the diverse practices of the same mathematical operations. Earlier works constrain available thinking processes by limited prediction strategies without considering their significance in acquiring mathematical knowledge. We introduce Attention-based THought Expansion Network Architecture (ATHENA) to tackle the challenges of real-world practices by mimicking human thought expansion mechanisms in the form of neural network propagation. A thought expansion recurrently generates the candidates carrying the thoughts of possible math expressions driven from the previous step and yields reasonable thoughts by selecting the valid pathways to the goal. Our experiments show that ATHENA achieves a new state-of-the-art stage toward the ideal model that is compelling in variant questions even when the informativeness in training examples is restricted.

09:00-10:30 (East Foyer / Virtual)
**#121 A Digital Language Coherence Marker for Monitoring Dementia**
*Dimitris Gkoumas, Adam Tsakalidis and Maria Liakata*
The use of spontaneous language to derive appropriate digital markers has become an emergent, promising and non-intrusive method to diagnose and monitor dementia. Here we propose methods to capture language coherence as a cost-effective, human-interpretable digital marker for monitoring cognitive changes in people with dementia. We introduce a novel task to learn the temporal logical consistency of utterances in short transcribed narratives and investigate a range of neural approaches. We compare such language coherence patterns between people with dementia and healthy controls and conduct a longitudinal evaluation against three clinical bio-markers to investigate the reliability of our proposed digital coherence marker. The coherence marker shows a significant difference between people with mild cognitive impairment, those with Alzheimer's Disease and healthy controls. Moreover our analysis shows high association between the coherence marker and the clinical bio-markers as well as generalisability potential to other related conditions.

09:00-10:30 (East Foyer / Virtual)
**#122 FAME: Flexible, Scalable Analogy Mappings Engine**
*Shahar Jacob, Chen Shani and Dafna Shahaf*
Analogy is one of the core capacities of human cognition; when faced with new situations, we often transfer prior experience from other domains. Most work on computational analogy relies heavily on complex, manually crafted input. In this work, we relax the input requirements, requiring only names of entities to be mapped. We automatically extract commonsense representations and use them to identify a mapping between the entities. Unlike previous works, our framework can handle partial analogies and suggest new entities to be added. Moreover, our method's output is easily interpretable, allowing for users to understand why a specific mapping was chosen. Experiments show that our model correctly maps 81.2% of classical 2x2 analogy problems (guess level=50%). On larger problems, it achieves 77.8% accuracy (mean guess level=13.1%). In another experiment, we show our algorithm outperforms human performance, and the automatic suggestions of new entities resemble those suggested by humans. We hope this work will advance computational analogy by paving the way to more flexible, realistic input requirements, with broader applicability.

09:00-10:30 (East Foyer / Virtual)
**#123 CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning**
*Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan and Chao Shen*
Machine-Generated Text (MGT) detection, a task that discriminates MGT from Human-Written Text (HWT), plays a crucial role in preventing misuse of text generative models, which excel in mimicking human writing style recently. Latest proposed detectors usually take coarse text sequences as input and fine-tune pretrained models with standard cross-entropy loss. However, these methods fail to consider the linguistic structure of texts. Moreover, they lack the ability to handle the low-resource problem which could often happen in practice considering the enormous amount of textual data online. In this paper, we present a coherence-based contrastive learning model named CoCo to detect the possible MGT under low-resource scenario. To exploit the linguistic feature, we encode coherence information in form of graph into text representation. To tackle the challenges of low data resource, we employ a contrastive learning framework and propose an improved contrastive loss for preventing performance degradation brought by simple samples. The experiment results on two public datasets and two self-constructed datasets prove our approach outperforms the state-of-art methods significantly. Also, we surprisingly find that MGTs originated from up-to-date language models could be easier to detect than these from previous models, in our experiments. And we propose some preliminary explanations for this counter-intuitive phenomena. All the codes and datasets are open-sourced.

09:00-10:30 (East Foyer / Virtual)
**#124 Relation-aware Ensemble Learning for Knowledge Graph Embedding**
*Ling Yue, Yongqi Zhang, Quanming Yao, Yong Li, Xian Wu, Ziheng Zhang, Zhenxi Lin and Yefeng Zheng*
Knowledge graph (KG) embedding is a fundamental task in natural language processing, and various methods have been proposed to explore semantic patterns in distinctive ways. In this paper, we propose to learn an ensemble by leveraging existing methods in a relation-aware manner. However, exploring these semantics using relation-aware ensemble leads to a much larger search space than general ensemble methods. To address this issue, we propose a divide-search-combine algorithm RelEns-DSC that searches the relation-wise ensemble weights independently. This algorithm has the same computation cost as general ensemble methods but with much better performance. Experimental

results on benchmark datasets demonstrate the effectiveness of the proposed method in efficiently searching relation-aware ensemble weights and achieving state-of-the-art embedding performance. The code is public at https://github.com/LARS-research/RelEns.

09:00-10:30 (East Foyer / Virtual)
**#125 Semantic Similarity Models for Depression Severity Estimation**
*Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar and Iryna Gurevych*
Depressive disorders constitute a severe public health issue worldwide. However, public health systems have limited capacity for case detection and diagnosis. In this regard, the widespread use of social media has opened up a way to access public information on a large scale. Computational methods can serve as support tools for rapid screening by exploiting this user-generated social media content. This paper presents an efficient semantic pipeline to study depression severity in individuals based on their social media writings. We select test user sentences for producing semantic rankings over an index of representative training sentences corresponding to depressive symptoms and severity levels. Then, we use the sentences from those results as evidence for predicting symptoms severity. For that, we explore different aggregation methods to answer one of four Beck Depression Inventory (BDI-II) options per symptom. We evaluate our methods on two Reddit-based benchmarks, achieving improvement over state of the art in terms of measuring depression level.

09:00-10:30 (East Foyer / Virtual)
**#126 Federated Meta-Learning for Emotion and Sentiment Aware Multi-modal Complaint Identification**
*Apoorva Singh, Siddarth Chandrasekar, Sriparna Saha and Tanmay Sen*
Automatic detection of consumers' complaints about items or services they buy can be critical for organizations and online merchants. Previous studies on complaint identification are limited to text. Images along with the reviews can provide cues to identify complaints better, thus emphasizing the importance of incorporating multi-modal inputs into the process. Generally, the customer's emotional state significantly impacts the complaint expression; thus, the effect of emotion and sentiment on complaint identification must also be investigated. Furthermore, different organizations are usually not allowed to share their privacy-sensitive records due to data security and privacy concerns. Due to these issues, traditional models find it hard to understand and identify complaint patterns, particularly in the financial and healthcare sectors. In this work, we created a new dataset - Multi-modal Complaint Dataset (MCD), a collection of reviews and images of the products posted on the website of the retail giant Amazon. We propose a federated meta-learning-based multi-modal multi-task framework for identifying complaints considering emotion recognition and sentiment analysis as two auxiliary tasks. Experimental results indicate that the proposed approach outperforms the baselines and the state-of-the-art approaches in centralized and federated meta-learning settings.

09:00-10:30 (East Foyer / Virtual)
**#127 Not all Fake News is Written: A Dataset and Analysis of Misleading Video Headlines**
*Yoo Yeon Sung, Jordan Lee Boyd-Graber and Naeemul Hassan*
Polarization and the marketplace for impressions have conspired to make navigating information online difficult for users, and while there has been a significant effort to detect false or misleading text, multimodal datasets have received considerably less attention. To complement existing resources, we present multimodal Video Misleading Headline (VMH), a dataset that consists of videos and whether annotators believe the headline is representative of the video's contents. After collecting and annotating this dataset, we analyze multimodal baselines for detecting misleading headlines. Our annotation process also focuses on why annotators view a video as misleading, allowing us to better understand the interplay of annotators' background and the content of the videos.

09:00-10:30 (East Foyer / Virtual)
**#128 Automated Fact-Checking in Dialogue: Are Specialized Models Needed?**
*Eric Chamoun, Marzieh Saeidi and Andreas Vlachos*
Prior research has shown that typical fact-checking models for stand-alone claims struggle with claims made in conversation. As a solution, fine-tuning these models on dialogue data has been proposed. However, creating separate models for each use case is impractical, and we show that fine-tuning models for dialogue results in poor performance on typical fact-checking. To overcome this challenge, we present techniques that allow us to use the same models for both dialogue and typical fact-checking. These mainly focus on retrieval adaptation and transforming conversational inputs so that they can be accurately processed by models trained on stand-alone claims. We demonstrate that a typical fact-checking model incorporating these techniques is competitive with state-of-the-art models for dialogue, while maintaining its performance on stand-alone claims.

09:00-10:30 (East Foyer / Virtual)
**#129 Exploring Distributional Shifts in Large Language Models for Code Analysis**
*Shushan Arakelyan, Rocktim Jyoti Das, Yi Mao and Xiang Ren*
We systematically study how three large language models with code capabilities - CodeT5, Codex, and ChatGPT - generalize to out-of-domain data. We consider two fundamental applications - code summarization, and code generation. We split data into domains following its natural boundaries - by an organization, by a project, and by a module within the software project. We establish that samples from each new domain present all the models with a significant challenge of distribution shift. We study how established methods adapt models to better generalize to new domains. Our experiments show that while multitask learning alone is a reasonable baseline, combining it with few-shot finetuning on examples retrieved from training data can achieve very strong performance. Moreover, this solution can outperform direct finetuning for very low-data scenarios. Finally, we consider variations of this approach to create a more broadly applicable method to adapt to multiple domains at once. We find that for code generation, a model adapted to multiple domains simultaneously performs on par with those adapted to a single domain.

09:00-10:30 (East Foyer / Virtual)
**#130 An End-to-End Contrastive Self-Supervised Learning Framework for Language Understanding**
*Pengtao Xie and Hongchao Fang*
Self-supervised learning (SSL) methods such as Word2vec, BERT, and GPT have shown great effectiveness in language understanding. Contrastive learning, as a recent SSL approach, has attracted increasing attention in NLP. Contrastive learning learns data representations by predicting whether two augmented data instances are generated from the same original data example. Previous contrastive learning methods perform data augmentation and contrastive learning separately. As a result, the augmented data may not be optimal for contrastive learning. To address this problem, we propose a four-level optimization framework which performs data augmentation and contrastive learning end-to-end, to enable the augmented data to be tailored to the contrastive learning task. This framework consists of four learning stages, including training machine translation models for sentence augmentation, pretraining a text encoder using contrastive learning, finetuning a text classification model, and updating weights of translation data by minimizing the validation loss of the classification model, which are performed in a unified way. Experiments on datasets in the GLUE benchmark and on datasets used in (Gururangan et al., 2020) demonstrate the effectiveness of our method.

09:00-10:30 (East Foyer / Virtual)
**#131 Reasoning over Public and Private Data in Retrieval-Based Systems**

*Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn and Christopher Ré*
Users and organizations are generating ever-increasing amounts of private data from a wide range of sources. Incorporating private context is important to personalize open-domain tasks such as question-answering, fact-checking, and personal assistants. State-of-the-art systems for these tasks explicitly retrieve information that is relevant to an input question from a background corpus before producing an answer. While today's retrieval systems assume relevant corpora are fully (e.g., publicly) accessible, users are often unable or unwilling to expose their private data to entities hosting public data. We define the Public-Private Autoregressive Information Retrieval (PAIR) problem involving retrieval over multiple privacy scopes. We introduce a foundational benchmark with which to study PAIR, as no existing benchmark includes data from a private distribution. Our dataset, ConcurrentQA, includes data from distinct public and private distributions and is the first textual QA benchmark requiring concurrent retrieval over multiple distributions. Finally, we show that existing retrieval approaches face significant performance degradations when applied to our proposed retrieval setting and investigate approaches with which these tradeoffs can be mitigated. We release the QA system and new benchmark.

09:00-10:30 (East Foyer / Virtual)
**#132 Erasure of Unaligned Attributes from Neural Representations**
*Yftah Ziser, Shay Cohen and Shun Shao*
We present the Assignment-Maximization Spectral Attribute removaL (AMSAL) algorithm, which removes information from neural representations when the information to be erased is implicit rather than directly being aligned to each input example. Our algorithm works by alternating between two steps. In one, it finds an assignment of the input representations to the information to be erased, and in the other, it creates projections of both the input representations and the information to be erased into a joint latent space. We test our algorithm on an extensive array of datasets, including a Twitter dataset with multiple guarded attributes, the BiasBios dataset and the BiasBench benchmark. The last benchmark includes four datasets with various types of protected attributes. Our results demonstrate that bias can often be removed in our setup. We also discuss the limitations of our approach when there is a strong entanglement between the main task and the information to be erased. Our code is available at https://github.com/jasonshaoshun/AMSAL.

09:00-10:30 (East Foyer / Virtual)
**#133 Introduction to Mathematical Language Processing: Informal Proofs, Word Problems, and Supporting Tasks**
*Jordan Meadows and Andre Freitas*
Automating discovery in mathematics and science will require sophisticated methods of information extraction and abstract reasoning, including models that can convincingly process relationships between mathematical elements and natural language, to produce problem solutions of real-world value. We analyze mathematical language processing methods across five strategic sub-areas (identifier-definition extraction, formula retrieval, natural language premise selection, math word problem solving, and informal theorem proving) in recent years, highlighting prevailing methodologies, existing limitations, overarching trends, and promising avenues for future research.

09:00-10:30 (East Foyer / Virtual)
**#134 Improving Multitask Retrieval by Promoting Task Specialization**
*Wenzheng Zhang, Chenyan Xiong, Karl Stratos and Arnold Overwijk*
In multitask retrieval, a single retriever is trained to retrieve relevant contexts for multiple tasks. Despite its practical appeal, naive multitask retrieval lags behind task-specific retrieval in which a separate retriever is trained for each task. We show that it is possible to train a multitask retriever that outperforms task-specific retrievers by promoting task specialization. The main ingredients are: (1) a better choice of pretrained model one that is explicitly optimized for multitasking along with compatible prompting, and (2) a novel adaptive learning method that encourages each parameter to specialize in a particular task. The resulting multitask retriever is highly performant on the KILT benchmark. Upon analysis, we find that the model indeed learns parameters that are more task-specialized compared to naive multitasking without prompting or adaptive learning.

09:00-10:30 (East Foyer / Virtual)
**#135 Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems**
*Lemei Zhang, Peng Liu and Jon Atle Gulla*
The emergency of Pre-trained Language Models (PLMs) has achieved tremendous success in the field of Natural Language Processing (NLP) by learning universal representations on large corpora in a self-supervised manner. The pre-trained models and the learned representations can be beneficial to a series of downstream NLP tasks. This training paradigm has recently been adapted to the recommendation domain and is considered a promising approach by both academia and industry. In this paper, we systematically investigate how to extract and transfer knowledge from pre-trained models learned by different PLM-related training paradigms to improve recommendation performance from various perspectives, such as generality, sparsity, efficiency and effectiveness. Specifically, we propose a comprehensive taxonomy to divide existing PLM-based recommender systems w.r.t. their training strategies and objectives. Then, we analyze and summarize the connection between PLM-based training paradigms and different input data types for recommender systems. Finally, we elaborate on open issues and future research directions in this vibrant field.

09:00-10:30 (East Foyer / Virtual)
**#136 General then Personal: Decoupling and Pre-training for Personalized Headline Generation**
*Hong-Han Shuai, Yun-Zhu Song, Yi-Syuan Chen and Lu Wang*
Personalized Headline Generation aims to generate unique headlines tailored to users' browsing history. In this task, understanding user preferences from click history and incorporating them into headline generation pose challenges. Existing approaches typically rely on predefined styles as control codes, but personal style lacks explicit definition or enumeration, making it difficult to leverage traditional techniques. To tackle these challenges, we propose General Then Personal (GTP), a novel framework comprising user modeling, headline generation, and customization. We train the framework using tailored designs that emphasize two central ideas: (a) task decoupling and (b) model pre-training. With the decoupling mechanism separating the task into generation and customization, two mechanisms, i.e., information self-boosting and mask user modeling, are further introduced to facilitate the training and text control. Additionally, we introduce a new evaluation metric to address existing limitations. Extensive experiments conducted on the PENS dataset, considering both zero-shot and few-shot scenarios, demonstrate that GTP outperforms state-of-the-art methods. Furthermore, ablation studies and analysis emphasize the significance of decoupling and pre-training. Finally, the human evaluation validates the effectiveness of our approaches.

## Findings 6

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)

**Investigating Multilingual Coreference Resolution by Universal Annotations**
*Haixia Chai and Michael Strube*
Multilingual coreference resolution (MCR) has been a long-standing and challenging task. With the newly proposed multilingual coreference dataset, CorefUD (Nedoluzhko et al., 2022), we conduct an investigation into the task by using its harmonized universal morphosyntactic and coreference annotations. First, we study coreference by examining the ground truth data at different linguistic levels, namely mention, entity and document levels, and across different genres, to gain insights into the characteristics of coreference across multiple languages. Second, we perform an error analysis of the most challenging cases that the SotA system fails to resolve in the CRAC 2022 shared task using the universal annotations. Last, based on this analysis, we extract features from universal morphosyntactic annotations and integrate these features into a baseline system to assess their potential benefits for the MCR task. Our results show that our best configuration of features improves the baseline by 0.9% F1 score.

09:00-10:30 (East Foyer / Virtual)
**Do "English" Named Entity Recognizers Work Well on Global Englishes?**
*Alexander Shan, John Bauer, Riley Carlson and Christopher D Manning*
The vast majority of the popular English named entity recognition (NER) datasets contain American or British English data, despite the existence of many global varieties of English. As such, it is unclear whether they generalize for analyzing use of English globally. To test this, we build a newswire dataset, the Worldwide English NER Dataset, to analyze NER model performance on low-resource English variants from around the world. We test widely used NER toolkits and transformer models, including models using the pre-trained contextual models RoBERTa and ELECTRA, on three datasets: a commonly used British English newswire dataset, CoNLL 2003, a more American focused dataset OntoNotes, and our global dataset. All models trained on the CoNLL or OntoNotes datasets experienced significant performance drops—over 10 F1 in some cases—when tested on the Worldwide English dataset. Upon examination of region-specific errors, we observe the greatest performance drops for Oceania and Africa, while Asia and the Middle East had comparatively strong performance. Lastly, we find that a combined model trained on the Worldwide dataset and either CoNLL or OntoNotes lost only 1-2 F1 on both test sets.

09:00-10:30 (East Foyer / Virtual)
**A Reference-free Segmentation Quality Index (SegReFree)**
*Evan Lucas, Dylan Kangas and Timothy Havens*
Topic segmentation, in the context of natural language processing, is the process of finding boundaries in a sequence of sentences that separate groups of adjacent sentences at shifts in semantic meaning. Currently, assessing the quality of a segmentation is done by comparing segmentation boundaries selected by a human or algorithm to those selected by a known good reference. This means that it is not possible to quantify the quality of a segmentation without a human annotator, which can be costly and time consuming. This work seeks to improve assessment of segmentation by proposing a reference-free segmentation quality index (SegReFree). The metric takes advantage of the fact that segmentation at a sentence level generally seeks to identify segment boundaries at semantic boundaries within the text. The proposed metric uses a modified cluster validity metric with semantic embeddings of the sentences to determine the quality of the segmentation. Multiple segmentation data sets are used to compare our proposed metric with existing reference-based segmentation metrics by progressively degrading the reference segmentation while computing all possible metrics; through this process, a strong correlation with existing segmentation metrics is shown. A Python library implementing the metric is released under the GNU General Public License and the repository is available at https://github.com/evan-person/reference_free_segmentation_metric.

09:00-10:30 (East Foyer / Virtual)
**Non-Autoregressive Sentence Ordering**
*Yi Bin, Wenhao Shi, Bin Ji, Jipeng Zhang, Yujuan Ding and Yang Yang*
Existing sentence ordering approaches generally employ encoder-decoder frameworks with the pointer net to recover the coherence by recurrently predicting each sentence step-by-step. Such an autoregressive manner only leverages unilateral dependencies during decoding and cannot fully explore the semantic dependency between sentences for ordering. To overcome these limitations, in this paper, we propose a novel Non-Autoregressive Ordering Network, dubbed *NAON*, which explores bilateral dependencies between sentences and predicts the sentence for each position in parallel. We claim that the non-autoregressive manner is not just applicable but also particularly suitable to the sentence ordering task because of two peculiar characteristics of the task: 1) each generation target is in deterministic length, and 2) the sentences and positions should match exclusively. Furthermore, to address the repetition issue of the naive non-autoregressive Transformer, we introduce an exclusive loss to constrain the exclusiveness between positions and sentences. To verify the effectiveness of the proposed model, we conduct extensive experiments on several common-used datasets and the experimental results show that our method outperforms all the autoregressive approaches and yields competitive performance compared with the state-of-the-arts. The codes are available at: https://github.com/steven640pixel/nonautoregressive-sentence-ordering.

09:00-10:30 (East Foyer / Virtual)
**"Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters**
*Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang and Nanyun Peng*
Large Language Models (LLMs) have recently emerged as an effective tool to assist individuals in writing various types of content, including professional documents such as recommendation letters. Though bringing convenience, this application also introduces unprecedented fairness concerns. Model-generated reference letters might be directly used by users in professional scenarios. If underlying biases exist in these model-constructed letters, using them without scrutinization could lead to direct societal harms, such as sabotaging application success rates for female applicants. In light of this pressing issue, it is imminent and necessary to comprehensively study fairness issues and associated harms in this real-world use case. In this paper, we critically examine gender biases in LLM-generated reference letters. Drawing inspiration from social science findings, we design evaluation methods to manifest biases through 2 dimensions: (1) biases in language style and (2) biases in lexical content. We further investigate the extent of bias propagation by analyzing the hallucination bias of models, a term that we define to be bias exacerbation in model-hallucinated contents. Through benchmarking evaluation on 2 popular LLMs- ChatGPT and Alpaca, we reveal significant gender biases in LLM-generated recommendation letters. Our findings not only warn against using LLMs for this application without scrutinization, but also illuminate the importance of thoroughly studying hidden biases and harms in LLM-generated professional documents.

09:00-10:30 (East Foyer / Virtual)
**Perceptual Structure in the absence of grounding: the impact of abstractedness and subjectivity in color language for LLMs**
*Pablo Loyola, Edison Marrese-Taylor and Andres Hoyos-Idrobo*
The need for grounding in language understanding is an active research topic. Previous work has suggested that color perception and color language appear as a suitable test bed to empirically study the problem, given its cognitive significance and showing that there is considerable alignment between a defined color space and the feature space defined by a language model. To further study this issue, we collect a large scale source of colors and their descriptions, containing almost a 1 million examples , and perform an empirical analysis to compare two kinds of alignments: (i) inter-space, by learning a mapping between embedding space and color space, and (ii) intra-space, by means of prompting comparatives between color descriptions. Our results show that while color space alignment holds for monolexemic, highly pragmatic color

descriptions, this alignment drops considerably in the presence of examples that exhibit elements of real linguistic usage such as subjectivity and abstractedness, suggesting that grounding may be required in such cases.

09:00-10:30 (East Foyer / Virtual)
**TalkUp: Paving the Way for Understanding Empowering Language**
*Lucille Njoo, Chan Young Park, Octavia Stappart, Marvin Thielk, Yi Chu and Yulia Tsvetkov*
Empowering language is important in many real-world contexts, from education to workplace dynamics to healthcare. Though language technologies are growing more prevalent in these contexts, empowerment has seldom been studied in NLP, and moreover, it is inherently challenging to operationalize because of its implicit nature. This work builds from linguistic and social psychology literature to explore what characterizes empowering language. We then crowdsource a novel dataset of Reddit posts labeled for empowerment, reasons why these posts are empowering to readers, and the social relationships between posters and readers. Our preliminary analyses show that this dataset, which we call TalkUp, can be used to train language models that capture empowering and disempowering language. More broadly, TalkUp provides an avenue to explore implication, presuppositions, and how social context influences the meaning of language.

09:00-10:30 (East Foyer / Virtual)
**Towards Zero-shot Relation Extraction in Web Mining: A Multimodal Approach with Relative XML Path**
*Zilong Wang and Jingbo Shang*
The rapid growth of web pages and the increasing complexity of their structure poses a challenge for web mining models. Web mining models are required to understand semi-structured web pages, particularly when little is known about the subject or template of a new page. Current methods migrate language models to web mining by embedding the XML source code into the transformer or encoding the rendered layout with graph neural networks. However, these approaches do not take into account the relationships between text nodes within and across pages. In this paper, we propose a new approach, ReXMiner, for zero-shot relation extraction in web mining. ReXMiner encodes the shortest relative paths in the Document Object Model (DOM) tree of the web page which is a more accurate and efficient signal for key-value pair extraction within a web page. It also incorporates the popularity of each text node by counting the occurrence of the same text node across different web pages. We use contrastive learning to address the issue of sparsity in relation extraction. Extensive experiments on public benchmarks show that our method, ReXMiner, outperforms the state-of-the-art baselines in the task of zero-shot relation extraction in web mining.

09:00-10:30 (East Foyer / Virtual)
**The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who**
*Michael Sejr Schlichtkrull, Nedjma Ousidhoum and Andreas Vlachos*
Automated fact-checking is often presented as an epistemic tool that fact-checkers, social media consumers, and other stakeholders can use to fight misinformation. Nevertheless, few papers thoroughly discuss *how*. We document this by analysing 100 highly-cited papers, and annotating epistemic elements related to intended use, i.e., means, ends, and stakeholders. We find that narratives leaving out some of these aspects are common, that many papers propose inconsistent means and ends, and that the feasibility of suggested strategies rarely has empirical backing. We argue that this vagueness actively hinders the technology from reaching its goals, as it encourages overclaiming, limits criticism, and prevents stakeholder feedback. Accordingly, we provide several recommendations for thinking and writing about the use of fact-checking artefacts.

09:00-10:30 (East Foyer / Virtual)
**Asking Clarification Questions to Handle Ambiguity in Open-Domain QA**
*Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee and Kyomin Jung*
Ambiguous questions persist in open-domain question answering, because formulating a precise question with a unique answer is often challenging. Previous works have tackled this issue by asking disambiguated questions for all possible interpretations of the ambiguous question. Instead, we propose to ask a clarification question, where the user's response will help identify the interpretation that best aligns with the user's intention. We first present CAmbigNQ, a dataset consisting of 5,653 ambiguous questions, each with relevant passages, possible answers, and a clarification question. The clarification questions were efficiently created by generating them using InstructGPT and manually revising them as necessary. We then define a pipeline of three tasks—(1) ambiguity detection, (2) clarification question generation, and (3) clarification-based QA. In the process, we adopt or design appropriate evaluation metrics to facilitate sound research. Lastly, we achieve F1 of 61.3, 25.1, and 40.5 on the three tasks, demonstrating the need for further improvements while providing competitive baselines for future work.

09:00-10:30 (East Foyer / Virtual)
**The language of prompting: What linguistic properties make a prompt successful?**
*Alina Leidinger, Robert van Rooij and Ekaterina Shutova*
The latest generation of LLMs can be prompted to achieve impressive zero-shot or few-shot performance in many NLP tasks. However, since performance is highly sensitive to the choice of prompts, considerable effort has been devoted to crowd-sourcing prompts or designing methods for prompt optimisation. Yet, we still lack a systematic understanding of how linguistic properties of prompts correlate with the task performance. In this work, we investigate how LLMs of different sizes, pre-trained and instruction-tuned, perform on prompts that are semantically equivalent, but vary in linguistic structure. We investigate both grammatical properties such as mood, tense, aspect and modality, as well as lexico-semantic variation through the use of synonyms. Our findings contradict the common assumption that LLMs achieve optimal performance on prompts which reflect language use in pretraining or instruction-tuning data. Prompts transfer poorly between datasets or models, and performance cannot generally be explained by perplexity, word frequency, word sense ambiguity or prompt length. Based on our results, we put forward a proposal for a more robust and comprehensive evaluation standard for prompting research.

09:00-10:30 (East Foyer / Virtual)
**PMIndiaSum: Multilingual and Cross-lingual Headline Summarization for Languages in India**
*Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay B Cohen, Manish Shrivastava and Barry Haddow*
This paper introduces PMIndiaSum, a multilingual and massively parallel summarization corpus focused on languages in India. Our corpus provides a training and testing ground for four language families, 14 languages, and the largest to date with 196 language pairs. We detail our construction workflow including data acquisition, processing, and quality assurance. Furthermore, we publish benchmarks for monolingual, cross-lingual, and multilingual summarization by fine-tuning, prompting, as well as translate-and-summarize. Experimental results confirm the crucial role of our data in aiding summarization between Indian languages. Our dataset is publicly available and can be freely modified and re-distributed.

09:00-10:30 (East Foyer / Virtual)
**KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion**
*Yanbin Wei, Qiushi Huang, Yu Zhang and James Kwok*
Knowledge Graph Completion (KGC) is crucial for addressing knowledge graph incompleteness and supporting downstream applications. Many models have been proposed for KGC and they can be categorized into two main classes, including triple-based and test-based approaches. Triple-based methods struggle with long-tail entities due to limited structural information and imbalanced distributions of entities.

Text-based methods alleviate this issue but require costly training for language models and specific finetuning for knowledge graphs, which limits their efficiency. To alleviate the limitations in the two approaches, in this paper, we propose KICGPT, a framework that integrates a large language model (LLM) and a triple-based KGC retriever, to alleviate the long-tail problem without incurring additional training overhead. In the proposed KICGPT model, we propose an in-context learning strategy called Knowledge Prompt, which encodes structural knowledge into demonstrations to guide LLM. Empirical results on benchmark datasets demonstrate the effectiveness of the proposed KICGPT model with lighter training overhead and no finetuning.

09:00-10:30 (East Foyer / Virtual)
**Adapting Pretrained Text-to-Text Models for Long Text Sequences**
*Wenhan Xiong, Anchit Gupta, Shubham Toshniwal, Yashar Mehdad and Scott Yih*
We present an empirical study of adapting an existing pretrained text-to-text model for long-sequence inputs. Through a comprehensive study along three axes of the pretraining pipeline – model architecture, optimization objective, and pretraining corpus, we propose an effective recipe to build long-context models from existing short-context models. Specifically, we replace the full attention in transformers with *pooling-augmented blockwise attention*, and pretrain the model with a masked-span prediction task with spans of varying lengths. In terms of the pretraining corpus, we find that using randomly concatenated short-documents from a large open-domain corpus results in better performance than using existing long document corpora, which are typically limited in their domain coverage. With these findings, we build a long-context model that achieves competitive performance on long-text QA tasks and establishes the new state of the art on *five* long-text summarization datasets, often outperforming previous methods with larger model sizes.

09:00-10:30 (East Foyer / Virtual)
**Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models**
*Hongzhan Lin, Ziyang Luo, Jing Ma and Long Chen*
The age of social media is rife with memes. Understanding and detecting harmful memes pose a significant challenge due to their implicit meaning that is not explicitly conveyed through the surface text and image. However, existing harmful meme detection approaches only recognize superficial harm-indicative signals in an end-to-end classification manner but ignore in-depth cognition of the meme text and image. In this paper, we attempt to detect harmful memes based on advanced reasoning over the interplay of multimodal information in memes. Inspired by the success of Large Language Models (LLMs) on complex reasoning, we first conduct abductive reasoning with LLMs. Then we propose a novel generative framework to learn reasonable thoughts from LLMs for better multimodal fusion and lightweight fine-tuning, which consists of two training stages: 1) Distill multimodal reasoning knowledge from LLMs; and 2) Fine-tune the generative framework to infer harmfulness. Extensive experiments conducted on three meme datasets demonstrate that our proposed approach achieves superior performance than state-of-the-art methods on the harmful meme detection task.

09:00-10:30 (East Foyer / Virtual)
**Data Augmentation for Code Translation with Comparable Corpora and Multiple References**
*Yiqing Xie, Atharva Naik, Daniel Fried and Carolyn Rose*
One major challenge of translating code between programming languages is that parallel training data is often limited. To overcome this challenge, we present two data augmentation techniques, one that builds comparable corpora (i.e., code pairs with similar functionality), and another that augments existing parallel data with multiple reference translations. Specifically, we build and analyze multiple types of comparable corpora, including programs generated from natural language documentation using a code generation model. Furthermore, to reduce overfitting to a single reference translation, we automatically generate additional translation references for available parallel data and filter the translations by unit tests, which increases variation in target translations. Experiments show that our data augmentation techniques significantly improve CodeT5 for translation between Java, Python, and C++ by an average of 7.5% Computational Accuracy (CA@1), which verifies the correctness of translations by execution. The code is available at https://github.com/Veronicium/CMTrans.

09:00-10:30 (East Foyer / Virtual)
**Improving the Robustness of Summarization Models by Detecting and Removing Input Noise**
*Kundan Krishna, Yao Zhao, Jie Ren, Balaji Lakshminarayanan, Jiaming Luo, Mohammad Saleh and Peter J Liu*
The evaluation of abstractive summarization models typically uses test data that is identically distributed as training data. In real-world practice, documents to be summarized may contain input noise caused by text extraction artifacts or data pipeline bugs. The robustness of model performance under distribution shift caused by such noise is relatively under studied. We present a large empirical study quantifying the sometimes severe loss in performance – up to 12 ROUGE-1 points – from different types of input noise for a range of datasets and model sizes. We then propose a light-weight method for detecting and removing such noise in the input during model inference without requiring any extra training, auxiliary models, or even prior knowledge of the type of noise. Our proposed approach effectively mitigates the loss in performance, recovering a large fraction of the performance drop, sometimes as large as 11 ROUGE-1 points.

09:00-10:30 (East Foyer / Virtual)
**Enhancing Abstractiveness of Summarization Models through Calibrated Distillation**
*Hwanjun Song, Igor Shalyminov, Hang Su, Siffi Singh, Kaisheng Yao and Saab Mansour*
In this paper, we propose a novel approach named DisCal to enhance the level of abstractiveness (measured by n-gram overlap) without sacrificing the informativeness (measured by ROUGE) of generated summaries. DisCal exposes diverse pseudo summaries with two supervision to the student model. Firstly, the best pseudo summary is identified in terms of abstractiveness and informativeness and used for sequence-level distillation. Secondly, their ranks are used to ensure the student model to assign higher prediction scores to summaries with higher ranks. Our experiments show that DisCal outperforms prior methods in abstractive summarization distillation, producing highly abstractive and informative summaries.

09:00-10:30 (East Foyer / Virtual)
**Evaluating and Enhancing the Robustness of Code Pre-trained Models through Structure-Aware Adversarial Samples Generation**
*Nuo Chen, Qiushi Sun, Jianing Wang, Ming Gao, Xiaoli Li and Xiang Li*
Code pre-trained models (CodePTMs) have significantly advanced the field of neural code intelligence. Despite their capabilities, these models are susceptible to adversarial attacks that subtly modify the model inputs, resulting in incorrect outputs or predictions. Previous methods of robustness evaluation for CodePTMs primarily stem from a textual perspective, without explicitly taking into account the structure of the code. Furthermore, prior studies fail to encompass a broad enough spectrum of tasks and models. In this paper, we propose a set of novel robustness evaluation methods based on the intrinsic structure of the code. Specifically, we first launch adversarial attacks on crucial identifier tokens and sub-tree structures to explore the impact of imperceptible perturbation. Then, we perform global restructuring of the code using different traversal methods for abstract syntax trees, aiming to explore the model's sensitivity to input samples with equivalent information. Moreover, for each scenario, we employ adversarial training methods to explore the possibility of restoring the performance of perturbed models. For both code understanding and generation, our proposed method has demonstrated its effectiveness across a wide range of models and tasks, thereby allowing us to make one step forward in our understanding of the inner mechanisms of CodePTMs.

09:00-10:30 (East Foyer / Virtual)
**Aligning Language Models to User Opinions**
*EunJeong Hwang, Bodhisattwa Prasad Majumder and Niket Tandon*
An important aspect of developing LLMs that interact with humans is to align models' behavior to their users. It is possible to prompt an LLM into behaving as a certain persona, especially a user group or ideological persona the model captured during its pertaining stage. But, how to best align an LLM with a specific user and not a demographic or ideological group remains an open question. Mining public opinion surveys (by PEW research), we find that the opinions of a user and their demographics and ideologies are not mutual predictors. We use this insight to align LLMs by modeling relevant past user opinions in addition to user demographics and ideology, achieving up to 7 points accuracy gains in predicting public opinions from survey questions across a broad set of topics. Our work opens up the research avenues to bring user opinions as an important ingredient in aligning language models.

09:00-10:30 (East Foyer / Virtual)
**GenKIE: Robust Generative Multimodal Document Key Information Extraction**
*Panfeng Cao, Ye Wang, Qiang Zhang and Zaiqiao Meng*
Key information extraction (KIE) from scanned documents has gained increasing attention because of its applications in various domains. Although promising results have been achieved by some recent KIE approaches, they are usually built based on discriminative models, which lack the ability to handle optical character recognition (OCR) errors and require laborious token-level labeling. In this paper, we propose a novel generative end-to-end model, named GenKIE, to address the KIE task. GenKIE is a sequence-to-sequence multimodal generative model that utilizes multimodal encoders to embed visual, layout and textual features and a decoder to generate the desired output. Well-designed prompts are leveraged to incorporate the label semantics as the weakly supervised signals and entice the generation of the key information. One notable advantage of the generative model is that it enables automatic correction of OCR errors. Besides, token-level granular annotation is not required. Extensive experiments on multiple public real-world datasets show that GenKIE effectively generalizes over different types of documents and achieves state-of-the-art results. Our experiments also validate the model's robustness against OCR errors, making GenKIE highly applicable in real-world scenarios.

09:00-10:30 (East Foyer / Virtual)
**Ultra-Fine Entity Typing with Prior Knowledge about Labels: A Simple Clustering Based Strategy**
*Na Li, Zied Bouraoui and Steven Schockaert*
Ultra-fine entity typing (UFET) is the task of inferring the semantic types from a large set of fine-grained candidates that apply to a given entity mention. This task is especially challenging because we only have a small number of training examples for many types, even with distant supervision strategies. State-of-the-art models, therefore, have to rely on prior knowledge about the type labels in some way. In this paper, we show that the performance of existing methods can be improved using a simple technique: we use pre-trained label embeddings to cluster the labels into semantic domains and then treat these domains as additional types. We show that this strategy consistently leads to improved results as long as high-quality label embeddings are used. Furthermore, we use the label clusters as part of a simple post-processing technique, which results in further performance gains. Both strategies treat the UFET model as a black box and can thus straightforwardly be used to improve a wide range of existing models.

09:00-10:30 (East Foyer / Virtual)
**COMET-M: Reasoning about Multiple Events in Complex Sentences**
*Sahithya Ravi, Raymond T. Ng and Vered Shwartz*
Understanding the speaker's intended meaning often involves drawing commonsense inferences to reason about what is not stated explicitly. In multi-event sentences, it requires understanding the relationships between events based on contextual knowledge. We propose COMET-M (Multi-Event), an event-centric commonsense model capable of generating commonsense inferences for a target event within a complex sentence. COMET-M builds upon COMET (Bosselut et al., 2019), which excels at generating event-centric inferences for simple sentences, but struggles with the complexity of multi-event sentences prevalent in natural text. To overcome this limitation, we curate a Multi-Event Inference (MEI) dataset of 35K human-written inferences. We train COMET-M on the human-written inferences and also create baselines using automatically labeled examples. Experimental results demonstrate the significant performance improvement of COMET-M over COMET in generating multi-event inferences. Moreover, COMET-M successfully produces distinct inferences for each target event, taking the complete context into consideration. COMET-M holds promise for downstream tasks involving natural text such as coreference resolution, dialogue, and story understanding.

09:00-10:30 (East Foyer / Virtual)
**More than Votes? Voting and Language based Partisanship in the US Supreme Court**
*Biaoyan Fang, Trevor Cohn, Timothy Baldwin and Lea Frermann*
Understanding the prevalence and dynamics of justice partisanship and ideology in the US Supreme Court is critical in studying jurisdiction. Most research quantifies partisanship based on voting behavior, and oral arguments in the courtroom — the last essential procedure before the final case outcome — have not been well studied for this purpose. To address this gap, we present a framework for analyzing the language of justices in the courtroom for partisan signals, and study how partisanship in speech aligns with voting patterns. Our results show that the affiliated party of justices can be predicted reliably from their oral contributions. We further show a strong correlation between language partisanship and voting ideology.

09:00-10:30 (East Foyer / Virtual)
**CASE: Commonsense-Augmented Score with an Expanded Answer Space**
*Wenkai Chen, Sahithya Ravi and Vered Shwartz*
LLMs have demonstrated impressive zero-shot performance on NLP tasks thanks to the knowledge they acquired in their training. In multiple-choice QA tasks, the LM probabilities are used as an imperfect measure of the plausibility of each answer choice. One of the major limitations of the basic score is that it treats all words as equally important. We propose CASE, a Commonsense-Augmented Score with an Expanded Answer Space. CASE addresses this limitation by assigning importance weights for individual words based on their semantic relations to other words in the input. The dynamic weighting approach outperforms basic LM scores, not only because it reduces noise from unimportant words, but also because it informs the model of implicit commonsense knowledge that may be useful for answering the question. We then also follow prior work in expanding the answer space by generating lexically-divergent answers that are conceptually-similar to the choices. When combined with answer space expansion, our method outperforms strong baselines on 5 commonsense benchmarks. We further show these two approaches are complementary and may be especially beneficial when using smaller LMs.

09:00-10:30 (East Foyer / Virtual)
**Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data**
*Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu and Elena Simperl*
Numerical data plays a crucial role in various real-world domains like finance, economics, and science. Thus, understanding and reasoning with numbers are essential in these fields. Recent benchmarks have assessed the numerical reasoning abilities of language models, revealing

their limitations in limited and specific numerical aspects. In this paper, we propose a complete hierarchical taxonomy for numerical reasoning skills, encompassing over ten reasoning types across four levels: representation, number sense, manipulation, and complex reasoning. We conduct a comprehensive evaluation of state-of-the-art models on all reasoning types. To identify challenging reasoning types for different model types, we develop a diverse and extensive set of numerical probes and measure performance shifts. By employing a semi-automated approach, we focus on the tabular Natural Language Inference (TNLI) task as a case study. While no single model excels in all reasoning types, FlanT5 (few-/zero-shot) and GPT3.5 (few-shot) demonstrate strong overall numerical reasoning skills compared to other models in our probes.

09:00-10:30 (East Foyer / Virtual)
**Smart "Chef": Verifying the Effect of Role-based Paraphrasing for Aspect Term Extraction**
*Jiaxiang Chen, Yu Hong, Qingting Xu and Jianmin Yao*
We tackle Aspect Term Extraction (ATE), a task of automatically extracting aspect terms from sentences. The current Pretrained Language Model (PLM) based extractors have achieved significant improvements. They primarily benefit from context-aware encoding. However, a considerable number of sentences in ATE corpora contain uninformative or low-quality contexts. Such sentences frequently act as "trouble-makers" during test. In this study, we explore the context-oriented quality improvement method. Specifically, we propose to automatically rewrite the sentences from the perspectives of virtual experts with different roles, such as a "chef" in the restaurant domain. On this basis, we perform ATE over the paraphrased sentences during test, using the well-trained extractors without any change. In the experiments, we leverage ChatGPT to determine virtual experts in the considered domains, and induce ChatGPT to generate paraphrases conditioned on the roles of virtual experts. We experiment on the benchmark SemEval datasets, including Laptop-domain L14 and Restaurant-domain R14-16. The experimental results show that our approach effectively recalls the inconspicuous aspect terms like "al di la", although it reduces the precision. In addition, it is proven that our approach can be substantially improved by redundancy elimination and multi-role voting. More importantly, our approach can be used to expand the predictions obtained on the original sentences. This yields state-of-the-art performance (i.e., F1-scores of 86.2%, 89.3%, 77.7%, 82.7% on L14 and R14-16) without retraining or fine-tuning the baseline extractors.

09:00-10:30 (East Foyer / Virtual)
**Multilingual Generation and Answering of Questions from Texts and Knowledge Graphs**
*Kelvin Han and Claire Gardent*
The ability to bridge Question Generation (QG) and Question Answering (QA) across structured and unstructured modalities has the potential for aiding different NLP applications. One key application is in QA-based methods that have recently been shown to be useful for automatically evaluating Natural Language (NL) texts generated from Knowledge Graphs (KG). While methods have been proposed for QG-QA across these modalities, these efforts have been in English only; in this work, we bring multilinguality (Brazilian Portuguese and Russian) to multimodal (KG and NL) QG-QA. Using synthetic data generation and machine translation to produce QG-QA data that is aligned between graph and text, we are able to train multimodal, multi-task models that can perform multimodal QG and QA in Portuguese and Russian. We show that our approach outperforms a baseline which is derived from previous work on English and adapted to handle these two languages.

09:00-10:30 (East Foyer / Virtual)
**Improving generalization in large langue model by learning prefix subspaces**
*Louis Falissard, Vincent Guigue and Laure Soulier*
This article focuses on large language models (LLMs) fine-tuning in the scarce data regime (also known as "few-shot learning setting"). We propose a method to increase the generalization capabilities of LLMs based on neural network subspaces. This optimization method, recently introduced in computer vision, aims to improve model generalization by identifying wider local optima through the joint optimization of an entire simplex of models in parameter space. Although this property would be highly beneficial in the context of training large language models in the "few-shot learning" setting, its adaptation to massive, pretrained transformers poses some challenges. First, their considerable number of parameters make it difficult to train several model jointly, and second, their deterministic parameter initialisation schemes make them unfit to the subspace method as originaly proposed. We show in this paper that its application to "Parameter Efficient Fine-Tuning" (PEFT) methods, however, is relatively natural, and we propose to apply it to prefix-tuning, by learning entire simplexes of continous prefixes. We test our method on a variant of the GLUE benchmark adapted to the few-shot learning setting, and show that both our contributions (learning prefix simplexes, and non-deterministic validation metric inference) jointly lead to a gain in average performances compared to state of the art methods.

09:00-10:30 (East Foyer / Virtual)
**InvGC: Robust Cross-Modal Retrieval by Inverse Graph Convolution**
*Xiangru Jian and Yimu Wang*
Over recent decades, significant advancements in cross-modal retrieval is mainly driven by breakthroughs in visual and linguistic modeling. However, a recent study shows that multi-modal data representations tend to cluster within a limited convex cone (as representation degeneration problem), which hinders retrieval performance due to the inseparability of these representations. In our study, we first empirically validate the presence of the representation degeneration problem across multiple cross-modal benchmarks and methods. Next, to address it, we introduce a novel method, called InvGC, a post-processing technique inspired by graph convolution and average pooling. Specifically, InvGC defines the graph topology within the datasets and then applies graph convolution in a subtractive manner. This method effectively separates representations by increasing the distances between data points. To improve the efficiency and effectiveness of InvGC, we propose an advanced graph topology, LocalAdj, which only aims to increase the distances between each data point and its nearest neighbors. To understand why InvGC works, we present a detailed theoretical analysis, proving that the lower bound of recall will be improved after deploying InvGC. Extensive empirical results show that InvGC and InvGC w/LocalAdj significantly mitigate the representation degeneration problem, thereby enhancing retrieval performance.

09:00-10:30 (East Foyer / Virtual)
**HeQ: a Large and Diverse Hebrew Reading Comprehension Benchmark**
*Amir David Nissan Cohen, Hilla Merhav-Fine, Yoav Goldberg and Reut Tsarfaty*
Current benchmarks for Hebrew Natural Language Processing (NLP) focus mainly on morpho-syntactic tasks, neglecting the semantic dimension of language understanding. To bridge this gap, we set out to deliver a Hebrew Machine Reading Comprehension (MRC) dataset, where MRC is to be realized as extractive Question Answering. The morphologically-rich nature of Hebrew poses a challenge to this endeavor: the indeterminacy and non-transparency of span boundaries in morphologically complex forms lead to annotation inconsistencies, disagreements, and flaws of standard evaluation metrics. To remedy this, we devise a novel set of guidelines, a controlled crowdsourcing protocol, and revised evaluation metrics, that are suitable for the morphologically rich nature of the language. Our resulting benchmark, HeQ (Hebrew QA), features 30,147 diverse question-answer pairs derived from both Hebrew Wikipedia articles and Israeli tech news. Our empirical investigation reveals that standard evaluation metrics such as F1 Scores and Exact Match (EM) are not appropriate for Hebrew (and other MRLs), and we propose a relevant enhancement. In addition, our experiments show low correlation between models' performance on morpho-syntactic tasks and on MRC, which suggests that models that are designed for the former might underperform on semantic-heavy tasks. The development and exploration of HeQ illustrate some of the challenges MRLs pose in natural language understanding (NLU), fostering progression towards more and better NLU models for Hebrew and other MRLs.

09:00-10:30 (East Foyer / Virtual)
**Unleashing the Multilingual Encoder Potential: Boosting Zero-Shot Performance via Probability Calibration**
*Ercong Nie, Helmut Schmid and Hinrich Schuetze*
Pretrained multilingual encoder models can directly perform zero-shot multilingual tasks or linguistic probing by reformulating the input examples into cloze-style prompts. This is accomplished by predicting the probabilities of the label words at the masked token position, without requiring any updates to the model parameters. However, the performance of this method is limited by the model's bias toward predicting label words which frequently occurred during the pretraining. These words typically receive high probabilities. To address this issue, we combine the models with calibration techniques which modify the probabilities of label words predicted by the models. We first validate the effectiveness of a proposed simple calibration method together with other existing techniques on monolingual encoders in both zero- and few-shot scenarios. We subsequently employ these calibration techniques on multilingual encoders, resulting in substantial performance improvements across a wide range of tasks.

09:00-10:30 (East Foyer / Virtual)
**Focus on the Core: Efficient Attention via Pruned Token Compression for Document Classification**
*Jungmin Yun, Mihyeon Kim and Youngbin Kim*
Transformer-based models have achieved dominant performance in numerous NLP tasks. Despite their remarkable successes, pre-trained transformers such as BERT suffer from a computationally expensive self-attention mechanism that interacts with all tokens, including the ones unfavorable to classification performance. To overcome these challenges, we propose integrating two strategies: token pruning and token combining. Token pruning eliminates less important tokens in the attention mechanism's key and value as they pass through the layers. Additionally, we adopt fuzzy logic to handle uncertainty and alleviate potential mispruning risks arising from an imbalanced distribution of each token's importance. Token combining, on the other hand, condenses input sequences into smaller sizes in order to further compress the model. By integrating these two approaches, we not only improve the model's performance but also reduce its computational demands. Experiments with various datasets demonstrate superior performance compared to baseline models, especially with the best improvement over the existing BERT model, achieving +5%p in accuracy and +5.6%p in F1 score. Additionally, memory cost is reduced to 0.61x, and a speedup of 1.64x is achieved.

09:00-10:30 (East Foyer / Virtual)
**PersonaLM: Language Model Personalization via Domain-distributed Span Aggregated K-Nearest N-gram Retrieval Augmentation**
*Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Keren, Zeeshan Ahmed, Dinesh Manocha and Xuedong Zhang*
We introduce PersonaLM - Domain-distributed Span-Aggregated K-nearest N-gram retrieval augmentation to improve language modeling for Automatic Speech Recognition (ASR) personalization. PersonaLM leverages contextually similar n-gram word frequencies for recognizing rare word patterns associated with unseen domains. It aggregates the next-word probability distribution based on the relative importance of different domains to the input query. To achieve this, we propose a Span Aggregated Group-Contrastive Neural (SCAN) retriever that learns to rank external domains/users by utilizing a group-wise contrastive span loss that pulls together span representations belonging to the same group while pushing away spans from unrelated groups in the semantic space. We propose ASAP benchmark for ASR LM personalization that consists of three user-specific speech-to-text tasks for meetings, TED talks, and financial earnings calls. Extensive experiments show that PersonaLM significantly outperforms strong baselines with a 10-16% improvement in perplexity and a 5-8% reduction in Word Error Rates on popular Wikitext-103, UserLibri, and our ASAP dataset. We further demonstrate the usefulness of the SCAN retriever for improving user-personalized text generation and classification by retrieving relevant context for zero-shot prompting and few-shot fine-tuning of LLMs by 7-12% on the LAMP benchmark.

09:00-10:30 (East Foyer / Virtual)
**Conditional Natural Language Inference**
*Youngwoo Kim, Razieh Rahimi and James Allan*
To properly explain sentence pairs that provide contradictory (different) information for different conditions, we introduce the task of conditional natural language inference (Cond-NLI) and focus on automatically extracting contradictory aspects and their conditions from a sentence pair. Cond-NLI can help to provide a full spectrum of information, such as when there are multiple answers to a question each addressing a specific condition, or reviews with different opinions for different conditions. We show that widely-used feature-attribution explanation models are not suitable for finding conditions, especially when sentences are long and are written independently. We propose a simple yet effective model for the original NLI task that can successfully extract conditions while not requiring token-level annotations. Our model enhances the interpretability of the NLI task while maintaining comparable accuracy. To evaluate models for the Cond-NLI, we build and release a token-level annotated dataset BioClaim which contains potentially contradictory claims from the biomedical domain. Our experiments show that our proposed model outperforms the full cross-encoder and other baselines in extracting conditions. It also performs on-par with GPT-3 which has an order of magnitude more parameters and trained on a huge amount of data.

09:00-10:30 (East Foyer / Virtual)
**Estimating Large Language Model Capabilities without Labeled Test Data**
*Harvey Yiyun Fu, Qinyuan Ye, Albert Xu, Xiang Ren and Robin Jia*
Large Language Models (LLMs) have exhibited an impressive ability to perform in-context learning (ICL) from only a few examples, but the success of ICL varies widely from task to task. Thus, it is important to quickly determine whether ICL is applicable to a new task, but directly evaluating ICL accuracy can be expensive in situations where test data is expensive to annotate—the exact situations where ICL is most appealing. In this paper, we propose the task of ICL accuracy estimation, in which we predict the accuracy of an LLM when doing in-context learning on a new task given only unlabeled test data for that task. To perform ICL accuracy estimation, we propose a method that trains a meta-model using LLM confidence scores as features. We compare our method to several strong accuracy estimation baselines on a new benchmark that covers 4 LLMs and 3 task collections. The meta-model improves over all baselines across 7 out of 12 settings and achieves the same estimation performance as directly evaluating on 40 collected labeled test examples per task. At the same time, no existing approach provides an accurate and reliable ICL accuracy estimation in every setting, highlighting the need for better ways to measure the uncertainty of LLM predictions.

09:00-10:30 (East Foyer / Virtual)
**Social Commonsense-Guided Search Query Generation for Open-Domain Knowledge-Powered Conversations**
*Revanth Gangi Reddy, Hao Bai, Wentao Yao, Sharath Chandra Etagi Suresh, Heng Ji and ChengXiang Zhai*
Open-domain dialog involves generating search queries that help obtain relevant knowledge for holding informative conversations. However, it can be challenging to determine what information to retrieve when the user is passive and does not express a clear need or request. To tackle this issue, we present a novel approach that focuses on generating internet search queries that are guided by social commonsense. Specifically, we leverage a commonsense dialog system to establish connections related to the conversation topic, which subsequently guides our query generation. Our proposed framework addresses passive user interactions by integrating topic tracking, commonsense response generation and instruction-driven query generation. Through extensive evaluations, we show that our approach overcomes limitations of existing query gen-

eration techniques that rely solely on explicit dialog information, and produces search queries that are more relevant, specific, and compelling, ultimately resulting in more engaging responses.

09:00-10:30 (East Foyer / Virtual)
**The Less the Merrier? Investigating Language Representation in Multilingual Models**
*Hellina Hailu Nigatu, Atnafu Lambebo Tonja and Jugal Kalita*
Multilingual Language Models offer a way to incorporate multiple languages in one model and utilize cross-language transfer learning to improve performance for different Natural Language Processing (NLP) tasks. Despite progress in multilingual models, not all languages are supported as well, particularly in low-resource settings. In this work, we investigate the linguistic representation of different languages in multilingual models. We start by asking the question which languages are supported in popular multilingual models and which languages are left behind. Then, for included languages, we look at models' learned representations based on language family and dialect and try to understand how models' learned representations for (1) seen and (2) unseen languages vary across different language groups. In addition, we test and analyze performance on downstream tasks such as text generation and Named Entity Recognition. We observe from our experiments that community-centered models—models that focus on languages of a given family or geographical location and are built by communities who speak them—perform better at distinguishing between languages in the same family for low-resource languages. Our paper contributes to the literature in understanding multilingual models and their shortcomings and offers insights on potential ways to improve them.

09:00-10:30 (East Foyer / Virtual)
**Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting**
*Fanghua Ye, Meng Fang, Shenghui Li and Emine Yilmaz*
Query rewriting plays a vital role in enhancing conversational search by transforming context-dependent user queries into standalone forms. Existing approaches primarily leverage human-rewritten queries as labels to train query rewriting models. However, human rewrites may lack sufficient information for optimal retrieval performance. To overcome this limitation, we propose utilizing large language models (LLMs) as query rewriters, enabling the generation of informative query rewrites through well-designed instructions. We define four essential properties for well-formed rewrites and incorporate all of them into the instruction. In addition, we introduce the role of rewrite editors for LLMs when initial query rewrites are available, forming a "rewrite-then-edit" process. Furthermore, we propose distilling the rewriting capabilities of LLMs into smaller models to reduce rewriting latency. Our experimental evaluation on the QReCC dataset demonstrates that informative query rewrites can yield substantially improved retrieval performance compared to human rewrites, especially with sparse retrievers.

09:00-10:30 (East Foyer / Virtual)
**MetaReVision: Meta-Learning with Retrieval for Visually Grounded Compositional Concept Acquisition**
*Guangyue Xu, Parisa Kordjamshidi and Joyce Chai*
Humans have the ability to learn novel compositional concepts by recalling primitive concepts acquired from past experience and generalizing these primitive concepts to novel compositions. Inspired by the above human's compositional learning procedure, in this paper, we propose MetaReVision, a retrievalenhanced meta-learning model to solve the visually grounded compositional concept learning problem. The proposed MetaReVision consists of a retrieval module and a meta-learning module which are designed to incorporate retrieved primitive concepts as supporting set to meta-train visual-language models for grounded compositional concept recognition. Through meta-learning from episodes constructed by the retriever, MetaReVision learns a generic compositional representation that can be fast updated to recognize novel composi tional concepts. We create CompCOCO and CompFlickr to benchmark the grounded compositional concept learning. Our experimental results show MetaReVision outperforms other competitive baselines and the retrieval module does plays an important role in this compositional learning process.

09:00-10:30 (East Foyer / Virtual)
**Coverage-based Example Selection for In-Context Learning**
*Shivanshu Gupta, Matt Gardner and Sameer Singh*
In-context learning (ICL), the ability of large language models to perform novel tasks by conditioning on a prompt with a few task examples, requires these examples to be informative about the test instance. The standard approach of independently ranking and selecting the most similar examples selects redundant examples while omitting important information. In this work, we show that BERTScore-Recall (BSR) selects better examples that demonstrate more of the salient aspects, e.g. reasoning patterns, of the test input. We further extend BSR and many standard metrics to easily optimizable set-level metrics, giving still better coverage of those salient aspects. On 15 datasets spanning 6 tasks and with 7 diverse LLMs, we show that (1) BSR is the superior metric for in-context example selection across the board, and (2) for compositional tasks, set selection using Set-BSR outperforms independent ranking by up to 17 points on average and, despite being training-free, surpasses methods that leverage task or LLM-specific training.

09:00-10:30 (East Foyer / Virtual)
**Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?**
*Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alex de la Iglesia, Megan Su, Xiao Lin Zheng and Joyce Chai*
Despite tremendous advances in AI, it remains a significant challenge to develop interactive task guidance systems that can offer situated, personalized guidance and assist humans in various tasks. These systems need to have a sophisticated understanding of the user as well as the environment, and make timely accurate decisions on when and what to say. To address this issue, we created a new multimodal benchmark dataset, Watch, Talk and Guide (WTaG) based on natural interaction between a human user and a human instructor. We further proposed two tasks: User and Environment Understanding, and Instructor Decision Making. We leveraged several foundation models to study to what extent these models can be quickly adapted to perceptually enabled task guidance. Our quantitative, qualitative, and human evaluation results show that these models can demonstrate fair performances in some cases with no task-specific training, but a fast and reliable adaptation remains a significant challenge. Our benchmark and baselines will provide a stepping stone for future work on situated task guidance.

09:00-10:30 (East Foyer / Virtual)
**IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning**
*Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy L. Bearman and Babak Damavandi*
We present IMU2CLIP, a novel pre-training approach to align Inertial Measurement Unit (IMU) motion sensor recordings with text and video, by projecting them into the joint representation space of Contrastive Language-Image Pre-training (CLIP). The proposed approach allows IMU2CLIP to translate human motions (as measured by IMU sensors) into their corresponding textual descriptions and videos – while preserving the transitivity across these modalities. We introduce several new IMU-based Wearable AI applications such as motion-based media search, or an LM-based multimodal reasoning with motion sensor data – all using text as the grounding platform. In addition, we show that IMU2CLIP significantly improves downstream performances when fine-tuned for each application, demonstrating its universal usage as a new pre-trained resource. Our code and models will be released publicly.

09:00-10:30 (East Foyer / Virtual)
**Test-time Augmentation for Factual Probing**

*Go Kamoda, Benjamin Heinzerling, Keisuke Sakaguchi and Kentaro Inui*
Factual probing is a method that uses prompts to test if a language model "knows" certain world knowledge facts. A problem in factual probing is that small changes to the prompt can lead to large changes in model output. Previous work aimed to alleviate this problem by optimizing prompts via text mining or fine-tuning. However, such approaches are relation-specific and do not generalize to unseen relation types. Here, we propose to use test-time augmentation (TTA) as a relation-agnostic method for reducing sensitivity to prompt variations by automatically augmenting and ensembling prompts at test time. Experiments show improved model calibration, i.e., with TTA, model confidence better reflects prediction accuracy. Improvements in prediction accuracy are observed for some models, but for other models, TTA leads to degradation. Error analysis identifies the difficulty of producing high-quality prompt variations as the main challenge for TTA.

09:00-10:30 (East Foyer / Virtual)
**NEWTON: Are Large Language Models Capable of Physical Reasoning?**
*Yi Ru Wang, Jiafei Duan, Dieter Fox and Siddhartha Srinivasa*
Large Language Models (LLMs), through their contextualized representations, have been empirically proven to encapsulate syntactic, semantic, word sense, and common-sense knowledge. However, there has been limited exploration of their physical reasoning abilities, specifically concerning the crucial attributes for comprehending everyday objects. To address this gap, we introduce NEWTON, a repository and benchmark for evaluating the physics reasoning skills of LLMs. Further, to enable domain-specific adaptation of this benchmark, we present a pipeline to enable researchers to generate a variant of this benchmark that has been customized to the objects and attributes relevant for their application. The NEWTON repository comprises a collection of 2800 object-attribute pairs, providing the foundation for generating infinite-scale assessment templates. The NEWTON benchmark consists of 160K QA questions, curated using the NEWTON repository to investigate the physical reasoning capabilities of several mainstream language models across foundational, explicit, and implicit reasoning tasks. Through extensive empirical analysis, our results highlight the capabilities of LLMs for physical reasoning. We find that LLMs like GPT-4 demonstrate strong reasoning capabilities in scenario-based tasks but exhibit less consistency in object-attribute reasoning compared to humans (50% vs. 84%). Furthermore, the NEWTON platform demonstrates its potential for evaluating and enhancing language models, paving the way for their integration into physically grounded settings, such as robotic manipulation. Project site: https://newtonreasoning.github.io

09:00-10:30 (East Foyer / Virtual)
**SuperTweetEval: A Challenging, Unified and Heterogeneous Benchmark for Social Media NLP Research**
*Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei and Jose Camacho-Collados*
Despite its relevance, the maturity of NLP for social media pales in comparison with general-purpose models, metrics and benchmarks. This fragmented landscape makes it hard for the community to know, for instance, given a task, which is the best performing model and how it compares with others. To alleviate this issue, we introduce a unified benchmark for NLP evaluation in social media, SuperTweetEval, which includes a heterogeneous set of tasks and datasets combined, adapted and constructed from scratch. We benchmarked the performance of a wide range of models on SuperTweetEval and our results suggest that, despite the recent advances in language modelling, social media remains challenging.

09:00-10:30 (East Foyer / Virtual)
**In What Languages are Generative Language Models the Most Formal? Analyzing Formality Distribution across Languages**
*Asım Ersoy, Gerson Vizcarra, Tahsin Mayeesha and Benjamin Muller*
Multilingual generative language models (LMs) are increasingly fluent in a large variety of languages. Trained on the concatenation of corpora in multiple languages, they enable powerful transfer from high-resource languages to low-resource ones. However, it is still unknown what cultural biases are induced in the predictions of these models. In this work, we focus on one language property highly influenced by culture: formality. We analyze the formality distributions of XGLM and BLOOM's predictions, two popular generative multilingual language models, in 5 languages. We classify 1,200 generations per language as formal, informal, or incohesive and measure the impact of the prompt formality on the predictions. Overall, we observe a diversity of behaviors across the models and languages. For instance, XGLM generates informal text in Arabic and Bengali when conditioned with informal prompts, much more than BLOOM. In addition, even though both models are highly biased toward the formal style when prompted neutrally, we find that the models generate a significant amount of informal predictions even when prompted with formal text. We release with this work 6,000 annotated samples, paving the way for future work on the formality of generative multilingual LMs.

09:00-10:30 (East Foyer / Virtual)
**Segmented Recurrent Transformer: An Efficient Sequence-to-Sequence Model**
*Yinghan Long, Sayeed Shafayet Chowdhury and Kaushik Roy*
Transformers have shown dominant performance across a range of domains including language and vision. However, their computational cost grows quadratically with the sequence length, making their usage prohibitive for resource-constrained applications. To counter this, our approach is to divide the whole sequence into segments and apply attention to the individual segments. We propose a segmented recurrent transformer (SRformer) that combines segmented (local) attention with recurrent attention. The loss caused by reducing the attention window length is compensated by aggregating information across segments with recurrent attention. SRformer leverages Recurrent Accumulate-and-Fire (RAF) neurons' inherent memory to update the cumulative product of keys and values. The segmented attention and lightweight RAF neurons ensure the efficiency of the proposed transformer. Such an approach leads to models with sequential processing capability at a lower computation/memory cost. We apply the proposed method to T5 and BART transformers. The modified models are tested on summarization datasets including CNN-dailymail, XSUM, ArXiv, and MediaSUM. Notably, using segmented inputs of varied sizes, the proposed model achieves 6-22% higher ROUGE1 scores than a segmented transformer and outperforms other recurrent transformer approaches. Furthermore, compared to full attention, the proposed model reduces the computational complexity of cross attention by around 40%.

09:00-10:30 (East Foyer / Virtual)
**Visually Grounded Continual Language Learning with Selective Specialization**
*Kyra Ahrens, Lennart Bengtson, Jae Hee Lee and Stefan Wermter*
A desirable trait of an artificial agent acting in the visual world is to continually learn a sequence of language-informed tasks while striking a balance between sufficiently specializing in each task and building a generalized knowledge for transfer. Selective specialization, i.e., a careful selection of model components to specialize in each task, is a strategy to provide control over this trade-off. However, the design of selection strategies requires insights on the role of each model component in learning rather specialized or generalizable representations, which poses a gap in current research. Thus, our aim with this work is to provide an extensive analysis of selection strategies for visually grounded continual language learning. Due to the lack of suitable benchmarks for this purpose, we introduce two novel diagnostic datasets that provide enough control and flexibility for a thorough model analysis. We assess various heuristics for module specialization strategies as well as quantifiable measures for two different types of model architectures. Finally, we design conceptually simple approaches based on our analysis that outperform common continual learning baselines. Our results demonstrate the need for further efforts towards better aligning continual learning algorithms with the learning behaviors of individual model parts.

09:00-10:30 (East Foyer / Virtual)
**MUX-PLMs: Data Multiplexing for High-throughput Language Models**
*Vishvak Murahari, Ameet Deshpande, Carlos E Jimenez, Izhak Shafran, Mingqiu Wang, Yuan Cao and Karthik R Narasimhan*
The widespread adoption of large language models such as ChatGPT and Bard has led to unprecedented demand for these technologies. The burgeoning cost of inference for ever-increasing model sizes coupled with hardware shortages has limited affordable access and poses a pressing need for efficiency approaches geared towards high throughput and performance. Multi-input multi-output (MIMO) algorithms such as data multiplexing, offer a promising solution with a many-fold increase in throughput by performing inference for multiple inputs at the cost of a single input. Yet these approaches are not currently performant enough to be deployed in modern systems. We change that by developing MUX-PLMs, a class of high throughput pre-trained language models (PLMs) trained with data multiplexing, that can be fine-tuned for any downstream task to yield high-throughput high-performance. Our novel multiplexing and demultiplexing modules proficiently entangle and disentangle inputs, and enable high-performance high throughput MUX-PLMs that are competitive with vanilla PLMs while achieving 2x/5x inference speedup with only a 1-4 % drop on a broad suite of tasks.

09:00-10:30 (East Foyer / Virtual)
**PaRaDe: Passage Ranking using Demonstrations with LLMs**
*Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler and Kai Hui*
Recent studies show that large language models (LLMs) can be instructed to effectively perform zero-shot passage re-ranking, in which the results of a first stage retrieval method, such as BM25, are rated and reordered to improve relevance. In this work, we improve LLM-based re-ranking by algorithmically selecting few-shot demonstrations to include in the prompt. Our analysis investigates the conditions where demonstrations are most helpful, and shows that adding even one demonstration is significantly beneficial. We propose a novel demonstration selection strategy based on difficulty rather than the commonly used semantic similarity. Furthermore, we find that demonstrations helpful for ranking are also effective at question generation. We hope our work will spur more principled research into question generation and passage ranking.

09:00-10:30 (East Foyer / Virtual)
**Interpreting Indirect Answers to Yes-No Questions in Multiple Languages**
*Zijie Wang, Md Mosharaf Hossain, Shivam Mathur, Terry Cruz Melo, Kadir Bulut Ozler, Keun Hee Park, Jacob Quintero, MohammadHossein Rezaei, Shreya Nupur Shakya, Md Nayem Uddin and Eduardo Blanco*
Yes-no questions expect a yes or no for an answer, but people often skip polar keywords. Instead, they answer with long explanations that must be interpreted. In this paper, we focus on this challenging problem and release new benchmarks in eight languages. We present a distant supervision approach to collect training data, and demonstrate that direct answers (i.e., with polar keywords) are useful to train models to interpret indirect answers (i.e., without polar keywords). We show that monolingual fine-tuning is beneficial if training data can be obtained via distant supervision for the language of interest (5 languages). Additionally, we show that cross-lingual fine-tuning is always beneficial (8 languages).

09:00-10:30 (East Foyer / Virtual)
**CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles**
*Shashank Sonkar, Naiming Liu, Debshila Basu Mallick and Richard Baraniuk*
We present a design framework called Conversational Learning with Analytical Step-by-Step Strategies (CLASS) for building advanced Intelligent Tutoring Systems (ITS) powered by high-performance Large Language Models (LLMs). The CLASS framework empowers ITS with two key capabilities. First, through a carefully curated scaffolding dataset, CLASS equips ITS with essential problem-solving strategies, enabling it to provide tutor-like, step-by-step guidance to students. Second, by using a dynamic conversational dataset, CLASS assists ITS in facilitating natural language interactions, fostering engaging student-tutor conversations. The CLASS framework also provides valuable insights into ITS's internal decision-making process which allows seamless integration of user feedback, thus enabling continuous refinement and improvement. We also present a proof-of-concept ITS, referred to as SPOCK, which is trained using the CLASS framework with a focus on introductory college level biology content. A carefully constructed protocol was developed for SPOCK's preliminary evaluation, examining aspects such as the factual accuracy and relevance of its responses. Experts in the field of biology offered favorable remarks, particularly highlighting SPOCK's capability to break down questions into manageable subproblems and provide encouraging responses to students.

09:00-10:30 (East Foyer / Virtual)
**Handshape-Aware Sign Language Recognition: Extended Datasets and Exploration of Handshape-Inclusive Methods**
*Xuan Zhang and Kevin Duh*
The majority of existing work on sign language recognition encodes signed videos without explicitly acknowledging the phonological attributes of signs. Given that handshape is a vital parameter in sign languages, we explore the potential of handshape-aware sign language recognition. We augment the PHOENIX14T dataset with gloss-level handshape labels, resulting in the new PHOENIX14T-HS dataset. Two unique methods are proposed for handshape-inclusive sign language recognition: a single-encoder network and a dual-encoder network, complemented by a training strategy that simultaneously optimizes both the CTC loss and frame-level cross-entropy loss. The proposed methodology consistently outperforms the baseline performance. The dataset and code can be accessed at: www.anonymous.com.

09:00-10:30 (East Foyer / Virtual)
**How Reliable Are AI-Generated-Text Detectors? An Assessment Framework Using Evasive Soft Prompts**
*Tharindu Sandaruwan Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland and Huan Liu*
In recent years, there has been a rapid proliferation of AI-generated text, primarily driven by the release of powerful pre-trained language models (PLMs). To address the issue of misuse associated with AI-generated text, various high-performing detectors have been developed, including the OpenAI detector and the Stanford DetectGPT. In our study, we ask how reliable these detectors are. We answer the question by designing a novel approach that can prompt any PLM to generate text that evades these high-performing detectors. The proposed approach suggests a universal evasive prompt, a novel type of soft prompt, which guides PLMs in producing "human-like" text that can mislead the detectors. The novel universal evasive prompt is achieved in two steps: First, we create an evasive soft prompt tailored to a specific PLM through prompt tuning; and then, we leverage the transferability of soft prompts to transfer the learned evasive soft prompt from one PLM to another. Employing multiple PLMs in various writing tasks, we conduct extensive experiments to evaluate the efficacy of the evasive soft prompts in their evasion of state-of-the-art detectors.

09:00-10:30 (East Foyer / Virtual)
**Intersectional Stereotypes in Large Language Models: Dataset and Analysis**
*Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang and Soroush Vosoughi*
Despite many stereotypes targeting intersectional demographic groups, prior studies on stereotypes within Large Language Models (LLMs) primarily focus on broader, individual categories. This research bridges this gap by introducing a novel dataset of intersectional stereotypes, curated with the assistance of the ChatGPT model and manually validated. Moreover, this paper offers a comprehensive analysis of intersec-

tional stereotype propagation in three contemporary LLMs by leveraging this dataset. The findings underscore the urgency of focusing on intersectional biases in ongoing efforts to reduce stereotype prevalence in LLMs.

09:00-10:30 (East Foyer / Virtual)
**GPT-4 as an Effective Zero-Shot Evaluator for Scientific Figure Captions**
*Ting-Yao Hsu, Chieh-Yang Huang, Ryan A. Rossi, Sungchul Kim, C. Lee Giles and Ting-Hao Kenneth Huang*
There is growing interest in systems that generate captions for scientific figures. However, assessing these systems' output poses a significant challenge. Human evaluation requires academic expertise and is costly, while automatic evaluation depends on often low-quality author-written captions. This paper investigates using large language models (LLMs) as a cost-effective, reference-free method for evaluating figure captions. We first constructed SCICAP-EVAL, a human evaluation dataset that contains human judgments for 3,600 scientific figure captions, both original and machine-made, for 600 arXiv figures. We then prompted LLMs like GPT-4 and GPT-3 to score (1-6) each caption based on its potential to aid reader understanding, given relevant context such as figure-mentioning paragraphs. Results show that GPT-4, used as a zero-shot evaluator, outperformed all other models and even surpassed assessments made by computer science undergraduates, achieving a Kendall correlation score of 0.401 with Ph.D. students' rankings.

09:00-10:30 (East Foyer / Virtual)
**Open Domain Multi-document Summarization: A Comprehensive Study of Model Brittleness under Retrieval**
*John Michael Giorgi, Luca Soldaini, Bo Wang, Gary D. Bader, Kyle Lo, Lucy Lu Wang and Arman Cohan*
Multi-document summarization (MDS) assumes a set of topic-related documents are provided as input. In practice, this document set is not always available; it would need to be retrieved given an information need, i.e. a question or topic statement, a setting we dub "open-domain" MDS. We study this more challenging setting by formalizing the task and bootstrapping it using existing datasets, retrievers and summarizers. Via extensive automatic and human evaluation, we determine: (1) state-of-the-art summarizers suffer large reductions in performance when applied to open-domain MDS, (2) additional training in the open-domain setting can reduce this sensitivity to imperfect retrieval, and (3) summarizers are insensitive to the retrieval of duplicate documents and the order of retrieved documents, but highly sensitive to other errors, like the retrieval of irrelevant documents. Based on our results, we provide practical guidelines to enable future work on open-domain MDS, e.g. how to choose the number of retrieved documents to summarize. Our results suggest that new retrieval and summarization methods and annotated resources for training and evaluation are necessary for further progress in the open-domain setting.

09:00-10:30 (East Foyer / Virtual)
**Video-Text Retrieval by Supervised Sparse Multi-Grained Learning**
*Yimu Wang and Peng Shi*
While recent progress in video-text retrieval has been advanced by the exploration of better representation learning, in this paper, we present a novel multi-grained sparse learning framework, S3MA, to learn an aligned sparse space shared between the video and the text for video-text retrieval. The shared sparse space is initialized with a finite number of sparse concepts, each of which refers to a number of words. With the text data at hand, we learn and update the shared sparse space in a supervised manner using the proposed similarity and alignment losses. Moreover, to enable multi-grained alignment, we incorporate frame representations for better modeling the video modality and calculating fine-grained and coarse-grained similarities. Benefiting from the learned shared sparse space and multi-grained similarities, extensive experiments on several video-text retrieval benchmarks demonstrate the superiority of S3MA over existing methods.

09:00-10:30 (East Foyer / Virtual)
**"You Are An Expert Linguistic Annotator": Limits of LLMs as Analyzers of Abstract Meaning Representation**
*Allyson Ettinger, Jena D. Hwang, Valentina Pyatkin, Chandra Bhagavatula and Yejin Choi*
Large language models (LLMs) demonstrate an amazing proficiency and fluency in the *use* of language. Does that mean that they have also acquired insightful linguistic knowledge *about* the language, to an extent that they can serve as an "expert linguistic annotator"? In this paper, we examine the successes and limitations of the GPT-3, ChatGPT, and GPT-4 models, focusing on the Abstract Meaning Representation (AMR) parsing formalism (Banarescu et al., 2013), which provides rich graphical representations of sentence meaning structure while abstracting away from surface forms. We compare models' analysis of this semantic structure across two settings: 1) direct production of AMR parses based on zero- and few-shot examples, and 2) indirect partial reconstruction of AMR via metalinguistic natural language queries (e.g., "Identify the primary event of this sentence, and the predicate corresponding to that event."). Across these settings, we find that models can reliably reproduce the basic format of AMR, as well as some core event, argument, and modifier structure—however, model outputs are prone to frequent and major errors, and holistic analysis of parse acceptability shows that even with few-shot demonstrations, models have virtually 0% success in producing fully accurate parses. Eliciting responses in natural language produces similar patterns of errors. Overall, our findings indicate that these models out-of-the-box can accurately identify some core aspects of semantic structure, but there remain key limitations in their ability to support fully accurate semantic analyses or parses.

09:00-10:30 (East Foyer / Virtual)
**Energy and Carbon Considerations of Fine-Tuning BERT**
*Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler and Sasha Luccioni*
Despite the popularity of the pre-train then fine-tune paradigm in the NLP community, existing work quantifying energy costs and associated carbon emissions has largely focused on language model pre-training. Although a single pre-training run draws substantially more energy than fine-tuning, fine-tuning is performed more frequently by many more individual actors, and thus must be accounted for when considering the energy and carbon footprint of NLP. In order to better characterize the role of fine-tuning in the landscape of energy and carbon emissions in NLP, we perform a careful empirical study of the computational costs of fine-tuning across tasks, datasets, hardware infrastructure and measurement modalities. Our experimental results allow us to place fine-tuning energy and carbon costs into perspective with respect to pre-training and inference, and outline recommendations to NLP researchers and practitioners who wish to improve their fine-tuning energy efficiency.

09:00-10:30 (East Foyer / Virtual)
**Evaluating Emotion Arcs Across Languages: Bridging the Global Divide in Sentiment Analysis**
*Daniela Teodorescu and Saif M. Mohammad*
Emotion arcs capture how an individual (or a population) feels over time. They are widely used in industry and research; however, there is little work on evaluating the automatically generated arcs. This is because of the difficulty of establishing the true (gold) emotion arc. Our work, for the first time, systematically and quantitatively evaluates automatically generated emotion arcs. We also compare two common ways of generating emotion arcs: Machine-Learning (ML) models and Lexicon-Only (LexO) methods. By running experiments on 18 diverse datasets in 9 languages, we show that despite being markedly poor at instance level emotion classification, LexO methods are highly accurate at generating emotion arcs when aggregating information from hundreds of instances. We also show, through experiments on six indigenous African languages, as well as Arabic, and Spanish, that automatic translations of English emotion lexicons can be used to generate high-quality emotion arcs in less-resource languages. This opens up avenues for work on emotions in languages from around the world; which is crucial for commerce, public policy, and health research in service of speakers often left behind. Code and resources:

https://github.com/dteodore/EmotionArcs

09:00-10:30 (East Foyer / Virtual)
**A Closer Look into Using Large Language Models for Automatic Evaluation**
*Cheng-Han Chiang and Hung-yi Lee*
Using large language models (LLMs) to evaluate text quality has recently gained popularity. Some existing prior works explore the idea of using LLMs for evaluation, while they differ in some details of the evaluation process. In this paper, we analyze *LLM evaluation* and *G-Eval*, and we discuss how those details in the evaluation process change how well the ratings given by LLMs correlate with human ratings. We find that the auto Chain-of-Thought (CoT) used in G-Eval does not always make G-Eval more aligned with human ratings. We also show that forcing the LLM to output only a numeric rating, as in G-Eval, is suboptimal. Last, we reveal that asking the LLM to explain its own ratings consistently improves the correlation between the ChatGPT and human ratings and pushes state-of-the-art (SoTA) correlations on two meta-evaluation datasets.

09:00-10:30 (East Foyer / Virtual)
**Sound of Story: Multi-modal Storytelling with Audio**
*Jaeyeon Bae, Seokhoon Jeong, Seokun Kang, Namgi Han, Jae-Yon Lee, Hyounghun Kim and Taehwan Kim*
Storytelling is multi-modal in the real world. When one tells a story, one may use all of the visualizations and sounds along with the story itself. However, prior studies on storytelling datasets and tasks have paid little attention to sound even though sound also conveys meaningful semantics of the story. Therefore, we propose to extend story understanding and telling areas by establishing a new component called background sound which is story context-based audio without any linguistic information. For this purpose, we introduce a new dataset, called Sound of Story (SoS), which has paired image and text sequences with corresponding sound or background music for a story. To the best of our knowledge, this is the largest well-curated dataset for storytelling with sound. Our SoS dataset consists of 27,354 stories with 19.6 images per story and 984 hours of speech-decoupled audio such as background music and other sounds. As benchmark tasks for storytelling with sound and the dataset, we propose retrieval tasks between modalities, and audio generation tasks from image-text sequences, introducing strong baselines for them. We believe the proposed dataset and tasks may shed light on the multi-modal understanding of storytelling in terms of sound.

09:00-10:30 (East Foyer / Virtual)
**RSVP: Customer Intent Detection via Agent Response Contrastive and Generative Pre-Training**
*Yu-Chien Tang, Wei-Yao Wang, An-Zi Yen and Wen-Chih Peng*
The dialogue systems in customer services have been developed with neural models to provide users with precise answers and round-the-clock support in task-oriented conversations by detecting customer intents based on their utterances. Existing intent detection approaches have highly relied on adaptively pre-training language models with large-scale datasets, yet the predominant cost of data collection may hinder their superiority. In addition, they neglect the information within the conversational responses of the agents, which have a lower collection cost, but are significant to customer intent as agents must tailor their replies based on the customers' intent. In this paper, we propose RSVP, a self-supervised framework dedicated to task-oriented dialogues, which utilizes agent responses for pre-training in a two-stage manner. Specifically, we introduce two pre-training tasks to incorporate the relations of utterance-response pairs: 1) Response Retrieval by selecting a correct response from a batch of candidates, and 2) Response Generation by mimicking agents to generate the response to a given utterance. Our benchmark results for two real-world customer service datasets show that RSVP significantly outperforms the state-of-the-art baselines by 4.95% for accuracy, 3.4% for MRR@3, and 2.75% for MRR@5 on average. Extensive case studies are investigated to show the validity of incorporating agent responses into the pre-training stage.

09:00-10:30 (East Foyer / Virtual)
**BERTwich: Extending BERT's Capabilities to Model Dialectal and Noisy Text**
*Aarohi Srivastava and David Chiang*
Real-world NLP applications often deal with nonstandard text (e.g., dialectal, informal, or misspelled text). However, language models like BERT deteriorate in the face of dialect variation or noise. How do we push BERT's modeling capabilities to encompass nonstandard text? Fine-tuning helps, but it is designed for specializing a model to a task and does not seem to bring about the deeper, more pervasive changes needed to adapt a model to nonstandard language. In this paper, we introduce the novel idea of sandwiching BERT's encoder stack between additional encoder layers trained to perform masked language modeling on noisy text. We find that our approach, paired with recent work on including character-level noise in fine-tuning data, can promote zero-shot transfer to dialectal text, as well as reduce the distance in the embedding space between words and their noisy counterparts.

09:00-10:30 (East Foyer / Virtual)
**Quick Back-Translation for Unsupervised Machine Translation**
*Benjamin Lincoln Brimacombe and Jiawei Zhou*
The field of unsupervised machine translation has seen significant advancement from the marriage of the Transformer and the back-translation algorithm. The Transformer is a powerful generative model, and back-translation leverages Transformer's high-quality translations for iterative self-improvement. However, the Transformer is encumbered by the run-time of autoregressive inference during back-translation, and back-translation is limited by a lack of synthetic data efficiency. We propose a two-for-one improvement to Transformer back-translation: Quick Back-Translation (QBT). QBT re-purposes the encoder as a generative model, and uses encoder-generated sequences to train the decoder in conjunction with the original autoregressive back-translation step, improving data throughput and utilization. Experiments on various WMT benchmarks demonstrate that a relatively small number of refining steps of QBT improve current unsupervised machine translation models, and that QBT dramatically outperforms standard back-translation only method in terms of training efficiency for comparable translation qualities.

09:00-10:30 (East Foyer / Virtual)
**RobustEmbed: Robust Sentence Embeddings Using Self-Supervised Contrastive Pre-Training**
*Javad Rafiei Asl, Eduardo Blanco and Daniel Takabi*
Pre-trained language models (PLMs) have demonstrated their exceptional performance across a wide range of natural language processing tasks. The utilization of PLM-based sentence embeddings enables the generation of contextual representations that capture rich semantic information. However, despite their success with unseen samples, current PLM-based representations suffer from poor robustness in adversarial scenarios. In this paper, we propose RobustEmbed, a self-supervised sentence embedding framework that enhances both generalization and robustness in various text representation tasks and against diverse adversarial attacks. By generating high-risk adversarial perturbations to promote higher invariance in the embedding space and leveraging the perturbation within a novel contrastive objective approach, RobustEmbed effectively learns high-quality sentence embeddings. Our extensive experiments validate the superiority of RobustEmbed over previous state-of-the-art self-supervised representations in adversarial settings, while also showcasing relative improvements in seven semantic textual similarity (STS) tasks and six transfer tasks. Specifically, our framework achieves a significant reduction in attack success rate from 75.51% to 39.62% for the BERTAttack attack technique, along with enhancements of 1.20% and 0.40% in STS tasks and transfer tasks, respectively.

09:00-10:30 (East Foyer / Virtual)
**Sub-network Discovery and Soft-masking for Continual Learning of Mixed Tasks**
*Zixuan Ke, Bing Liu, Wenhan Xiong, Asli Celikyilmaz and Haoran Li*
Continual learning (CL) has two main objectives: preventing catastrophic forgetting (CF) and encouraging knowledge transfer (KT). The existing literature mainly focused on overcoming CF. Some work has also been done on KT when the tasks are similar. To our knowledge, only one method has been proposed to learn a sequence of mixed tasks. However, these techniques still suffer from CF and/or limited KT. This paper proposes a new CL method to achieve both. It overcomes CF by isolating the knowledge of each task via discovering a sub-network for it. A soft-masking mechanism is also proposed to preserve the previous knowledge and to enable the new task to leverage the past knowledge to achieve KT. Experiments using classification, generation, information extraction, and their mixture (i.e., heterogeneous tasks) show that the proposed method consistently outperforms strong baselines.

09:00-10:30 (East Foyer / Virtual)
**Calibrated Seq2seq Models for Efficient and Generalizable Ultra-fine Entity Typing**
*Yanlin Feng, Adithya Pratapa and David R Mortensen*
Ultra-fine entity typing plays a crucial role in information extraction by predicting fine-grained semantic types for entity mentions in text. However, this task poses significant challenges due to the massive number of entity types in the output space. The current state-of-the-art approaches, based on standard multi-label classifiers or cross-encoder models, suffer from poor generalization performance or inefficient inference speed. In this paper, we present CASENT, a seq2seq model designed for ultra-fine entity typing that predicts ultra-fine types with calibrated confidence scores. Our model takes an entity mention as input and employs constrained beam search to generate multiple types autoregressively. The raw sequence probabilities associated with the predicted types are then transformed into confidence scores using a novel calibration method. We conduct extensive experiments on the UFET dataset which contains over $10k$ types. Our method outperforms the previous state-of-the-art in terms of F1 score and calibration error, while achieving an inference speedup of over 50 times. Additionally, we demonstrate the generalization capabilities of our model by evaluating it in zero-shot and few-shot settings on five specialized domain entity typing datasets that are unseen during training. Remarkably, our model outperforms large language models with 10 times more parameters in the zero-shot setting, and when fine-tuned on 50 examples, it significantly outperforms ChatGPT on all datasets.

09:00-10:30 (East Foyer / Virtual)
**Gradually Excavating External Knowledge for Implicit Complex Question Answering**
*Chang Liu, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Edmund Y. Lam and Ngai Wong*
Recently, large language models (LLMs) have gained much attention for the emergence of human-comparable capabilities and huge potential. However, for open-domain implicit question-answering problems, LLMs may not be the ultimate solution due to the reasons of: 1) uncovered or out-of-date domain knowledge, 2) one-shot generation and hence restricted comprehensiveness. To this end, this work proposes a gradual knowledge excavation framework for open-domain complex question answering, where LLMs iteratively and actively acquire extrinsic information, then reason based on acquired historical knowledge. Specifically, during each step of the solving process, the model selects an action to execute, such as querying external knowledge or performing a single logical reasoning step, to gradually progress toward a final answer. Our method can effectively leverage plug-and-play external knowledge and dynamically adjust the strategy for solving complex questions. Evaluated on the StrategyQA dataset, our method achieves 78.17% accuracy with less than 6% parameters of its competitors, setting new SOTA in the $\sim$10B LLM class.

09:00-10:30 (East Foyer / Virtual)
**Mandarin classifier systems optimize to accommodate communicative pressures**
*Yamei Wang and Géraldine Walther*
Previous work on noun classification implies that gender systems are inherently optimized to accommodate communicative pressures on human language learning and processing (Dye. et al 2017, 2018). They state that languages make use of either grammatical (e.g., gender) or probabilistic (pre-nominal modifiers) to smoothe the entropy of nouns in context. We show that even languages that are considered genderless, like Mandarin Chinese, possess a noun classification device that plays the same functional role as gender markers. Based on close to 1M Mandarin noun phrases extracted from the Leipzig Corpora Collection (Goldhahn et al. 2012) and their corresponding fastText embeddings (Bojanowski et al. 2016), we show that noun-classifier combinations are sensitive to same frequency, similarity, and co-occurrence interactions that structure gender systems. We also present the first study of the effects of the interaction between grammatical and probabilisitic noun classification.

09:00-10:30 (East Foyer / Virtual)
**Improving Span Representation by Efficient Span-Level Attention**
*Pengyu Ji, Songlin Yang and Kewei Tu*
High-quality span representations are crucial to natural language processing tasks involving span prediction and classification. Most existing methods derive a span representation by aggregation of token representations within the span. In contrast, we aim to improve span representations by considering span-span interactions as well as more comprehensive span-token interactions. Specifically, we introduce layers of span-level attention on top of a normal token-level transformer encoder. Given that attention between all span pairs results in $O(n^4)$ complexity ($n$ being the sentence length) and not all span interactions are intuitively meaningful, we restrict the range of spans that a given span could attend to, thereby reducing overall complexity to $O(n^3)$. We conduct experiments on various span-related tasks and show superior performance of our model surpassing baseline models. Our code is publicly available at https://github.com/jipy0222/Span-Level-Attention.

09:00-10:30 (East Foyer / Virtual)
**Target-Aware Spatio-Temporal Reasoning via Answering Questions in Dynamic Audio-Visual Scenarios**
*Yuanyuan Jiang and Jianqin Yin*
Audio-visual question answering (AVQA) is a challenging task that requires multistep spatio-temporal reasoning over multimodal contexts. Recent works rely on elaborate target-agnostic parsing of audio-visual scenes for spatial grounding while mistreating audio and video as separate entities for temporal grounding. This paper proposes a new target-aware joint spatio-temporal grounding network for AVQA. It consists of two key components: the target-aware spatial grounding module (TSG) and the single-stream joint audio-visual temporal grounding module (JTG). The TSG can focus on audio-visual cues relevant to the query subject by utilizing explicit semantics from the question. Unlike previous two-stream temporal grounding modules that required an additional audio-visual fusion module, JTG incorporates audio-visual fusion and question-aware temporal grounding into one module with a simpler single-stream architecture. The temporal synchronization between audio and video in the JTG is facilitated by our proposed cross-modal synchrony loss (CSL). Extensive experiments verified the effectiveness of our proposed method over existing state-of-the-art methods.

09:00-10:30 (East Foyer / Virtual)

**BotPercent: Estimating Bot Populations in Twitter Communities**
*Zhaoxuan Tan, Shangbin Feng, Melanie Sclar, Herun Wan, Minnan Luo, Yejin Choi and Yulia Tsvetkov*
Twitter bot detection is vital in combating misinformation and safeguarding the integrity of social media discourse. While malicious bots are becoming more and more sophisticated and personalized, standard bot detection approaches are still agnostic to social environments (henceforth, communities) the bots operate at. In this work, we introduce community-specific bot detection, estimating the percentage of bots given the context of a community. Our method—BotPercent—is an amalgamation of Twitter bot detection datasets and feature-, text-, and graph-based models, adjusted to a particular community on Twitter. We introduce an approach that performs confidence calibration across bot detection models, which addresses generalization issues in existing community-agnostic models targeting individual bots and leads to more accurate community-level bot estimations. Experiments demonstrate that BotPercent achieves state-of-the-art performance in community-level Twitter bot detection across both balanced and imbalanced class distribution settings, presenting a less biased estimator of Twitter bot populations within the communities we analyze. We then analyze bot rates in several Twitter groups, including users who engage with partisan news media, political communities in different countries, and more. Our results reveal that the presence of Twitter bots is not homogeneous, but exhibiting a spatial-temporal distribution with considerable heterogeneity that should be taken into account for content moderation and social media policy making. The implementation of BotPercent is available at https://github.com/TamSiuhin/BotPercent.

09:00-10:30 (East Foyer / Virtual)
**Crossing the Aisle: Unveiling Partisan and Counter-Partisan Events in News Reporting**
*Kaijian Zou, Xinliang Frederick Zhang, Winston Wu, Nicholas Beauchamp and Lu Wang*
News media is expected to uphold unbiased reporting. Yet they may still affect public opinion by selectively including or omitting events that support or contradict their ideological positions. Prior work in NLP has only studied media bias via linguistic style and word usage. In this paper, we study to which degree media balances news reporting and affects consumers through event inclusion or omission. We first introduce the task of detecting both partisan and counter-partisan events: events that support or oppose the author's political ideology. To conduct our study, we annotate a high-quality dataset, PAC, containing $8,511$ (counter-)partisan event annotations in $304$ news articles from ideologically diverse media outlets. We benchmark PAC to highlight the challenges of this task. Our findings highlight both the ways in which the news subtly shapes opinion and the need for large language models that better understand events within a broader context. Our dataset can be found at https://github.com/launchnlp/Partisan-Event-Dataset.

09:00-10:30 (East Foyer / Virtual)
**Dense Retrieval as Indirect Supervision for Large-space Decision Making**
*Nan Xu, Fei Wang, Mingtao Dong and Muhao Chen*
Many discriminative natural language understanding (NLU) tasks have large label spaces. Learning such a process of large-space decision making is particularly challenging due to the lack of training instances per label and the difficulty of selection among many fine-grained labels. Inspired by dense retrieval methods for passage finding in open-domain QA, we propose a reformulation of large-space discriminative NLU tasks as a learning-to-retrieve task, leading to a novel solution named Dense Decision Retrieval (DDR). Instead of predicting fine-grained decisions as logits, DDR adopts a dual-encoder architecture that learns to predict by retrieving from a decision thesaurus. This approach not only leverages rich indirect supervision signals from easy-to-consume learning resources for dense retrieval, it also leads to enhanced prediction generalizability with a semantically meaningful representation of the large decision space. When evaluated on tasks with decision spaces ranging from hundreds to hundred-thousand scales, DDR outperforms strong baselines greatly by 27.54% in P @1 on two extreme multi-label classification tasks, 1.17% in F1 score ultra-fine entity typing, and 1.26% in accuracy on three few-shot intent classification tasks on average.

09:00-10:30 (East Foyer / Virtual)
**RegaVAE: A Retrieval-Augmented Gaussian Mixture Variational Auto-Encoder for Language Modeling**
*Jingcheng Deng, Liang Pang, Huawei Shen and Xueqi Cheng*
Retrieval-augmented language models show promise in addressing issues like outdated information and hallucinations in language models (LMs). However, current research faces two main problems: 1) determining what information to retrieve, and 2) effectively combining retrieved information during generation. We argue that valuable retrieved information should not only be related to the current source text but also consider the future target text, given the nature of LMs that model future tokens. Moreover, we propose that aggregation using latent variables derived from a compact latent space is more efficient than utilizing explicit raw text, which is limited by context length and susceptible to noise. Therefore, we introduce RegaVAE, a retrieval-augmented language model built upon the variational auto-encoder (VAE). It encodes the text corpus into a latent space, capturing current and future information from both source and target text. Additionally, we leverage the VAE to initialize the latent space and adopt the probabilistic form of the retrieval generation paradigm by expanding the Gaussian prior distribution into a Gaussian mixture distribution. Theoretical analysis provides an optimizable upper bound for RegaVAE. Experimental results on various datasets demonstrate significant improvements in text generation quality and hallucination removal.

09:00-10:30 (East Foyer / Virtual)
**Hierarchical Prompting Assists Large Language Model on Web Navigation**
*Abishek Sridhar, Robert Lo, Frank F. Xu, Hao Zhu and Shuyan Zhou*
Large language models (LLMs) struggle on processing complicated observations in interactive decision making. To alleviate this issue, we propose a simple hierarchical prompting approach. Diverging from previous prompting approaches that always put the full observation (a web page) to the prompt, we propose to first construct an action-aware observation which is more condensed and relevant with a dedicated Summarizer prompt. The Actor prompt then predicts the next action based on the summarized history. While our method has broad applicability, we particularly demonstrate its efficacy in the complex domain of web navigation where a full observation often contains redundant and irrelevant information. Our approach outperforms the previous state-of-the-art prompting mechanism with the same LLM by 6.2% on task success rate, demonstrating its potential on interactive decision making tasks with long observation traces.

09:00-10:30 (East Foyer / Virtual)
**Conic10K: A Challenging Math Problem Understanding and Reasoning Dataset**
*Haoyi Wu, Wenyang Hui, Yezeng Chen, Weiqi Wu, Kewei Tu and Yi Zhou*
Mathematical understanding and reasoning are crucial tasks for assessing the capabilities of artificial intelligence (AI). However, existing benchmarks either require just a few steps of reasoning, or only contain a small amount of data in one specific topic, making it hard to analyse AI's behaviour with reference to different problems within a specific topic in detail. In this work, we propose Conic10K, a challenging math problem dataset on conic sections in Chinese senior high school education. Our dataset contains various problems with different reasoning depths, while only the knowledge from conic sections is required. Since the dataset only involves a narrow range of knowledge, it is easy to separately analyse the knowledge a model possesses and the reasoning ability it has. For each problem, we provide a high-quality formal representation, the reasoning steps, and the final solution. Experiments show that existing large language models, including GPT-4, exhibit weak performance on complex reasoning. We hope that our findings could inspire more advanced techniques for precise natural language understanding and reasoning. Our dataset and codes are available at https://github.com/whyNLP/Conic10K.

09:00-10:30 (East Foyer / Virtual)
**Quality Estimation-Assisted Automatic Post-Editing**
*Sourabh Dattatray Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe and Pushpak Bhattacharyya*
Automatic Post-Editing (APE) systems are prone to over-correction of the Machine Translation (MT) outputs. While Word-level Quality Estimation (QE) system can provide a way to curtail the over-correction, a significant performance gain has not been observed thus far by utilizing existing APE and QE combination strategies. In this paper, we propose joint training of a model on APE and QE tasks to improve the APE. Our proposed approach utilizes a multi-task learning (MTL) methodology, which shows significant improvement while treating both tasks as a 'bargaining game' during training. Moreover, we investigate various existing combination strategies and show that our approach achieves state-of-the-art performance for a 'distant' language pair, viz., English-Marathi. We observe an improvement of 1.09 TER and 1.37 BLEU points over a baseline QE-Unassisted APE system for English-Marathi, while also observing 0.46 TER and 0.62 BLEU points for English-German. Further, we discuss the results qualitatively and show how our approach helps reduce over-correction, thereby improving the APE performance. We also observe that the degree of integration between QE and APE directly correlates with the APE performance gain. We release our code and models publicly.

09:00-10:30 (East Foyer / Virtual)
**Dissecting In-Context Learning of Translations in GPT-3**
*Vikas Raunak, Arul Menezes and Hany Hassan Awadalla*
Most of the recent work in leveraging Large Language Models (LLMs) such as GPT-3 for Machine Translation (MT) has focused on selecting the few-shot samples for prompting. In this work, we try to better understand the role of demonstration attributes for the in-context learning of translations through perturbations of high-quality, in-domain demonstrations. We find that asymmetric perturbation of the source-target mappings yield vastly different results. We show that the perturbation of the source side has surprisingly little impact, while target perturbation can drastically reduce translation quality, suggesting that it is the output text distribution that provides the most important learning signal during in-context learning of translations. We propose a method named Zero-Shot-Context to add this signal automatically in Zero-Shot prompting. We demonstrate that it improves upon the zero-shot translation performance of GPT-3, even making it competitive with few-shot prompted translations.

09:00-10:30 (East Foyer / Virtual)
**Time-Considerable Dialogue Models via Reranking by Time Dependency**
*Yuiko Tsunomori, Masakazu Ishihata and Hiroaki Sugiyama*
In the last few years, generative dialogue models have shown excellent performance and have been used for various applications. As chatbots become more prevalent in our daily lives, more and more people expect them to behave more like humans, but existing dialogue models do not consider the time information that people are constantly aware of. In this paper, we aim to construct a time-considerable dialogue model that actively utilizes time information. First, we categorize responses by their naturalness at different times and introduce a new metric to classify responses into our categories. Then, we propose a new reranking method to make the existing dialogue model time-considerable using the proposed metric and subjectively evaluate the performances of the obtained time-considerable dialogue models by humans.

09:00-10:30 (East Foyer / Virtual)
**Manifold-Preserving Transformers are Effective for Short-Long Range Encoding**
*Ayan Sengupta, Md Shad Akhtar and Tanmoy Chakraborty*
Multi-head self-attention-based Transformers have shown promise in different learning tasks. Albeit these models exhibit significant improvement in understanding short-term and long-term contexts from sequences, encoders of Transformers and their variants fail to preserve layer-wise contextual information. Transformers usually project tokens onto sparse manifolds and fail to preserve mathematical equivalence among the token representations. In this work, we propose TransJect, an encoder model that guarantees a theoretical bound for layer-wise distance preservation between a pair of tokens. We propose a simple alternative to dot-product attention to ensure Lipschitz continuity. This allows TransJect to learn injective mappings to transform token representations to different manifolds with similar topology and preserve Euclidean distance between every pair of tokens in subsequent layers. Evaluations across multiple benchmark short- and long-sequence classification tasks show maximum improvements of 6.8% and 5.9%, respectively, over the variants of Transformers. Additionally, TransJect displays 79% better performance than Transformer on the language modeling task. We further highlight the shortcomings of multi-head self-attention from the statistical physics viewpoint. Although multi-head self-attention was incepted to learn different abstraction levels within the networks, our empirical analyses suggest that different attention heads learn randomly and unorderly. In contrast, TransJect adapts a mixture of experts for regularization; these experts are more orderly and balanced and learn different sparse representations from the input sequences. TransJect exhibits very low entropy and can be efficiently scaled to larger depths.

09:00-10:30 (East Foyer / Virtual)
**A Parallel Corpus for Vietnamese Central-Northern Dialect Text Transfer**
*Thang Le and Anh Tuan Luu*
The Vietnamese language embodies dialectal variants closely attached to the nation's three macro-regions: the Northern, Central and Southern regions. As the northern dialect forms the basis of the standard language, it's considered the prestige dialect. While the northern dialect differs from the remaining two in certain aspects, it almost shares an identical lexicon with the southern dialect, making the textual attributes nearly interchangeable. In contrast, the central dialect possesses a number of unique vocabularies and is less mutually intelligible to the standard dialect. Through preliminary experiments, we observe that current NLP models do not possess understandings of the Vietnamese central dialect text, which most likely originates from the lack of resources. To facilitate research on this domain, we introduce a new parallel corpus for Vietnamese central-northern dialect text transfer. Via exhaustive benchmarking, we discover monolingual language models' superiority over their multilingual counterparts on the dialect transfer task. We further demonstrate that fine-tuned transfer models can seamlessly improve the performance of existing NLP systems on the central dialect domain with dedicated results in translation and text-image retrieval tasks.

## Industry 6

09:00-10:30 (East Foyer / Virtual)

---

09:00-10:30 (East Foyer / Virtual)
**CDD: A Large Scale Dataset for Legal Intelligence Research**
*Changzhen Ji, Yating Zhang, Adam Jatowt and Haipang Wu*
As an important application of Artificial Intelligence, legal intelligence has recently attracted the attention of many researchers. Previous works investigated diverse issues like predicting crimes, predicting outcomes of judicial debates, or extracting information/knowledge from various kinds of legal documents. Although many advances have been made, the research on supporting prediction of court judgments remains relatively scarce, while the lack of large-scale data resources limits the development of this research.In this paper, we present a novel, large-

size Court Debate Dataset (CDD), which includes 30,481 court cases, totaling 1,144,425 utterances. CDD contains real-world conversations involving judges, plaintiffs and defendants in court trials. To construct this dataset we have invited experienced judges to design appropriate labels for data records. We then asked law school students to provide annotations based on the defined labels. The dataset can be applied to several downstream tasks, such as text summarization, dialogue generation, text classification, etc. We introduce the details of the different tasks in the rapidly developing field of legal intelligence, the research of which can be fostered thanks to our dataset, and we provide the corresponding benchmark performance.

# Coffee Break

10:30-11:00 - Location: West Foyer

# Session 10: Oral & Poster - 11:00-12:30

## NLP Applications 2

11:00-12:30 (East Ballroom)

---

11:00-11:15 (East Ballroom)

**UniMath: A Foundational and Multimodal Mathematical Reasoner**

*Zhenwen Liang, Tianyu Yang, Jipeng Zhang and Xiangliang Zhang*

While significant progress has been made in natural language processing (NLP), existing methods exhibit limitations in effectively interpreting and processing diverse mathematical modalities. Therefore, we introduce UniMath, a versatile and unified system designed for multimodal mathematical reasoning tasks. Tackling complex problem-solving in arithmetic, geometry, and table-based math, UniMath utilizes a fine-tuned T5 model augmented with a variational autoencoder (VAE)-based image tokenizer. By jointly training and evaluating the model on three diverse datasets - SVAMP, GeoQA, and TableMWP, UniMath achieves state-of-the-art performance. The model's generalization ability is further demonstrated via fine-tuning on two additional datasets, MathQA and Geo-Proving. Through comprehensive evaluations, we showcase that joint training across diverse math tasks improves overall model performance and enhances its ability to generalize across different mathematical reasoning tasks. This pioneering approach provides a blueprint and inspires further efforts on unified mathematical reasoning with deep learning systems.

11:15-11:30 (East Ballroom)

**Predictive Chemistry Augmented with Text Retrieval**

*Yujie Qian, Zhening Li, Zhengkai Tu, Connor W. Coley and Regina Barzilay*

This paper focuses on using natural language descriptions to enhance predictive models in the chemistry field. Conventionally, chemoinformatics models are trained with extensive structured data manually extracted from the literature. In this paper, we introduce TextReact, a novel method that directly augments predictive chemistry with texts retrieved from the literature. TextReact retrieves text descriptions relevant for a given chemical reaction, and then aligns them with the molecular representation of the reaction. This alignment is enhanced via an auxiliary masked LM objective incorporated in the predictor training. We empirically validate the framework on two chemistry tasks: reaction condition recommendation and one-step retrosynthesis. By leveraging text retrieval, TextReact significantly outperforms state-of-the-art chemoinformatics models trained solely on molecular data.

11:30-11:45 (East Ballroom)

**Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration**

*Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu and Kun Kuang*

Legal Judgment Prediction (LJP) has become an increasingly crucial task in Legal AI, i.e., predicting the judgment of the case in terms of case fact description. Precedents are the previous legal cases with similar facts, which are the basis for the judgment of the subsequent case in national legal systems. Thus, it is worthwhile to explore the utilization of precedents in the LJP. Recent advances in deep learning have enabled a variety of techniques to be used to solve the LJP task. These can be broken down into two categories: large language models (LLMs) and domain-specific models. LLMs are capable of interpreting and generating complex natural language, while domain models are efficient in learning task-specific information. In this paper, we propose the precedent-enhanced LJP framework (PLJP) – a system that leverages the strength of both LLM and domain models in the context of precedents. Specifically, the domain models are designed to provide candidate labels and find the proper precedents efficiently, and the large models will make the final prediction with an in-context precedents comprehension. Experiments on the real-world dataset demonstrate the effectiveness of our PLJP. Moreover, our work shows a promising direction for LLM and domain-model collaboration that can be generalized to other vertical domains.

11:45-12:00 (East Ballroom)

**Hidding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection**

*Xinlin Peng, Ying Zhou, Ben He, Le Sun and Yingfei Sun*

Large language models (LLMs) have exhibited remarkable capabilities in text generation tasks. However, the utilization of these models carries inherent risks, including but not limited to plagiarism, the dissemination of fake news, and issues in educational exercises. Although several detectors have been proposed to address these concerns, their effectiveness against adversarial perturbations, specifically in the context of student essay writing, remains largely unexplored. This paper aims to bridge this gap by constructing AIG-ASAP, an AI-generated student essay dataset, employing a range of text perturbation methods that are expected to generate high-quality essays while evading detection. Through empirical experiments, we assess the performance of current AIGC detectors on the AIG-ASAP dataset. The results reveal that the existing detectors can be easily circumvented using straightforward automatic adversarial attacks. Specifically, we explore word substitution and sentence substitution perturbation methods that effectively evade detection while maintaining the quality of the generated essays. This highlights the urgent need for more accurate and robust methods to detect AI-generated student essays in the education domain. Code and data are released for public use.

12:00-12:15 (East Ballroom)

**Analyzing Norm Violations in Live-Stream Chat**

*Jihyung Moon, Dong-Ho Lee, Hyundong Justin Cho, Woojeong Jin, Chan Young Park, Minwoo Kim, Jonathan May, Jay Pujara and Sungjoon Park*

Toxic language, such as hate speech, can deter users from participating in online communities and enjoying popular platforms. Previous approaches to detecting toxic language and norm violations have been primarily concerned with conversations from online forums and social media, such as Reddit and Twitter. These approaches are less effective when applied to conversations on live-streaming platforms, such as Twitch and YouTube Live, as each comment is only visible for a limited time and lacks a thread structure that establishes its relationship with

other comments. In this work, we share the first NLP study dedicated to detecting norm violations in conversations on live-streaming platforms. We define norm violation categories in live-stream chats and annotate 4,583 moderated comments from Twitch. We articulate several facets of live-stream data that differ from other forums, and demonstrate that existing models perform poorly in this setting. By conducting a user study, we identify the informational context humans use in live-stream moderation, and train models leveraging context to identify norm violations. Our results show that appropriate contextual information can boost moderation performance by 35%.

12:15-12:30 (East Ballroom)
**ALCAP: Alignment-Augmented Music Captioner**
*Zihao He, Weituo Hao, Wei-Tsung Lu, Changyou Chen, Kristina Lerman and Xuchen Song*
Music captioning has gained significant attention in the wake of the rising prominence of streaming media platforms. Traditional approaches often prioritize either the audio or lyrics aspect of the music, inadvertently ignoring the intricate interplay between the two. However, a comprehensive understanding of music necessitates the integration of both these elements. In this study, we delve into this overlooked realm by introducing a method to systematically learn multimodal alignment between audio and lyrics through contrastive learning. This not only recognizes and emphasizes the synergy between audio and lyrics but also paves the way for models to achieve deeper cross-modal coherence, thereby producing high-quality captions. We provide both theoretical and empirical results demonstrating the advantage of the proposed method, which achieves new state-of-the-art on two music captioning datasets.

## Resources and Evaluation 2

11:00-12:30 (Central 1 Ballroom)

---

11:00-11:15 (Central 1 Ballroom)
**Unveiling the Essence of Poetry: Introducing a Comprehensive Dataset and Benchmark for Poem Summarization**
*Ridwan Mahbub, Ifrad Towhid Khan, Samiha Shafiq Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar and Sabbir Ahmed*
While research in natural language processing has progressed significantly in creative language generation, the question of whether language models can interpret the intended meaning of creative language largely remains unanswered. Poetry as a creative art form has existed for generations, and summarization of such content requires deciphering the figurative patterns to find out the actual intent and message of the poet. This task can provide the researchers an opportunity to evaluate the creative language interpretation capacity of the language models. Unlike typical text, summarization of poems is a challenging task as poems carry a deeper meaning, which can be easily lost if only the literal meaning is considered. That being said, we propose a new task in the field of natural language understanding called 'Poem Summarization'. As a starting, we propose the first-ever dataset for this task, named 'PoemSum', consisting of 3011 samples of poetry and its corresponding summarized interpretation in the English language. We have benchmarked the performance of different state-of-the-art summarization models and provided observations on their limitations. The dataset and all relevant code used in this work have been made publicly available.

11:15-11:30 (Central 1 Ballroom)
**Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark**
*Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu and David Jurgens*
Large language models (LLMs) have been shown to perform well at a variety of syntactic, discourse, and reasoning tasks. While LLMs are increasingly deployed in many forms including conversational agents that interact with humans, we lack a grounded benchmark to measure how well LLMs understand social language. Here, we introduce a new theory-driven benchmark, SocKET, that contains 58 NLP tasks testing social knowledge which we group into five categories: humor & sarcasm, offensiveness, sentiment & emotion, and trustworthiness. In tests on the benchmark, we demonstrate that current models attain only moderate performance but reveal significant potential for task transfer among different types and categories of tasks, which were predicted from theory. Through zero-shot evaluations, we show that pretrained models already possess some innate but limited capabilities of social language understanding and training on one category of tasks can improve zero-shot testing on others. Our benchmark provides a systematic way to analyze model performance on an important dimension of language and points to clear room for improvement to build more socially-aware LLMs. The resources are released at https://github.com/minjechoi/SOCKET.

11:30-11:45 (Central 1 Ballroom)
**It Ain't Over: A Multi-aspect Diverse Math Word Problem Dataset**
*Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak and Jongwuk Lee*
The math word problem (MWP) is a complex task that requires natural language understanding and logical reasoning to extract key knowledge from natural language narratives. Previous studies have provided various MWP datasets but lack diversity in problem types, lexical usage patterns, languages, and annotations for intermediate solutions. To address these limitations, we introduce a new MWP dataset, named DMath (Diverse Math Word Problems), offering a wide range of diversity in problem types, lexical usage patterns, languages, and intermediate solutions. The problems are available in English and Korean and include an expression tree and Python code as intermediate solutions. Through extensive experiments, we demonstrate that the DMath dataset provides a new opportunity to evaluate the capability of large language models, i.e., GPT-4 only achieves about 75% accuracy on the DMath dataset.

11:45-12:00 (Central 1 Ballroom)
**Syllogistic Reasoning for Legal Judgment Analysis**
*Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen and Pengjie Ren*
Legal judgment assistants are developing fast due to impressive progress of large language models (LLMs). However, people can hardly trust the results generated by a model without reliable analysis of legal judgement. For legal practitioners, it is common practice to utilize syllogistic reasoning to select and evaluate the arguments of the parties as part of the legal decision-making process. But the development of syllogistic reasoning for legal judgment analysis is hindered by the lack of resources: (1) there is no large-scale syllogistic reasoning dataset for legal judgment analysis, and (2) there is no set of established benchmarks for legal judgment analysis. In this paper, we construct and manually correct a syllogistic reasoning dataset for legal judgment analysis. The dataset contains 11,239 criminal cases which cover 4 criminal elements, 80 charges and 124 articles. We also select a set of large language models as benchmarks, and conduct a in-depth analysis of the capacity of their legal judgment analysis.

12:00-12:15 (Central 1 Ballroom)
**TempTabQA: Temporal Question Answering for Semi-Structured Tables**
*Vivek Gupta, Pranshu Kandoi, Mahek Bhavesh Vora, Shuo Zhang, Yujie He, Ridho Reinanda and Vivek Srikumar*
Semi-structured data, such as Infobox tables, often include temporal information about entities, either implicitly or explicitly. Can current NLP systems reason about such information in semi-structured tables? To tackle this question, we introduce the task of temporal question answering on semi-structured tables. We present a dataset, TEMPTABQA, which comprises 11,454 question-answer pairs extracted from 1,208 Wikipedia Infobox tables spanning more than 90 distinct domains. Using this dataset, we evaluate several state-of-the-art models for

temporal reasoning. We observe that even the top-performing LLMs lag behind human performance by more than 13.5 F1 points. Given these results, our dataset has the potential to serve as a challenging benchmark to improve the temporal reasoning capabilities of NLP models.

12:15-12:30 (Central 1 Ballroom)
**Multilingual Previously Fact-Checked Claim Retrieval**
*Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek and Maria Bielikova*
Fact-checkers are often hampered by the sheer amount of online content that needs to be fact-checked. NLP can help them by retrieving already existing fact-checks relevant to the content being investigated. This paper introduces a new multilingual dataset for previously fact-checked claim retrieval. We collected 28k posts in 27 languages from social media, 206k fact-checks in 39 languages written by professional fact-checkers, as well as 31k connections between these two groups. This is the most extensive and the most linguistically diverse dataset of this kind to date. We evaluated how different unsupervised methods fare on this dataset and its various dimensions. We show that evaluating such a diverse dataset has its complexities and proper care needs to be taken before interpreting the results. We also evaluated a supervised fine-tuning approach, improving upon the unsupervised method significantly.

## Semantics 2

11:00-12:30 (Central 3 Ballroom)

11:00-11:15 (Central 3 Ballroom)
**On Bilingual Lexicon Induction with Large Language Models**
*Yaoyiran Li, Anna Korhonen and Ivan Vulić*
Bilingual Lexicon Induction (BLI) is a core task in multilingual NLP that still, to a large extent, relies on calculating cross-lingual word representations. Inspired by the global paradigm shift in NLP towards Large Language Models (LLMs), we examine the potential of the latest generation of LLMs for the development of bilingual lexicons. We ask the following research question: Is it possible to prompt and fine-tune multilingual LLMs (mLLMs) for BLI, and how does this approach compare against and complement current BLI approaches? To this end, we systematically study 1) zero-shot prompting for unsupervised BLI and 2) few-shot in-context prompting with a set of seed translation pairs, both without any LLM fine-tuning, as well as 3) standard BLI-oriented fine-tuning of smaller LLMs. We experiment with 18 open-source text-to-text mLLMs of different sizes (from 0.3B to 13B parameters) on two standard BLI benchmarks covering a range of typologically diverse languages. Our work is the first to demonstrate strong BLI capabilities of text-to-text mLLMs. The results reveal that few-shot prompting with in-context examples from nearest neighbours achieves the best performance, establishing new state-of-the-art BLI scores for many language pairs. We also conduct a series of in-depth analyses and ablation studies, providing more insights on BLI with (m)LLMs, also along with their limitations.

11:15-11:30 (Central 3 Ballroom)
**Random Entity Quantization for Parameter-Efficient Compositional Knowledge Graph Representation**
*Jiaang Li, Quan Wang, Yi Liu, Licheng Zhang and Zhendong Mao*
Representation Learning on Knowledge Graphs (KGs) is essential for downstream tasks. The dominant approach, KG Embedding (KGE), represents entities with independent vectors and faces the scalability challenge. Recent studies propose an alternative way for parameter efficiency, which represents entities by composing entity-corresponding codewords matched from predefined small-scale codebooks. We refer to the process of obtaining corresponding codewords of each entity as entity quantization, for which previous works have designed complicated strategies. Surprisingly, this paper shows that simple random entity quantization can achieve similar results to current strategies. We analyze this phenomenon and reveal that entity codes, the quantization outcomes for expressing entities, have higher entropy at the code level and Jaccard distance at the codeword level under random entity quantization. Therefore, different entities become more easily distinguished, facilitating effective KG representation. The above results show that current quantization strategies are not critical for KG representation, and there is still room for improvement in entity distinguishability beyond current strategies.

11:30-11:45 (Central 3 Ballroom)
**Systematic word meta-sense extension**
*Lei Yu*
The meaning of polysemous words often varies in a highly productive yet predictable way. Generalizing the regularity between conventional senses to derive novel word meaning is crucial for automated processing of non-literal language uses such as figurative expressions. We introduce a novel task called systematic word meta-sense extension (SWORME) to test and improve language models' ability to extend word meaning to denote new semantic domains (also called meta-senses) that bear regular semantic relations with existing senses. We found that language models prefer incremental lexical semantic change toward conceptually similar meta-senses such as logical metonymy, and are much worse at predicting highly non-literal meaning extensions such as metaphors. We propose a novel analogy-based method of word meaning extension, and show that it effectively improves language model systematicity in making both gradual and radical types of meta-sense extension. We further demonstrate that learning systematic meta-sense extensions benefits language models on multiple benchmarks of figurative language understanding.

11:45-12:00 (Central 3 Ballroom)
**Pragmatic Reasoning Unlocks Quantifier Semantics for Foundation Models**
*Yiyuan Li, Rakesh R Menon, Sayan Ghosh and Shashank Srivastava*
Generalized quantifiers (e.g., *few*, *most*) are used to indicate the proportions predicates satisfy (for example, *some* apples are red). One way to interpret quantifier semantics is to explicitly bind these satisfactions with percentage scopes (e.g., 30%-40% of apples are red). This approach can be helpful for tasks like logic formalization and surface-form quantitative reasoning (Gordon and Schubert, 2010; Roy et al., 2015). However, it remains unclear if recent foundation models (Bommasani et al., 2021) possess this ability due to the absence of direct training signals. To explore this, we introduce QuRe, a crowd-sourced dataset of human-annotated generalized quantifiers in Wikipedia sentences featuring percentage-equipped predicates. We explore quantifier comprehension using PRESQUE, a framework that combines natural language inference and the Rational Speech Acts framework. Experimental results on the HVD dataset (Herbelot and Vecchi, 2015) and QuRe demonstrate PRESQUE's superiority over a literal listener baseline, showing a 20% relative improvement in F1 in predicting percentage scopes for quantifiers, even with no additional training.

12:00-12:15 (Central 3 Ballroom)
**Non-Programmers Can Label Programs Indirectly via Active Examples: A Case Study with Text-to-SQL**
*Ruiqi Zhong, Charlie Victor Snell, Dan Klein and Jason Eisner*
Can non-programmers annotate natural language utterances with complex programs that represent their meaning? We introduce APEL, a

framework in which non-programmers select among candidate programs generated by a seed semantic parser (e.g., Codex). Since they cannot understand the candidate programs, we ask them to select indirectly by examining the programs' input-ouput examples. For each utterance, APEL actively searches for a simple input on which the candidate programs tend to produce different outputs. It then asks the non-programmers only to choose the appropriate output, thus allowing us to infer which program is correct and could be used to fine-tune the parser. As a first case study, we recruited human non-programmers to use APEL to re-annotate SPIDER, a text-to-SQL dataset. Our approach achieved the same annotation accuracy as the original expert annotators (75%) and exposed many subtle errors in the original annotations.

12:15-12:30 (Central 3 Ballroom)
**Improving Language Models' Meaning Understanding and Consistency by Learning Conceptual Roles from Dictionary**
*Myeongjun Erik Jang and Thomas Lukasiewicz*
The non-humanlike behaviour of contemporary pre-trained language models (PLMs) is a leading cause undermining their trustworthiness. A striking phenomenon of such faulty behaviours is the generation of inconsistent predictions, which produces logically contradictory results, such as generating different predictions for texts delivering the same meaning or violating logical properties. Previous studies exploited data augmentation or implemented specialised loss functions to alleviate the issue. However, their usage is limited, because they consume expensive training resources for large-sized PLMs and can only handle a certain consistency type. To this end, we propose a practical approach that alleviates the inconsistent behaviour issue by fundamentally improving PLMs' meaning awareness. Based on the conceptual role theory, our method allows PLMs to capture accurate meaning by learning precise interrelationships between concepts from word-definition pairs in a dictionary. Next, we propose an efficient parameter integration technique that updates only a few additional parameters to combine the learned interrelationship with PLMs' pre-trained knowledge. Our experimental results reveal that the approach can concurrently improve multiple types of consistency, enables efficient knowledge integration, and easily applies to other languages.

## Speech & Multimodality 2

11:00-12:30 (West 1 Ballroom)

---

11:00-11:15 (West 1 Ballroom)
**Rethinking and Improving Multi-task Learning for End-to-end Speech Translation**
*Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang and Jingbo Zhu*
Significant improvements in end-to-end speech translation (ST) have been achieved through the application of multi-task learning. However, the extent to which auxiliary tasks are highly consistent with the ST task, and how much this approach truly helps, have not been thoroughly studied. In this paper, we investigate the consistency between different tasks, considering different times and modules. We find that the textual encoder primarily facilitates cross-modal conversion, but the presence of noise in speech impedes the consistency between text and speech representations. Furthermore, we propose an improved multi-task learning (IMTL) approach for the ST task, which bridges the modal gap by mitigating the difference in length and representation. We conduct experiments on the MuST-C dataset. The results demonstrate that our method attains state-of-the-art results. Moreover, when additional data is used, we achieve the new SOTA result on MuST-C English to Spanish task with 20.8% of the training time required by the current SOTA method.

11:15-11:30 (West 1 Ballroom)
**Unsupervised Sounding Pixel Learning**
*Yining Zhang, Yanli Ji and Yang Yang*
Sounding source localization is a challenging cross-modal task due to the difficulty of cross-modal alignment. Although supervised cross-modal methods achieve encouraging performance, heavy manual annotations are expensive and inefficient. Thus it is valuable and meaningful to develop unsupervised solutions. In this paper, we propose an **U**nsupervised **S**ounding **P**ixel **L**earning (USPL) approach which enables a pixel-level sounding source localization in unsupervised paradigm. We first design a mask augmentation based multi-instance contrastive learning to realize unsupervised cross-modal coarse localization, which aligns audio-visual features to obtain coarse sounding maps. Secondly, we present an *Unsupervised Sounding Map Refinement (SMR)* module which employs the visual semantic affinity learning to explore inter-pixel relations of adjacent coordinate features. It contributes to recovering the boundary of coarse sounding maps and obtaining fine sounding maps. Finally, a *Sounding Pixel Segmentation (SPS)* module is presented to realize audio-supervised semantic segmentation. Extensive experiments are performed on the AVSBench-S4 and VGGSound datasets, exhibiting encouraging results compared with previous SOTA methods.

11:30-11:45 (West 1 Ballroom)
**Homophone Disambiguation Reveals Patterns of Context Mixing in Speech Transformers**
*Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema and Afra Alishahi*
Transformers have become a key architecture in speech processing, but our understanding of how they build up representations of acoustic and linguistic structure is limited. In this study, we address this gap by investigating how measures of 'context-mixing' developed for text models can be adapted and applied to models of spoken language. We identify a linguistic phenomenon that is ideal for such a case study: homophony in French (e.g. livre vs livres), where a speech recognition model has to attend to syntactic cues such as determiners and pronouns in order to disambiguate spoken words with identical pronunciations and transcribe them while respecting grammatical agreement. We perform a series of controlled experiments and probing analyses on Transformer-based speech models. Our findings reveal that representations in encoder-only models effectively incorporate these cues to identify the correct transcription, whereas encoders in encoder-decoder models mainly relegate the task of capturing contextual dependencies to decoder modules.

11:45-12:00 (West 1 Ballroom)
**Whispering LLaMA: A Cross-Modal Generative Error Correction Framework for Speech Recognition**
*Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero and Jesper Tegnér*
We introduce a new cross-modal fusion technique designed for generative error correction in automatic speech recognition (ASR). Our methodology leverages both acoustic information and external linguistic representations to generate accurate speech transcription contexts. This marks a step towards a fresh paradigm in generative error correction within the realm of n-best hypotheses. Unlike the existing ranking-based rescoring methods, our approach adeptly uses distinct initialization techniques and parameter-efficient algorithms to boost ASR performance derived from pre-trained speech and text models. Through evaluation across diverse ASR datasets, we assess our fusion technique, demonstrating a 37.66% improvement in word error rate (WER) relative performance compared to the n-best Oracle. To encourage future research, we have made our code and pre-trained models open source at [https://github.com/Srijith-rkr/Whispering-LLaMA](https://github.com/Srijith-rkr/Whispering-LLaMA)

12:00-12:15 (West 1 Ballroom)
**Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction**

*Cam Van Thi Nguyen, Tuan Anh Mai, Son Le The, Dang Hai Kieu and Duc-Trong Le*
Emotion recognition is a crucial task for human conversation understanding. It becomes more challenging with the notion of multimodal data, e.g., language, voice, and facial expressions. As a typical solution, the global- and the local context information are exploited to predict the emotional label for every single sentence, i.e., utterance, in the dialogue. Specifically, the global representation could be captured via modeling of cross-modal interactions at the conversation level. The local one is often inferred using the temporal information of speakers or emotional shifts, which neglects vital factors at the utterance level. Additionally, most existing approaches take fused features of multiple modalities in an unified input without leveraging modality-specific representations. Motivating from these problems, we propose the Relational Temporal Graph Neural Network with Auxiliary Cross-Modality Interaction (CORECT), an novel neural network framework that effectively captures conversation-level cross-modality interactions and utterance-level temporal dependencies with the modality-specific manner for conversation understanding. Extensive experiments demonstrate the effectiveness of CORECT via its state-of-the-art results on the IEMOCAP and CMU-MOSEI datasets for the multimodal ERC task.

12:15-12:30 (West 1 Ballroom)
**Visual Spatial Reasoning**
*Fangyu Liu, Guy Emerson and Nigel Collier*
Spatial relations are a basic part of human cognition. However, they are expressed in natural language in a variety of ways, and previous work has suggested that current vision-and-language models (VLMs) struggle to capture relational information. In this paper, we present Visual Spatial Reasoning (VSR), a dataset containing more than 10k natural text-image pairs with 66 types of spatial relations in English (such as: under, in front of, facing). While using a seemingly simple annotation format, we show how the dataset includes challenging linguistic phenomena, such as varying reference frames. We demonstrate a large gap between human and model performance: the human ceiling is above 95%, while state-of-the-art models only achieve around 70%. We observe that VLMs by-relation performances have little correlation with the number of training examples and the tested models are in general incapable of recognising relations concerning the orientations of objects.

## Theme Track: Large Language Models and the Future of NLP 2

11:00-12:30 (West 2 Ballroom)

11:00-11:15 (West 2 Ballroom)
**Lion: Adversarial Distillation of Proprietary Large Language Models**
*Yuxin Jiang, Chunkit Chan, Mingyang Chen and Wei Wang*
The practice of transferring knowledge from a sophisticated, proprietary large language model (LLM) to a compact, open-source LLM has garnered considerable attention. Previous works have focused on a unidirectional knowledge distillation way by aligning the responses of the student model with those of the teacher model to a set of instructions. Nevertheless, they overlooked the possibility of incorporating any "feedback"–identifying challenging instructions where the student model's performance falls short–to boost the student model's proficiency iteratively. To this end, we propose a novel adversarial distillation framework for a more efficient knowledge transfer. Leveraging the versatile role adaptability of LLMs, we prompt the teacher model to identify "hard" instructions and generate new "hard" instructions for the student model, creating a three-stage adversarial loop of imitation, discrimination, and generation. By applying this adversarial framework, we successfully transfer knowledge from ChatGPT to a student model (named Lion), using a mere 70k training data. Our results show that Lion-13B not only achieves comparable open-ended generation capabilities to ChatGPT but surpasses conventional state-of-the-art (SOTA) instruction-tuned models like Vicuna-13B by 55.4% in challenging zero-shot reasoning benchmarks such as BIG-Bench Hard (BBH) and 16.7% on AGIEval.

11:15-11:30 (West 2 Ballroom)
**EpiK-Eval: Evaluation for Language Models as Epistemic Models**
*Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani and Sarath Chandar*
In the age of artificial intelligence, the role of large language models (LLMs) is becoming increasingly central. Despite their growing prevalence, their capacity to consolidate knowledge from different training documents—a crucial ability in numerous applications—remains unexplored. This paper presents the first study examining the capability of LLMs to effectively combine such information within their parameter space. We introduce EpiK-Eval, a novel question-answering benchmark tailored to evaluate LLMs' proficiency in formulating a coherent and consistent knowledge representation from segmented narratives. Evaluations across various LLMs reveal significant weaknesses in this domain. We contend that these shortcomings stem from the intrinsic nature of prevailing training objectives. Consequently, we advocate for refining the approach towards knowledge consolidation, as it harbors the potential to dramatically improve their overall effectiveness and performance. The findings from this study offer insights for developing more robust and reliable LLMs. Our code and benchmark are available at https://github.com/chandar-lab/EpiK-Eval

11:30-11:45 (West 2 Ballroom)
**To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing**
*Sireesh Gururaja, Amanda Bertsch, Clara Na, David Gray Widder and Emma Strubell*
NLP is in a period of disruptive change that is impacting our methodologies, funding sources, and public perception. In this work, we seek to understand how to shape our future by better understanding our past. We study factors that shape NLP as a field, including culture, incentives, and infrastructure by conducting long-form interviews with 26 NLP researchers of varying seniority, research area, institution, and social identity. Our interviewees identify cyclical patterns in the field, as well as new shifts without historical parallel, including changes in benchmark culture and software infrastructure. We complement this discussion with quantitative analysis of citation, authorship, and language use in the ACL Anthology over time. We conclude by discussing shared visions, concerns, and hopes for the future of NLP. We hope that this study of our field's past and present can prompt informed discussion of our community's implicit norms and more deliberate action to consciously shape the future.

11:45-12:00 (West 2 Ballroom)
**Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding**
*Bram van Dijk, Tom Kouwenhoven, Marco Spruit and Max Johannes van Duijn*
Current Large Language Models (LLMs) are unparalleled in their ability to generate grammatically correct, fluent text. LLMs are appearing rapidly, and debates on LLM capacities have taken off, but reflection is lagging behind. Thus, in this position paper, we first zoom in on the debate and critically assess three points recurring in critiques of LLM capacities: i) that LLMs only parrot statistical patterns in the training data; ii) that LLMs master formal but not functional language competence; and iii) that language learning in LLMs cannot inform human language learning. Drawing on empirical and theoretical arguments, we show that these points need more nuance. Second, we outline a pragmatic perspective on the issue of 'real' understanding and intentionality in LLMs. Understanding and intentionality pertain to unobservable mental

states we attribute to other humans because they have pragmatic value: they allow us to abstract away from complex underlying mechanics and predict behaviour effectively. We reflect on the circumstances under which it would make sense for humans to similarly attribute mental states to LLMs, thereby outlining a pragmatic philosophical context for LLMs as an increasingly prominent technology in society.

12:00-12:15 (West 2 Ballroom)
**FreeAL: Towards Human-Free Active Learning in the Era of Large Language Models**
*Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen and Haobo Wang*
Collecting high-quality labeled data for model training is notoriously time-consuming and labor-intensive for various NLP tasks. While copious solutions, such as active learning for small language models (SLMs) and prevalent in-context learning in the era of large language models (LLMs), have been proposed and alleviate the labeling burden to some extent, their performances are still subject to human intervention. It is still underexplored how to reduce the annotation cost in the LLMs era. To bridge this, we revolutionize traditional active learning and propose an innovative collaborative learning framework FreeAL to interactively distill and filter the task-specific knowledge from LLMs. During collaborative training, an LLM serves as an active annotator inculcating its coarse-grained knowledge, while a downstream SLM is incurred as a student to filter out high-quality in-context samples to feedback LLM for the subsequent label refinery. Extensive experiments on eight benchmark datasets demonstrate that FreeAL largely enhances the zero-shot performances for both SLM and LLM without any human supervision.

12:15-12:30 (West 2 Ballroom)
**Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU**
*Fajri Koto, Nurul Aisyah, Haonan Li and Timothy Baldwin*
Although large language models (LLMs) are often pre-trained on large-scale multilingual texts, their reasoning abilities and real-world knowledge are mainly evaluated based on English datasets. Assessing LLM capabilities beyond English is increasingly vital but hindered due to the lack of suitable datasets. In this work, we introduce IndoMMLU, the first multi-task language understanding benchmark for Indonesian culture and languages, which consists of questions from primary school to university entrance exams in Indonesia. By employing professional teachers, we obtain 14,981 questions across 64 tasks and education levels, with 46% of the questions focusing on assessing proficiency in the Indonesian language and knowledge of nine local languages and cultures in Indonesia. Our empirical evaluations show that GPT-3.5 only manages to pass the Indonesian primary school level, with limited knowledge of local Indonesian languages and culture. Other smaller models such as BLOOMZ and Falcon perform at even lower levels.

## Industry track

11:00-12:30 (West 3 Ballroom)

---

11:00-11:15 (West 3 Ballroom)
**Self-Criticism: Aligning Large Language Models with their Understanding of Helpfulness, Honesty, and Harmlessness**
*Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu and Yuan Qi*
Recently, there has been a notable surge in the significance of large language models (LLMs) that engage in conversational-style interactions, such as ChatGPT and Claude, as they contribute significantly to the progress of artificial general intelligence (AGI). Typically, these models undergo a two-phase fine-tuning process: instruction fine-tuning (IF) and reinforcement learning from human feedback (RLHF). These methods aim to align the LLMs to be helpful, honest, and harmless (HHH). However, RLHF, which incorporates independent reward models trained on high-quality human feedback datasets, incurs high costs in terms of hardware resources and human efforts. Therefore, we explore the possibility of aligning LLMs with their own understanding of HHH through IF and in-context learning (ICL). In this study, we propose a novel framework called Self-Criticism, which allows LLMs to align themselves with HHH based on the definition they learned from a large-scale text corpus. We begin by employing IF on a given instruction set and learning HHH discrimination through few-shot ICL. Subsequently, the LLMs evaluate their own generated responses and learn to produce "better" responses based on self-judgment. Finally, the model is retrained based on the self-generated responses to distill the whole process. By analyzing our proposed method, we also find interesting connections between Self-Criticism and goal-conditioned reinforcement learning, and pseudo-labeling. Experimental results demonstrate that this method achieves nearly identical performance to RLHF in terms of both human evaluation and evaluation by other LLMs, with only a minimal alignment tax.

11:15-11:30 (West 3 Ballroom)
**PILLOW: Enhancing Efficient Instruction Fine-tuning via Prompt Matching**
*Zhenting Qi, Xiaoyu Tan, Shaojie Shi, Chao Qu, Yinghui Xu and Yuan Qi*
Instruction fine-tuning has conventionally been employed to adapt Large Language Models (LLMs) to a variety of diverse tasks. Nonetheless, this technique often necessitates substantial computational resources, making it impractical for deployment by individuals or small-scale entities. Recently, Low-Rank Adaptation (LoRA) has become a promising alternative, offering tuning capabilities with reduced resource overhead. However, attaining satisfactory performance through the fine-tuning of LoRA is a non-trivial challenge. In this paper, we propose PILLOW, which aims to improve LoRA's performance by leveraging LLM's in-context learning capability through prompt matching via reinforcement learning in resource-constrained environments. Specifically, PILLOW incorporates a matching network that selects prompts from a user-defined pool, concatenates the optimal prompts given the user instruction, and performs inference using the LoRA-fine-tuned LLMs. Compared with typical instruction fine-tuning methods, PILLOW exhibits commensurate performance on various evaluation metrics, utilizing only consumer-grade GPU resources and exhibiting a large increase in training efficiency.

11:30-11:45 (West 3 Ballroom)
**Lattice Path Edit Distance: A Romanization-aware Edit Distance for Extracting Misspelling-Correction Pairs from Japanese Search Query Logs**
*Nobuhiro Kaji*
Edit distance has been successfully used to extract training data, i.e., misspelling-correction pairs, of spelling correction models from search query logs in languages including English. However, the success does not readily apply to Japanese, where misspellings are often dissimilar to correct spellings due to the romanization-based input methods. To address this problem, we introduce lattice path edit distance, which utilizes romanization lattices to efficiently consider all possible romanized forms of input strings. Empirical experiments using Japanese search query logs demonstrated that the lattice path edit distance outperformed baseline methods including the standard edit distance combined with an existing transliterator and morphological analyzer. A training data collection pipeline that uses the lattice path edit distance has been deployed in production at our search engine for over a year.

11:45-12:00 (West 3 Ballroom)
**LLM4Vis: Explainable Visualization Recommendation using ChatGPT**

*Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim and Yong Wang*
Data visualization is a powerful tool for exploring and communicating insights in various domains. To automate visualization choice for datasets, a task known as visualization recommendation has been proposed. Various machine-learning-based approaches have been developed for this purpose, but they often require a large corpus of dataset-visualization pairs for training and lack natural explanations for their results. To address this research gap, we propose LLM4Vis, a novel ChatGPT-based prompting approach to perform visualization recommendation and return human-like explanations using very few demonstration examples. Our approach involves feature description, demonstration example selection, explanation generation, demonstration example construction, and inference steps. To obtain demonstration examples with high-quality explanations, we propose a new explanation generation bootstrapping to iteratively refine generated explanations by considering the previous generation and template-based hint. Evaluations on the VizML dataset show that LLM4Vis outperforms or performs similarly to supervised learning models like Random Forest, Decision Tree, and MLP, in both few-shot and zero-shot settings. The qualitative evaluation also shows the effectiveness of explanations generated by LLM4Vis.

12:00-12:15 (West 3 Ballroom)
**A Pretrained Language Model for Cyber Threat Intelligence**
*Youngja Park and Weiqiu You*
We present a new BERT model for the cybersecurity domain, CTI-BERT, which can improve the accuracy of cyber threat intelligence (CTI) extraction, enabling organizations to better defend against potential cyber threats. We provide detailed information about the domain corpus collection, the training methodology and its effectiveness for a variety of NLP tasks for the cybersecurity domain. The experiments show that CTI-BERT significantly outperforms several general-domain and security-domain models for these cybersecurity applications indicating that the training data and methodology have a significant impact on the model performance.

12:15-12:30 (West 3 Ballroom)
**Investigating the Role and Impact of Disfluency on Summarization**
*Varun Nathan, Ayush Kumar and Jithendra Vepa*
Contact centers handle both chat and voice calls for the same domain. As part of their workflow, it is a standard practice to summarize the conversations once they conclude. A significant distinction between chat and voice communication lies in the presence of disfluencies in voice calls, such as repetitions, restarts, and replacements. These disfluencies are generally considered noise for downstream natural language understanding (NLU) tasks. While a separate summarization model for voice calls can be trained in addition to chat specific model for the same domain, it requires manual annotations for both the channels and adds complexity arising due to maintaining two models. Therefore, it's crucial to investigate if a model trained on fluent data can handle disfluent data effectively. While previous research explored impact of disfluency on question-answering and intent detection, its influence on summarization is inadequately studied. Our experiments reveal up to 6.99-point degradation in Rouge-L score, along with reduced fluency, consistency, and relevance when a fluent-trained model handles disfluent data. Replacement disfluencies have the highest negative impact. To mitigate this, we examine Fused-Fine Tuning by training the model with a combination of fluent and disfluent data, resulting in improved performance on both public and real-life datasets. Our work highlights the significance of incorporating disfluency in training summarization models and its advantages in an industrial setting.

## Demo session 7

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**DRGCoder: Explainable Clinical Coding for the Early Prediction of Diagnostic-Related Groups**
*Daniel Hajialigol, Derek Kaknes, Tanner Barbour, Daphne Yao, Chris North, Jimeng Sun, David Liem and Xuan Wang*
Medical claim coding is the process of transforming medical records, usually presented as free texts written by clinicians, or discharge summaries, into structured codes in a classification system such as ICD-10 (International Classification of Diseases, Tenth Revision) or DRG (Diagnosis-Related Group) codes. This process is essential for medical billing and transitional care; however, manual coding is time-consuming, error-prone, and expensive. To solve these issues, we propose DRGCoder, an explainability-enhanced clinical claim coding system for the early prediction of medical severity DRGs (MS-DRGs), a classification system that categorizes patients' hospital stays into various DRG groups based on the severity of illness and mortality risk. The DRGCoder framework introduces a novel multi-task Transformer model for MS-DRG prediction, modeling both the DRG labels of the discharge summaries and the important, or salient words within he discharge summaries. We allow users to inspect DRGCoder's reasoning by visualizing the weights for each word of the input. Additionally, DRGCoder allows users to identify diseases within discharge summaries and compare across multiple discharge summaries. Our demo is available at https://huggingface.co/spaces/danielhajialigol/DRGCoder. A video demonstrating the demo can be found at https://www.youtube.com/watch?v=pcdiG6VwqIA

11:00-12:30 (East Foyer / Virtual)
**CAMRA: Copilot for AMR Annotation**
*Jon Cai, Shafiuddin Rehan Ahmed, Julia Bonn, Kristin Wright-Bettner, Martha Palmer and James H. Martin*
In this paper, we introduce CAMRA (Copilot for AMR Annotatations), a cutting-edge web-based tool designed for constructing Abstract Meaning Representation (AMR) from natural language text. CAMRA offers a novel approach to deep lexical semantics annotation such as AMR, treating AMR annotation akin to coding in programming languages. Leveraging the familiarity of programming paradigms, CAMRA encompasses all essential features of existing AMR editors, including example lookup, while going a step further by integrating Propbank roleset lookup as an autocomplete feature within the tool. Notably, CAMRA incorporates AMR parser models as coding co-pilots, greatly enhancing the efficiency and accuracy of AMR annotators.

11:00-12:30 (East Foyer / Virtual)
**Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data**
*Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao and Jiawei Han*
Chemical reactions, as a core entity in the realm of chemistry, hold crucial implications in diverse areas ranging from hands-on laboratory research to advanced computational drug design. Despite a burgeoning interest in employing NLP techniques to extract these reactions, aligning this task with the real-world requirements of chemistry practitioners remains an ongoing challenge. In this paper, we present Reaction Miner, a system specifically designed to interact with raw scientific literature, delivering precise and more informative chemical reactions. Going beyond mere extraction, Reaction Miner integrates a holistic workflow: it accepts PDF files as input, bypassing the need for pre-processing and bolstering user accessibility. Subsequently, a text segmentation module ensures that the refined text encapsulates complete chemical reactions, augmenting the accuracy of extraction. Moreover, Reaction Miner broadens the scope of existing pre-defined reaction roles, including vital attributes previously neglected, thereby offering a more comprehensive depiction of chemical reactions. Evaluations conducted by chemistry

domain users highlight the efficacy of each module in our system, demonstrating Reaction Miner as a powerful tool in this field.

11:00-12:30 (East Foyer / Virtual)
**Prompt2Model: Generating Deployable Models from Natural Language Instructions**
*Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu and Graham Neubig*
Large language models (LLMs) enable system builders today to create competent NLP systems through prompting, where they only need to describe the task in natural language and provide a few examples. However, in other ways, LLMs are a step backward from traditional special-purpose NLP models; they require extensive computational resources for deployment and can be gated behind APIs. In this paper, we propose Prompt2Model, a general-purpose method that takes a natural language task description like the prompts provided to LLMs, and uses it to train a special-purpose model that is conducive to deployment. This is done through a multi-step process of retrieval of existing datasets and pretrained models, dataset generation using LLMs, and supervised fine-tuning on these retrieved and generated datasets. Over three tasks, we demonstrate that given the same few-shot prompt as input, Prompt2Model trains models that outperform the results of a strong LLM, gpt-3.5-turbo, by an average of 20% while being up to 700 times smaller. We also show that this data can be used to obtain reliable performance estimates of model performance, enabling model developers to assess model reliability before deployment. Prompt2Model is available open-source at https://github.com/neulab/prompt2model. Our demo video is posted at youtu.be/LYYQ_EhGd-Q.

11:00-12:30 (East Foyer / Virtual)
**NewsSense: Reference-free Verification via Cross-document Comparison**
*Jeremiah Milbauer, Ziqi Ding, Zhijin Wu and Tongshuang Wu*
We present NewsSense, a novel sensemaking tool and reading interface designed to collect and integrate information from multiple news articles on a central topic. NewsSense provides "reference-free verification," augmenting a central grounding article of the user's choice by: (1) linking to related articles from different sources; and (2) providing inline highlights on how specific claims are either supported or contradicted by information from other articles. Using NewsSense, users can seamlessly digest and cross-check multiple information sources without disturbing their natural reading flow. Our pilot study shows that NewsSense has the potential to help users identify key information, verify the credibility of news articles, explore different perspectives, and understand what content is supported, contradicted, or missing.

11:00-12:30 (East Foyer / Virtual)
**NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails**
*Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien and Jonathan Cohen*
NeMo Guardrails is an open-source toolkit for easily adding programmable guardrails to LLM-based conversational systems. Guardrails (or rails for short) are a specific way of controlling the output of an LLM, such as not talking about topics considered harmful, following a predefined dialogue path, using a particular language style, and more. There are several mechanisms that allow LLM providers and developers to add guardrails that are embedded into a specific model at training, e.g. using model alignment. Using a runtime inspired from dialogue management, NeMo Guardrails provides a different approach by allowing developers to add programmable rails to LLM applications - these are user-defined, independent of the underlying LLM, and interpretable. Our initial results show that the proposed approach can be used with several LLM providers to develop controllable and safe LLM applications using programmable rails.

11:00-12:30 (East Foyer / Virtual)
**CocoSciSum: A Scientific Summarization Toolkit with Compositional Controllability**
*Yixi Ding, Yanxia Qin, Qian Liu and Min-Yen Kan*
We present a novel toolkit for controlled summarization of scientific documents, designed for the specific needs of the scientific community. Our system generates summaries based on user preferences, adjusting key attributes specifically of length and keyword inclusion. A distinguishing feature is its ability to manage multiple attributes concurrently, demonstrating Compositional Controllability for Scientific Summarization (CocoSciSum). Benchmarked against the strong Flan-T5 baseline, CocoSciSum exhibits superior performance on both the quality of summaries generated and the control over single and multiple attributes. Moreover, CocoSciSum is a user-centric toolkit, supporting user preferences expressed in natural language instructions, and accommodating diverse input document formats. CocoSciSum is available on GitHub (https://github.com/WING-NUS/SciAssist/tree/CocoSciSum) with an introduction video (https://youtu.be/YC1YDeEjAbQ).

## Industry 7

11:00-12:30 (East Foyer / Virtual)

---

11:00-12:30 (East Foyer / Virtual)
**A Comparative Analysis of Task-Agnostic Distillation Methods for Compressing Transformer Language Models**
*Takuma Udagawa, Aashka Trivedi, Michele Merler and Bishwaranjan Bhattacharjee*
Large language models have become a vital component in modern NLP, achieving state of the art performance in a variety of tasks. However, they are often inefficient for real-world deployment due to their expensive inference costs. Knowledge distillation is a promising technique to improve their efficiency while retaining most of their effectiveness. In this paper, we reproduce, compare and analyze several representative methods for task-agnostic (general-purpose) distillation of Transformer language models. Our target of study includes Output Distribution (OD) transfer, Hidden State (HS) transfer with various layer mapping strategies, and Multi-Head Attention (MHA) transfer based on MiniLMv2. Through our extensive experiments, we study the effectiveness of each method for various student architectures in both monolingual (English) and multilingual settings. Overall, we show that MHA transfer based on MiniLMv2 is generally the best option for distillation and explain the potential reasons behind its success. Moreover, we show that HS transfer remains as a competitive baseline, especially under a sophisticated layer mapping strategy, while OD transfer consistently lags behind other approaches. Findings from this study helped us deploy efficient yet effective student models for latency-critical applications.

11:00-12:30 (East Foyer / Virtual)
**AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications**
*Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo and Preethi Lahoti*
Adversarially testing large language models (LLMs) is crucial for their safe and responsible deployment in practice. We introduce an AI-assisted approach for automated generation of adversarial evaluation datasets to test the safety of LLM generations on new downstream applications. We call it AART AI-assisted Red-Teaming - an automated alternative to current manual red-teaming efforts. AART offers a data generation and augmentation pipeline of reusable and customizable recipes that reduce significantly human effort and enable integration of adversarial testing earlier in new product development. AART generates evaluation datasets with high diversity of content characteristics critical for effective adversarial testing (e.g. sensitive and harmful concepts, specific to a wide range of cultural and geographic regions and application scenarios). The data generation is steered by AI-assisted recipes to define, scope and prioritize diversity within a new application context. This feeds into a structured LLM-generation process that scales up evaluation priorities. This provides transparency of developers

evaluation intentions and enables quick adaptation to new use cases and newly discovered model weaknesses. Compared to some of the state-of-the-art tools AART shows promising results in terms of concept coverage and data quality.

11:00-12:30 (East Foyer / Virtual)
**AdaBERT-CTC: Leveraging BERT-CTC for Text-Only Domain Adaptation in ASR**
*Tyler Vuong, Karel Mundnich, Dhanush Bekal, Veera Raghavendra Elluru, Srikanth Ronanki and Sravan Bodapati*
End-to-end (E2E) automatic speech recognition (ASR) models are becoming increasingly popular in commercial applications, such as virtual assistants, closed captioning, and dictation systems. The accuracy of the ASR is crucial to their success. However, E2E models still struggle to recognize out-of-domain words such as proper nouns and domain-specific terms. In this paper we introduce AdaBERT-CTC, a domain adaptation technique that relies solely on textual data. Our method allows for text-only adaptation by fine-tuning a pre-trained self-supervised text encoder model. Additionally, we show that our method can be made parameter-efficient by adding bottleneck adapters to the pre-trained model. This allows for adaptation with less than a 5% increase in parameters and minimal computational overhead during inference. We demonstrate that our approach outperforms the base BERT-CTC model by up to 14% relative word error rate improvement on several out-of-domain, publicly available datasets.

11:00-12:30 (East Foyer / Virtual)
**AdapterDistillation: Non-Destructive Task Composition with Knowledge Distillation**
*Junjie Wang, Yicheng Chen, Wangshu Zhang, Sen Hu, Teng Xu and Jing Zheng*
Leveraging knowledge from multiple tasks through introducing a small number of task specific parameters into each transformer layer, also known as adapters, receives much attention recently. However, adding an extra fusion layer to implement knowledge composition not only increases the inference time but also is non-scalable for some applications. To avoid these issues, we propose a two-stage knowledge distillation algorithm called AdapterDistillation. In the first stage, we extract task specific knowledge by using local data to train a student adapter. In the second stage, we distill the knowledge from the existing teacher adapters into the student adapter to help its inference. Extensive experiments on frequently asked question retrieval in task-oriented dialog systems validate the efficiency of AdapterDistillation. We show that AdapterDistillation outperforms existing algorithms in terms of accuracy, resource consumption and inference time.

11:00-12:30 (East Foyer / Virtual)
**Adaptive Hyper-parameter Learning for Deep Semantic Retrieval**
*Mingming Li, Chunyuan Yuan, Huimu Wang, Peng Wang, Jingwei Zhuo, Binbin Wang, Lin Liu and Sulong Xu*
Deep semantic retrieval has achieved remarkable success in online E-commerce applications. The majority of methods aim to distinguish positive items and negative items for each query by utilizing margin loss or softmax loss. Despite their decent performance, these methods are highly sensitive to hyper-parameters, i.e., margin and temperature $\tau$, which measure the similarity of negative pairs and affect the distribution of items in metric space. How to design and choose adaptively parameters for different pairs is still an open challenge. Recently several methods have attempted to alleviate the above problem by learning each parameter through trainable/statistical methods in the recommendation. We argue that those are not suitable for retrieval scenarios, due to the agnosticism and diversity of the queries. To fully overcome this limitation, we propose a novel adaptive metric learning method that designs a simple and universal hyper-parameter-free learning method to improve the performance of retrieval. Specifically, we first propose a method that adaptive obtains the hyper-parameters by relying on the batch similarity without fixed or extra-trainable hyper-parameters. Subsequently, we adopt a symmetric metric learning method to mitigate model collapse issues. Furthermore, the proposed method is general and sheds a highlight on other fields. Extensive experiments demonstrate our method significantly outperforms previous methods on a real-world dataset, highlighting the superiority and effectiveness of our method. This method has been successfully deployed on an online E-commerce search platform and brought substantial economic benefits.

11:00-12:30 (East Foyer / Virtual)
**An Auxiliary Task Boosted Multi-task Learning Method for Service Account Retrieval with Limited Human Annotation**
*Yuanzhou Yao, Zhao Zhang, Kaijia Yang, Huasheng Liang, Qiang Yan and Yongjun Xu*
Service accounts, including organizations' official accounts and mini-programs, provide various convenient services for users, and have become crucial components of a number of applications. Therefore, retrieving service accounts quickly and accurately is vital. However, this task suffers from the problem of limited human annotation, i.e., manually assessing account functionality and assigning ratings based on user experience is both labor-intensive and time-consuming. To this end, this paper proposes a novel approach, the Auxiliary task Boosted Multi-Task Learning method (AuxBoost-MTL). Specifically, the proposed method introduces multiple auxiliary tasks, which is able to utilized the log data from our application as supervision, and enhance the performance of the main task, service account retrieval. Furthermore, we introduce an Adaptive Hierarchical Fusion Module (AHF module) into our approach. This module is designed to adaptively perform hierarchical fusion of embeddings from auxiliary tasks into the main task, thereby enhancing the model efficacy. Experiments on two real-world industrial datasets demonstrate the effectiveness of our proposed approach.

11:00-12:30 (East Foyer / Virtual)
**Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks**
*Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu and Sameena Shah*
The most recent large language models (LLMs) such as ChatGPT and GPT-4 have shown exceptional capabilities of generalist models, achieving state-of-the-art performance on a wide range of NLP tasks with little or no adaptation. How effective are such models in the finance domain? Understanding this basic question would have a significant impact on many downstream financial analytical tasks. In this paper, we conduct empirical studies and provide experimental evidences of their performance on a wide variety of financial text analytical problems, using eight benchmark datasets from five categories of tasks. We report both the strengths and limitations of the current models by comparing them to the state-of-the-art fine-tuned approaches and the recently released domain-specific pretrained models. We hope our study can help to understand the capability of the existing models in the financial domain and facilitate further improvements.

11:00-12:30 (East Foyer / Virtual)
**Batch Prompting: Efficient Inference with Large Language Model APIs**
*Zhoujun Cheng, Jungo Kasai and Tao Yu*
Performing inference on large volumes of samples with large language models (LLMs) can be computationally and financially costly in industry and real-world use. We propose batch prompting, a simple yet effective prompting approach that enables the LLM to run inference in batches, instead of one sample at a time. Our method reduces both token and time costs while retaining downstream performance. We theoretically demonstrate that under a few-shot in-context learning setting, the inference costs decrease almost inverse linearly with the number of samples in each batch. We extensively validate the effectiveness of batch prompting on ten datasets across commonsense QA, arithmetic reasoning, and NLI/NLU: batch prompting significantly (up to $5\times$ with six samples in batch) reduces the LLM (Codex) inference token and time costs while achieving better or comparable performance. For state-of-the-art Chat-based LLMs, e.g., GPT-3.5 and GPT-4, we show the benefits of batch prompting also hold. Further analysis shows that the number of samples in each batch and the complexity of tasks affect its performance. Moreover, batch prompting can be applied across different reasoning methods using LLMs. Our code is released at the site https://github.com/xlang-ai/batch-prompting.

11:00-12:30 (East Foyer / Virtual)
**BeautifulPrompt: Towards Automatic Prompt Engineering for Text-to-Image Synthesis**
*Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu and Jun Huang*
Recently, diffusion-based deep generative models (e.g., Stable Diffusion) have shown impressive results in text-to-image synthesis. However, current text-to-image models often require multiple passes of prompt engineering by humans in order to produce satisfactory results for real-world applications. We propose BeautifulPrompt, a deep generative model to produce high-quality prompts from very simple raw descriptions, which enables diffusion-based models to generate more beautiful images. In our work, we first fine-tuned the BeautifulPrompt model over low-quality and high-quality collecting prompt pairs. Then, to ensure that our generated prompts can generate more beautiful images, we further propose a Reinforcement Learning with Visual AI Feedback technique to fine-tune our model to maximize the reward values of the generated prompts, where the reward values are calculated based on the PickScore and the Aesthetic Scores. Our results demonstrate that learning from visual AI feedback promises the potential to improve the quality of generated prompts and images significantly. We further showcase the integration of BeautifulPrompt to a cloud-native AI platform to provide better text-to-image generation service in the cloud.

11:00-12:30 (East Foyer / Virtual)
**Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective**
*Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen and Shashi Bhushan TN*
This paper studies how to effectively build meeting summarization systems for real-world usage using large language models (LLMs). For this purpose, we conduct an extensive evaluation and comparison of various closed-source and open-source LLMs, namely, GPT-4, GPT-3.5, PaLM-2, and LLaMA-2. Our findings reveal that most closed-source LLMs are generally better in terms of performance. However, much smaller open-source models like LLaMA-2 (7B and 13B) could still achieve performance comparable to the large closed-source models even in zero-shot scenarios. Considering the privacy concerns of closed-source models for only being accessible via API, alongside the high cost associated with using fine-tuned versions of the closed-source models, the opensource models that can achieve competitive performance are more advantageous for industrial use. Balancing performance with associated costs and privacy concerns, the LLaMA-2-7B model looks more promising for industrial usage. In sum, this paper offers practical insights on using LLMs for real-world business meeting summarization, shedding light on the trade-offs between performance and cost.

11:00-12:30 (East Foyer / Virtual)
**CarExpert: Leveraging Large Language Models for In-Car Conversational Question Answering**
*Md Rashad Al Hasan Rony, Christian Suess, Sinchana Ramakanth Bhat, Viju Sudhi, Julia Schneider, Maximilian Vogel, Roman Teucher, Ken E. Friedl and Soumya Sahoo*
Large language models (LLMs) have demonstrated remarkable performance by following natural language instructions without fine-tuning them on domain-specific tasks and data. However, leveraging LLMs for domain-specific question answering suffers from severe limitations. The generated answer tends to hallucinate due to the training data collection time (when using off-the-shelf), complex user utterance and wrong retrieval (in retrieval-augmented generation). Furthermore, due to the lack of awareness about the domain and expected output, such LLMs may generate unexpected and unsafe answers that are not tailored to the target domain. In this paper, we propose CarExpert, an in-car retrieval-augmented conversational question-answering system leveraging LLMs for different tasks. Specifically, CarExpert employs LLMs to control the input, provide domain-specific documents to the extractive and generative answering components, and controls the output to ensure safe and domain-specific answers. A comprehensive empirical evaluation exhibits that CarExpert outperforms state-of-the-art LLMs in generating natural, safe and car-specific answers.

11:00-12:30 (East Foyer / Virtual)
**CL-QR: Cross-Lingual Enhanced Query Reformulation for Multi-lingual Conversational AI Agents**
*Zhongkai Sun, Zhengyang Zhao, Sixing Lu, Chengyuan Ma, Xiaohu Liu, Xing Fan, Wei Shen and Chenlei Guo*
The growing popularity of conversational AI agents such as Alexa, Google Assistant, and Siri rely on accurate spoken language comprehension. The query reformulation (QR) method, which reformulates defective user queries, has been broadly adopted to mitigate the challenges posed by understanding user's intent from imperfect spoken recognition result. However, due to the scarcity of non-English QR labels, providing high-quality QR for non-English users still remains a challenge. This work proposes a novel cross-lingual QR framework, CL-QR, to leverage the abundant reformulation resources in English to improve non-English QR performance. The proposed work also proposes a Module-wise Mutually-supervised Feedback learning (MMF) algorithm to enable the continually self-improving of the CL-QR, which alleviates the lack of cross-lingual QR training data and enhances the delivery of high-quality reformulations learned in English for multilingual queries. Both offline evaluation and online A/B testing demonstrates the effectiveness of the proposed method.

11:00-12:30 (East Foyer / Virtual)
**Compute-Efficient Churn Reduction for Conversational Agents**
*Christopher Hidey and Sarthak Sarthak*
Model churn occurs when re-training a model yields different predictions despite using the same data and hyper-parameters. Churn reduction is crucial for industry conversational systems where users expect consistent results for the same queries. In this setting, compute resources are often limited due to latency requirements during serving and overall time constraints during re-training. To address this issue, we propose a compute-efficient method that mitigates churn without requiring extra resources for training or inference. Our approach involves a lightweight data pre-processing step that pairs semantic parses based on their "function call signature" and encourages similarity through an additional loss based on Jensen-Shannon Divergence. We validate the effectiveness of our method in three scenarios: academic (+3.93 percent improvement on average in a churn reduction metric), simulated noisy data (+8.09), and industry (+5.28) settings.

11:00-12:30 (East Foyer / Virtual)
**Conversing with databases: Practical Natural Language Querying**
*Denis Kochedykov, Fenglin Yin and Sreevidya Khatravath*
In this work, we designed, developed and released in production DataQue – a hybrid NLQ (Natural Language Querying) system for conversational DB querying. We address multiple practical problems that are not accounted for in public Text-to-SQL solutions – numerous complex implied conditions in user questions, jargon and abbreviations, custom calculations, non-SQL operations, a need to inject all those into pipeline fast and to have guaranteed parsing results for demanding users, cold-start problem. The DataQue processing pipeline for Text-to-SQL translation consists of 10-15 model-based and rule-based components that allows to tightly control the processing.

11:00-12:30 (East Foyer / Virtual)
**Coordinated Replay Sample Selection for Continual Federated Learning**
*Jack Good, Jimit Majmudar, Christophe Dupuy, Jixuan Wang, Charith Peris, Clement Chung, Richard Zemel and Rahul Gupta*
Continual Federated Learning (CFL) combines Federated Learning (FL), the decentralized learning of a central model on a number of client devices that may not communicate their data, and Continual Learning (CL), the learning of a model from a continual stream of data without keeping the entire history. In CL, the main challenge is forgetting what was learned from past data. While replay-based algorithms that keep

a small pool of past training data are effective to reduce forgetting, only simple replay sample selection strategies have been applied to CFL in prior work, and no previous work has explored coordination among clients for better sample selection. To bridge this gap, we adapt a replay sample selection objective based on loss gradient diversity to CFL and propose a new relaxation-based selection of samples to optimize the objective. Next, we propose a practical algorithm to coordinate gradient-based replay sample selection across clients without communicating private data. We benchmark our coordinated and uncoordinated replay sample selection algorithms against random sampling-based baselines with language models trained on a large scale de-identified real-world text dataset. We show that gradient-based sample selection methods both boost performance and reduce forgetting compared to random sampling methods, with our coordination method showing gains early in the low replay size regime (when the budget for storing past data is small).

11:00-12:30 (East Foyer / Virtual)
**Creator Context for Tweet Recommendation**
*Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, Marc Najork, Matt Colen, Sergey Levi, Vladimir Ofitserov and Tanvir Amin*
When discussing a tweet, people usually not only refer to the content it delivers, but also to the person behind the tweet. In other words, grounding the interpretation of the tweet in the context of its creator plays an important role in deciphering the true intent and the importance of the tweet. In this paper, we attempt to answer the question of how creator context should be used to advance tweet understanding. Specifically, we investigate the usefulness of different types of creator context, and examine different model structures for incorporating creator context in tweet modeling. We evaluate our tweet understanding models on a practical use case – recommending relevant tweets to news articles. This use case already exists in popular news apps, and can also serve as a useful assistive tool for journalists. We discover that creator context is essential for tweet understanding, and can improve application metrics by a large margin. However, we also observe that not all creator contexts are equal. Creator context can be time sensitive and noisy. Careful creator context selection and deliberate model structure design play an important role in creator context effectiveness.

11:00-12:30 (East Foyer / Virtual)
**Deep Metric Learning to Hierarchically Rank - An Application in Product Retrieval**
*Kee Kiat Koo, Ashutosh Joshi, Nishaanth Reddy, Karim Bouyarmane, Ismail Tutar, Vaclav Petricek and Changhe Yuan*
Most e-commerce search engines use customer behavior signals to augment lexical matching and improve search relevance. Many e-commerce companies like Amazon, Alibaba, Ebay etc. operate in multiple countries with country specific stores. However, customer behavior data is sparse in newer stores. To compensate for sparsity of behavioral data in low traffic stores, search engines often use cross-listed products in some form. However, cross-listing across stores is not uniform and in many cases itself sparse. In this paper, we develop a model to identify duplicate and near-duplicate products across stores. Such a model can be used to unify product catalogs worldwide, improve product meta-data or as in our case, use near-duplicate products across multiple to improve search relevance. To capture the product similarity hierarchy, we develop an approach that integrates retrieval and ranking tasks across multiple languages in a single step based on a novel Hierarchical Ranked Multi Similarity (HRMS) Loss that combines Multi-Similarity (MS) loss and Hierarchical Triplet Loss to learn a hierarchical metric space. Our method outperforms strong baselines in terms of catalog coverage and precision of the mappings. We also show via online A/B tests that the product mappings found by our method are successful at improving search quality in low traffic stores, measured in rate of searches with at least one click, significantly by 0.8% and improving cold start product engagement measured as new product clicks significantly by 1.72% in established stores.

11:00-12:30 (East Foyer / Virtual)
**DELPHI: Data for Evaluating LLMs' Performance in Handling Controversial Issues**
*David Sun, Artem Abzaliev, Hadas Kotek, Christopher Klein, Zidi Xiu and Jason D Williams*
Controversy is a reflection of our zeitgeist, and an important aspect to any discourse. The rise of large language models (LLMs) as conversational systems has increased public reliance on these systems for answers to their various questions. Consequently, it is crucial to systematically examine how these models respond to questions that pertaining to ongoing debates. However, few such datasets exist in providing human-annotated labels reflecting the contemporary discussions. To foster research in this area, we propose a novel construction of a controversial questions dataset, expanding upon the publicly released Quora Question Pairs Dataset. This dataset presents challenges concerning knowledge recency, safety, fairness, and bias. We evaluate different LLMs using a subset of this dataset, illuminating how they handle controversial issues and the stances they adopt. This research ultimately contributes to our understanding of LLMs' interaction with controversial issues, paving the way for improvements in their comprehension and handling of complex societal debates.

11:00-12:30 (East Foyer / Virtual)
**DocumentNet: Bridging the Data Gap in Document Pre-training**
*Lijun Yu, Jin Miao, Xiaoyu Sun, Jiayi Chen, Alexander Hauptmann, Hanjun Dai and Wei Wei*
Document understanding tasks, in particular, Visually-rich Document Entity Retrieval (VDER), have gained significant attention in recent years thanks to their broad applications in enterprise AI. However, publicly available data have been scarce for these tasks due to strict privacy constraints and high annotation costs. To make things worse, the non-overlapping entity spaces from different datasets hinder the knowledge transfer between document types. In this paper, we propose a method to collect massive-scale and weakly labeled data from the web to benefit the training of VDER models. The collected dataset, named DocumentNet, does not depend on specific document types or entity sets, making it universally applicable to all VDER tasks. The current DocumentNet consists of 30M documents spanning nearly 400 document types organized in a four-level ontology. Experiments on a set of broadly adopted VDER tasks show significant improvements when DocumentNet is incorporated into the pre-training for both classic and few-shot learning settings. With the recent emergence of large language models (LLMs), DocumentNet provides a large data source to extend their multimodal capabilities for VDER.

11:00-12:30 (East Foyer / Virtual)
**Does Named Entity Recognition Truly Not Scale Up to Real-world Product Attribute Extraction?**
*Wei-Te Chen, Keiji Shinzato, Naoki Yoshinaga and Yandi Xia*
The key challenge in the attribute-value extraction (AVE) task from e-commerce sites is the scalability to diverse attributes for a large number of products in real-world e-commerce sites. To make AVE scalable to diverse attributes, recent researchers adopted a question-answering (QA)-based approach that additionally inputs the target attribute as a query to extract its values, and confirmed its advantage over a classical approach based on named-entity recognition (NER) on real-word e-commerce datasets. In this study, we argue the scalability of the NER-based approach compared to the QA-based approach, since researchers have compared BERT-based QA-based models to only a weak BiLSTM-based NER baseline trained from scratch in terms of only accuracy on datasets designed to evaluate the QA-based approach. Experimental results using a publicly available real-word dataset revealed that, under a fair setting, BERT-based NER models rival BERT-based QA models in terms of the accuracy, and their inference is faster than the QA model that processes the same product text several times to handle multiple target attributes.

11:00-12:30 (East Foyer / Virtual)
**DUBLIN: Visual Document Understanding By Language-Image Network**

*Kriti Aggarwal, Aditi Khandelwal, Kumar Tanmay, Owais Khan Mohammed, Qiang Liu, Monojit Choudhury, Hardik Hansrajbhai Chauhan, Subhojit Som, Vishrav Chaudhary and Saurabh Tiwary*
In this paper, we present DUBLIN, a pixel-based model for visual document understanding that does not rely on OCR. DUBLIN can process both images and texts in documents just by the pixels and handle diverse document types and tasks. DUBLIN is pretrained on a large corpus of document images with novel tasks that enhance its visual and linguistic abilities. We evaluate DUBLIN on various benchmarks and show that it achieves state-of-the-art performance on extractive tasks such as DocVQA, InfoVQA, AI2D, OCR-VQA, RefExp, and CORD, as well as strong performance on abstraction datasets such as VisualMRC and text captioning. Our model demonstrates the potential of OCR-free document processing and opens new avenues for applications and research.

11:00-12:30 (East Foyer / Virtual)
**E2E Spoken Entity Extraction for Virtual Agents**
*Karan Singla, Yeon-Jun Kim and Srinivas Bangalore*
In human-computer conversations, extracting entities such as names, street addresses and email addresses from speech is a challenging task. In this paper, we study the impact of fine-tuning pre-trained speech encoders on extracting spoken entities in human-readable form directly from speech without the need for text transcription. We illustrate that such a direct approach optimizes the encoder to transcribe only the entity relevant portions of speech ignoring the superfluous portions such as carrier phrases, or spell name entities. In the context of dialog from an enterprise virtual agent, we demonstrate that the 1-step approach outperforms the typical 2-step approach which first generates lexical transcriptions followed by text-based entity extraction for identifying spoken entities.

11:00-12:30 (East Foyer / Virtual)
**EELBERT: Tiny Models through Dynamic Embeddings**
*Gabrielle Cohn, Rishika Agarwal, Deepanshu Gupta and Siddharth Patwardhan*
We introduce EELBERT, an approach for compression of transformer-based models (e.g., BERT), with minimal impact on the accuracy of downstream tasks. This is achieved by replacing the input embedding layer of the model with dynamic, i.e. on-the-fly, embedding computations. Since the input embedding layer occupies a large portion of the model size, especially for the smaller BERT variants, replacing this layer with an embedding computation function helps us reduce the model size significantly. Empirical evaluation on the GLUE benchmark shows that our BERT variants (EELBERT) suffer minimal regression compared to the traditional BERT models. Through this approach, we are able to develop our smallest model UNO-EELBERT, which achieves a GLUE score within 4% of fully trained BERT-tiny, while being 15x smaller (1.2 MB) in size.

11:00-12:30 (East Foyer / Virtual)
**Efficient Transformer Knowledge Distillation: A Performance Review**
*Nathan Brown, Ashton Williamson, Tahj Anderson and Logan Lawrence*
As pretrained transformer language models continue to achieve state-of-the-art performance, the Natural Language Processing community has pushed for advances in model compression and efficient attention mechanisms to address high computational requirements and limited input sequence length. Despite these separate efforts, no investigation has been done into the intersection of these two fields. In this work, we provide an evaluation of model compression via knowledge distillation on efficient attention transformers. We provide cost-performance trade-offs for the compression of state-of-the-art efficient attention architectures and the gains made in performance in comparison to their full attention counterparts. Furthermore, we introduce a new long-context Named Entity Recognition dataset, GONERD, to train and test the performance of NER models on long sequences. We find that distilled efficient attention transformers can preserve a significant amount of original model performance, preserving up to **98.6%** across short-context tasks (GLUE, SQUAD, CoNLL-2003), up to **94.6%** across long-context Question-and-Answering tasks (HotpotQA, TriviaQA), and up to **98.8%** on long-context Named Entity Recognition (GONERD), while decreasing inference times by up to **57.8%**. We find that, for most models on most tasks, performing knowledge distillation is an effective method to yield high-performing efficient attention models with low costs.

11:00-12:30 (East Foyer / Virtual)
**Empower Large Language Model to Perform Better on Industrial Domain-Specific Question Answering**
*Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan and Dongmei Zhang*
Large Language Model (LLM) has gained popularity and achieved remarkable results in open-domain tasks, but its performance in real industrial domain-specific scenarios is average due to its lack of specific domain knowledge. This issue has attracted widespread attention, but there are few relevant benchmarks available. In this paper, we provide a benchmark Question Answering (QA) dataset named MSQA, centered around Microsoft products and IT technical problems encountered by customers. This dataset contains industry cloud-specific QA knowledge, an area not extensively covered in general LLMs, making it well-suited for evaluating methods aiming to enhance LLMs' domain-specific capabilities. In addition, we propose a new model interaction paradigm that can empower LLM to achieve better performance on domain-specific tasks where it is not proficient. Extensive experiments demonstrate that the approach following our method outperforms the commonly used LLM with retrieval methods. We make our source code and sample data available at: https://aka.ms/Microsoft_QA.

11:00-12:30 (East Foyer / Virtual)
**Enhancing Extreme Multi-Label Text Classification: Addressing Challenges in Model, Data, and Evaluation**
*Dan Li, Zi Long Zhu, Janneke van de Loo, Agnes Masip Gomez, Vikrant Yadav, Georgios Tsatsaronis and Zubair Afzal*
Extreme multi-label text classification is a prevalent task in industry, but it frequently encounters challenges in terms of machine learning perspectives, including model limitations, data scarcity, and time-consuming evaluation. This paper aims to mitigate these issues by introducing novel approaches. Firstly, we propose a label ranking model as an alternative to the conventional SciBERT-based classification model, enabling efficient handling of large-scale labels and accommodating new labels. Secondly, we present an active learning-based pipeline that addresses the data scarcity of new labels during the update of a classification system. Finally, we introduce ChatGPT to assist with model evaluation. Our experiments demonstrate the effectiveness of these techniques in enhancing the extreme multi-label text classification task.

11:00-12:30 (East Foyer / Virtual)
**Enhancing Language Model with Unit Test Techniques for Efficient Regular Expression Generation**
*Chenhui Mao, Xiexiong Lin, Xin Jin and Xin Zhang*
Recent research has investigated the use of generative language models to produce regular expressions with semantic-based approaches. However, these approaches have shown shortcomings in practical applications, particularly in terms of functional correctness, which refers to the ability to reproduce the intended function inputs by the user. To address this issue, we present a novel method called Unit-Test Driven Reinforcement Learning (UTD-RL). Our approach differs from previous methods by taking into account the crucial aspect of functional correctness and transforming it into a differentiable gradient feedback using policy gradient techniques. In which functional correctness can be evaluated through Unit Tests, a testing method that ensures regular expressions meets its design and performs as intended. Experiments conducted on three public datasets demonstrate the effectiveness of the proposed method in generating regular expressions. This method has been employed in a regulatory scenario where regular expressions can be utilized to ensure that all online content is free from non-compliant elements, thereby significantly reducing the workload of relevant personnel.

11:00-12:30 (East Foyer / Virtual)
**Gatekeeper to save COGS and improve efficiency of Text Prediction**
*Nidhi Tiwari, Sneha Kola, Milos Milunovic, Si-qing Chen and Marjan Slavkovski*
The text prediction (TP) workflow calls a Large Language Model (LLM), almost, after every character to get subsequent sequence of characters, till user accepts a suggestion. The confidence score of the prediction is commonly used for filtering the results to ensure that only correct predictions are shown to user. As LLMs require massive amounts of computation and storage, such an approach incurs network and high execution cost. So, we propose a Model gatekeeper (GK) to stop the LLM calls that will result in incorrect predictions at client application level itself. This way a GK can save cost of model inference and improve user experience by not showing the incorrect predictions. We demonstrate that use of a model gatekeeper saved approx 46.6% of COGS for TP, at the cost of approx 4.5% loss in character saving. Use of GK also improved the efficiency (suggestion rate) of TP model by 73%.

11:00-12:30 (East Foyer / Virtual)
**Generative Models for Product Attribute Extraction**
*Ansel Blume, Nasser Zalmout, Heng Ji and Xian Li*
Product attribute extraction is an emerging field in information extraction and e-commerce, with applications including knowledge base construction, product recommendation, and enhancing customer experiences. In this work, we explore the use of generative models for product attribute extraction. We analyze their utility with hard and soft prompting methods, and demonstrate their ability to generate implicit attribute values, which state-of-the-art sequence tagging models are unable to extract. We perform a wide range of experiments on Amazon and MAVE product attribute datasets, and are the first to present results on multilingual attribute extraction. Our results show that generative models can outperform state- of-the-art tagging models for explicit product attribute extraction while having greater data efficiency, that they have the unique ability to perform implicit attribute extraction, and that in certain settings large language models can perform competitively with finetuned models with as little as two in-context examples.

11:00-12:30 (East Foyer / Virtual)
**Gold Standard Bangla OCR Dataset: An In-Depth Look at Data Preprocessing and Annotation Processes**
*Hasmot Ali, AKM Shahariar Azad Rabby, Md Majedul Islam, A.K.M Mahamud, Nazmul Hasan and Fuad Rahman*
This research paper focuses on developing an improved Bangla Optical Character Recognition (OCR) system, addressing the challenges posed by the complexity of Bangla text structure, diverse handwriting styles, and the scarcity of comprehensive datasets. Leveraging recent advancements in Deep Learning and OCR techniques, we anticipate a significant enhancement in the performance of Bangla OCR by utilizing a large and diverse collection of labeled Bangla text image datasets. This study introduces the most extensive gold standard corpus for Bangla characters and words, comprising over 4 million human-annotated images. Our dataset encompasses various document types, such as Computer Compose, Letterpress, Typewriters, Outdoor Banner-Poster, and Handwritten documents, gathered from diverse sources. The entire corpus has undergone meticulous human annotation, employing a controlled annotation procedure consisting of three-step annotation and one-step validation, ensuring adherence to gold standard criteria. This paper provides a comprehensive overview of the complete data collection procedure. The ICT Division, Government of the People's Republic of Bangladesh, will make the dataset publicly available, facilitating further research and development in Bangla OCR and related domains.

11:00-12:30 (East Foyer / Virtual)
**Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding**
*Zheng Chen, Ziyan Jiang, Fan Yang, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu and Aram Galstyan*
A Personalized Query Rewriting system strives to minimize defective queries to ensure robust conversational functionality by considering individual user behavior and preferences. It's designed as a search-based system, maintaining a user index of past successful interactions with the conversational AI. However, this method faces challenges with unseen interactions, which refers to novel user interactions not covered by the user's historical index. This paper introduces our Collaborative Query Rewriting approach, which utilizes underlying topological information to assist in rewriting defective queries arising from unseen user interactions. This approach begins by constructing a "User Feedback Interaction Graph" (FIG) using historical user-entity interactions. Subsequently, we traverse through the graph edges to establish an enhanced user index, referred to as the "collaborative user index". This paper then further explores the use of Large Language Models (LLMs) in conjunction with graph traversal, leading to a significant increase in index coverage for unseen interactions. The effectiveness of our proposed approach has been proven through experiments on a large-scale real-world dataset and online A/B experiments.

11:00-12:30 (East Foyer / Virtual)
**Harnessing LLMs for Temporal Data - A Study on Explainable Financial Time Series Forecasting**
*Xinli Yu, Zheng Chen and Yanbin Lu*
Applying machine learning to financial time series has been an active area of industrial research enabling innovation in market insights, risk management, strategic decision-making, and policy formation. This paper explores the novel use of Large Language Models (LLMs) for explainable financial time series forecasting, addressing challenges in cross-sequence reasoning, multi-modal data integration, and result interpretation that are inherent in traditional approaches. Focusing on NASDAQ-100 stocks, we utilize public historical stock data, company metadata, and economic/financial news. Our experiments employ GPT-4 for zero-shot/few-shot inference and Open LLaMA for instruction-based fine-tuning. The study demonstrates LLMs' ability to generate well-reasoned decisions by leveraging cross-sequence information and extracting insights from text and price time series. We show that our LLM-based approach outperforms classic ARMA-GARCH and gradient-boosting tree models. Furthermore, fine-tuned public LLMs, such as Open-LLaMA, can generate reasonable and explainable forecasts, although they underperform compared to GPT-4.

11:00-12:30 (East Foyer / Virtual)
**Improving Contextual Query Rewrite for Conversational AI Agents through User-preference Feedback Learning**
*Zhongkai Sun, Yingxue Zhou, Jie Hao, Xing Fan, Yanbin Lu, Chengyuan Ma, Wei Shen and Chenlei Guo*
Contextual query rewriting (CQR) is a crucial component in Conversational AI agents, leveraging the contextual information from previous user-agent conversations to improve the comprehension of current user intent. However, traditional CQR methods often concentrate on supervised fine-tuning only, neglecting the opportunities to learn from user feedback to align with user preferences. Inspired by recent advances in learning from human feedback (LHF), this paper proposes a novel Preference Aligned Contextual Query Rewriting (PA-CQR) framework to enhance the CQR model's capability in generating user preference-aligned rewrites. This paper also investigates the efficacy of various state-of-the-art feedback learning algorithms on the CQR task, and proposes a novel Dynamic Direct Preference Optimization (Dynamic DPO) algorithm to better adapt the DPO algorithm to large-scale CQR training. Experiments on large-scale real-world CQR data set demonstrate the superiority of the proposed PA-CQR framework and the Dynamic DPO.

11:00-12:30 (East Foyer / Virtual)
**InsightNet : Structured Insight Mining from Customer Feedback**
*Sandeep Sricharan Mukku, Manan Soni, Chetan Aggarwal, Jitenkumar Rana, Promod Yenigalla, Rashmi Patange and Shyam Mohan*

We propose InsightNet, a novel approach for the automated extraction of structured insights from customer reviews. Our end-to-end machine learning framework is designed to overcome the limitations of current solutions, including the absence of structure for identified topics, non-standard aspect names, and lack of abundant training data. The proposed solution builds a semi-supervised multi-level taxonomy from raw reviews, a semantic similarity heuristic approach to generate labelled data and employs a multi-task insight extraction architecture by fine-tuning an LLM. InsightNet identifies granular actionable topics with customer sentiments and verbatim for each topic. Evaluations on real-world customer review data show that InsightNet performs better than existing solutions in terms of structure, hierarchy and completeness. We empirically demonstrate that InsightNet outperforms the current state-of-the-art methods in multi-label topic classification, achieving an F1 score of 0.85, which is an improvement of 11% F1-score over the previous best results. Additionally, InsightNet generalises well for unseen aspects and suggests new topics to be added to the taxonomy.

11:00-12:30 (East Foyer / Virtual)
**InstructPTS: Instruction-Tuning LLMs for Product Title Summarization**
*Besnik Fetahu, Zhiyu Chen, Oleg Rokhlenko and Shervin Malmasi*
E-commerce product catalogs contain billions of items. Most products have lengthy titles, as sellers pack them with product attributes to improve retrieval, and highlight key product aspects. This results in a gap between such unnatural products titles, and how customers refer to them. It also limits how e-commerce stores can use these seller-provided titles for recommendation, QA, or review summarization. Inspired by recent work on instruction-tuned LLMs, we present InstructPTS, a controllable approach for the task of Product Title Summarization (PTS). Trained using a novel instruction fine-tuning strategy, our approach is able to summarize product titles according to various criteria (e.g. number of words in a summary, inclusion of specific phrases, etc.). Extensive evaluation on a real-world e-commerce catalog shows that compared to simple fine-tuning of LLMs, our proposed approach can generate more accurate product name summaries, with an improvement of over 14 and 8 BLEU and ROUGE points, respectively.

11:00-12:30 (East Foyer / Virtual)
**Investigating Table-to-Text Generation Capabilities of Large Language Models in Real-World Information Seeking Scenarios**
*Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang and Arman Cohan*
Tabular data is prevalent across various industries, necessitating significant time and effort for users to understand and manipulate for their information-seeking purposes. The advancements in large language models (LLMs) have shown enormous potential to improve user efficiency. However, the adoption of LLMs in real-world applications for table information seeking remains underexplored. In this paper, we investigate the table-to-text capabilities of different LLMs using four datasets within two real-world information seeking scenarios. These include the LogicNLG and our newly-constructed LoTNLG datasets for data insight generation, along with the FeTaQA and our newly-constructed F2WTQ datasets for query-based generation. We structure our investigation around three research questions, evaluating the performance of LLMs in table-to-text generation, automated evaluation, and feedback generation, respectively. Experimental results indicate that the current high-performing LLM, specifically GPT-4, can effectively serve as a table-to-text generator, evaluator, and feedback generator, facilitating users' information seeking purposes in real-world scenarios. However, a significant performance gap still exists between other open-sourced LLMs (e.g., Vicuna and LLaMA-2) and GPT-4 models. Our data and code are publicly available at https://github.com/yale-nlp/LLM-T2T.

11:00-12:30 (East Foyer / Virtual)
**JarviX: A LLM No code Platform for Tabular Data Analysis and Optimization**
*Shang-Ching Liu, ShengKun Wang, Tsungyao Chang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo and Jianwei Zhang*
In this study, we introduce JarviX, a sophisticated data analytics framework. JarviX is designed to employ Large Language Models (LLMs) to facilitate an automated guide and execute high-precision data analyzes on tabular datasets. This framework emphasizes the significance of varying column types, capitalizing on state-of-the-art LLMs to generate concise data insight summaries, propose relevant analysis inquiries, visualize data effectively, and provide comprehensive explanations for results drawn from an extensive data analysis pipeline. Moreover, JarviX incorporates an automated machine learning (AutoML) pipeline for predictive modeling. This integration forms a comprehensive and automated optimization cycle, which proves particularly advantageous for optimizing machine configuration. The efficacy and adaptability of JarviX are substantiated through a series of practical use case studies.

11:00-12:30 (East Foyer / Virtual)
**Joint Dialogue Topic Segmentation and Categorization: A Case Study on Clinical Spoken Conversations**
*Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy and Nancy Chen*
Utilizing natural language processing techniques in clinical conversations is effective to improve the efficiency of health management workflows for medical staff and patients. Dialogue segmentation and topic categorization are two fundamental steps for processing verbose spoken conversations and highlighting informative spans for downstream tasks. However, in practical use cases, due to the variety of segmentation granularity and topic definition, and the lack of diverse annotated corpora, no generic models are readily applicable for domain-specific applications. In this work, we introduce and adopt a joint model for dialogue segmentation and topic categorization, and conduct a case study on healthcare follow-up calls for diabetes management; we provide insights from both data and model perspectives toward performance and robustness.

11:00-12:30 (East Foyer / Virtual)
**KD-Boost: Boosting Real-Time Semantic Matching in E-commerce with Knowledge Distillation**
*Sanjay Agrawal, Vivek Sembium and Ankith M S*
Real-time semantic matching is vital to web and product search. Transformer-based models have shown to be highly effective at encoding queries into an embedding space where semantically similar entities (queries or results) are in close proximity. However, the computational complexity of large transformer models limits their utilization for real-time matching. In this paper, we propose KD-Boost, a novel knowledge distillation algorithm designed for real-time semantic matching. KD-Boost trains low latency accurate student models by leveraging soft labels from a teacher model as well as ground truth via pairwise query-product and query-query signal derived from direct audits, user behavior, and taxonomy-based data using custom loss functions. Experiments on internal and external e-commerce datasets demonstrate an improvement of 2-3% ROC-AUC compared to training student models directly, outperforming teacher and SOTA knowledge distillation benchmarks. Simulated online A/B tests using KD-Boost for automated Query Reformulation (QR) indicate a 6.31% increase in query-to-query matching, 2.76% increase in product coverage, and a 2.19% improvement in relevance.

11:00-12:30 (East Foyer / Virtual)
**Learning Multilingual Sentence Representations with Cross-lingual Consistency Regularization**
*Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu and Haifeng Wang*
Multilingual sentence representations are the foundation for similarity-based bitext mining, which is crucial for scaling multilingual neural machine translation (NMT) system to more languages. In this paper, we introduce MuSR: a one-for-all Multilingual Sentence Representation model that supports 223 languages. Leveraging billions of English-centric parallel corpora, we train a multilingual Transformer encoder,

coupled with an auxiliary Transformer decoder, by adopting a multilingual NMT framework with CrossConST, a cross-lingual consistency regularization technique proposed in Gao et al. (2023). Experimental results on multilingual similarity search and bitext mining tasks show the effectiveness of our approach. Specifically, MuSR achieves superior performance over LASER3 (Heffernan et al., 2022) which consists of 148 independent multilingual sentence encoders.

11:00-12:30 (East Foyer / Virtual)
**Multi-teacher Distillation for Multilingual Spelling Correction**
*Jingfen Zhang, Xuan Guo, Sravan Bodapati and Christopher Potts*
Accurate spelling correction is a critical step in modern search interfaces, especially in an era of mobile devices and speech-to-text interfaces. For services that are deployed around the world, this poses a significant challenge for multilingual NLP: spelling errors need to be caught and corrected in all languages, and even in queries that use multiple languages. In this paper, we tackle this challenge using multi-teacher distillation. On our approach, a monolingual teacher model is trained for each language/locale, and these individual models are distilled into a single multilingual student model intended to serve all languages/locales. In experiments using open-source data as well as customer data from a worldwide search service, we show that this leads to highly effective spelling correction models that can meet the tight latency requirements of deployed services.

11:00-12:30 (East Foyer / Virtual)
**Multi-word Tokenization for Sequence Compression**
*Leonidas Gee, Leonardo Rigutini, Marco Ernandes and Andrea Zugarini*
Large Language Models have proven highly successful at modelling a variety of tasks. However, this comes at a steep computational cost that hinders wider industrial uptake. In this paper, we present MWT: a Multi-Word Tokenizer that goes beyond word boundaries by representing frequent multi-word expressions as single tokens. MWTs produce a more compact and efficient tokenization that yields two benefits: (1) Increase in performance due to a greater coverage of input data given a fixed sequence length budget; (2) Faster and lighter inference due to the ability to reduce the sequence length with negligible drops in performance. Our results show that MWT is more robust across shorter sequence lengths, thus allowing for major speedups via early sequence truncation.

11:00-12:30 (East Foyer / Virtual)
**MUST&P-SRL: Multi-lingual and Unified Syllabification in Text and Phonetic Domains for Speech Representation Learning**
*Noé Tits*
In this paper, we present a methodology for linguistic feature extraction, focusing particularly on automatically syllabifying words in multiple languages, with a design to be compatible with a forced-alignment tool, the Montreal Forced Aligner (MFA). In both the textual and phonetic domains, our method focuses on the extraction of phonetic transcriptions from text, stress marks, and a unified automatic syllabification (in text and phonetic domains). The system was built with open-source components and resources. Through an ablation study, we demonstrate the efficacy of our approach in automatically syllabifying words from several languages (English, French and Spanish). Additionally, we apply the technique to the transcriptions of the CMU ARCTIC dataset, generating valuable annotations available online (https://github.com/noetits/MUST_P-SRL) that are ideal for speech representation learning, speech unit discovery, and disentanglement of speech factors in several speech-related fields.

11:00-12:30 (East Foyer / Virtual)
**On Sample-Efficient Code Generation**
*Hojae Han, Yu Jin Kim, Byoungjip Kim, Youngwon Lee, Kyungjae Lee, Kyungmin Lee, Moontae Lee, Kyunghoon Bae and Seung-won Hwang*
Large language models often struggle to predict runtime behavior in code generation tasks, leading to a reliance on rejection sampling (best-of-n) to generate multiple code snippets then select the best. Our distinction is reducing sampling costs, without compromising generation quality. We introduce EFFICODE, a novel framework that prioritizes sampling on test problems that models can solve. We show how EFFICODE estimates solvability to optimize computational costs during multiple sampling. Based on empirical evidence, EFFICODE consistently demonstrates reduced sampling budgets while maintaining comparable code generation performance, especially when problems are challenging. In addition, utilizing EFFICODE to rank sampled code snippets also shows its effectiveness in answer code selection for reducing temporal costs, by not requiring any execution or test case generation.

11:00-12:30 (East Foyer / Virtual)
**ORANGE: Text-video Retrieval via Watch-time-aware Heterogeneous Graph Contrastive Learning**
*Yucheng Lin, Tim Chang, Yaning Chang, Jianqiang Ma, Donghui Li, Ting Peng, Zang Li, Zhiyi Zhou and Feng Wang*
With the explosive growth of short-video data on industrial video-sharing platforms such as TikTok and YouTube, text-video retrieval techniques have become increasingly important. Most existing works for text-video retrieval focus on designing informative representation learning methods and delicate matching mechanisms, which leverage the content information of queries and videos themselves (i.e., textual information of queries and multimodal information of videos). However, real-world scenarios often involve brief, ambiguous queries and low-quality videos, making content-based retrieval less effective. In order to accommodate various search requirements and enhance user satisfaction, this study introduces a novel Text-video Retrieval method via Watch-time-aware Heterogeneous Graph Contrastive Learning (termed ORANGE). This approach aims to learn informative embeddings for queries and videos by leveraging both content information and the abundant relational information present in video-search scenarios. Specifically, we first construct a heterogeneous information graph where nodes represent domain objects (e.g., query, video, tag) and edges represent rich relations among these objects. Afterwards, a meta-path-guided heterogeneous graph attention encoder with the awareness of video watch time is devised to encode various semantic aspects of query and video nodes. To train our model, we introduce a meta-path-wise contrastive learning paradigm that facilitates capturing dependencies across multiple semantic relations, thereby enhancing the obtained embeddings. Finally, when deployed online, for new queries non-existent in the constructed graph, a bert-based query encoder distilled from our ORANGE is employed. Offline experiments conducted on a real-world dataset demonstrate the effectiveness of our ORANGE. Moreover, it has been implemented in the matching stage of an industrial online video-search service, where it exhibited statistically significant improvements over the online baseline in an A/B test.

11:00-12:30 (East Foyer / Virtual)
**Personalized Dense Retrieval on Global Index for Voice-enabled Conversational Systems**
*Masha Belyi, Charlotte Dzialo, Chaitanya Dwivedi, Prajit Reddy Muppidi and Kanna Shimizu*
Voice-controlled AI dialogue systems are susceptible to noise from phonetic variations and failure to resolve ambiguous entities. Typically, personalized entity resolution (ER) and/or query rewrites (QR) are deployed to recover from these error modes. Previous work in this field achieves personalization by constraining retrieval search space to personalized indices built from user's historical interactions with the device. While constrained retrieval achieves high precision, predictions are limited to entities in recent user history, which offers low coverage of future requests. Further, maintaining individual indices for millions of users is memory intensive and difficult to scale. In this work, we propose a personalized entity retrieval system that is robust to phonetic noise and ambiguity but is not limited to a personalized index. We achieve this by embedding user listening preferences into a contextual query embedding used in retrieval. We demonstrate our model's ability to correct multiple error modes and show 91% improvement over baseline on the entity retrieval task. Finally, we optimize the end-to-end

approach to fit within online latency constraints while maintaining gains in performance.

11:00-12:30 (East Foyer / Virtual)
**PROMINET: Prototype-based Multi-View Network for Interpretable Email Response Prediction**
*Yuqing Wang, Prashanth Vijayaraghavan and Ehsan Degan*
Email is a widely used tool for business communication, and email marketing has emerged as a cost-effective strategy for enterprises. While previous studies have examined factors affecting email marketing performance, limited research has focused on understanding email response behavior by considering email content and metadata. This study proposes a Prototype-based Multi-view Network (PROMINET) that incorporates semantic and structural information from email data. By utilizing prototype learning, the PROMINET model generates latent exemplars, enabling interpretable email response prediction. The model maps learned semantic and structural exemplars to observed samples in the training data at different levels of granularity, such as document, sentence, or phrase. The approach is evaluated on two real-world email datasets: the Enron corpus and an in-house Email Marketing corpus. Experimental results demonstrate that the PROMINET model outperforms baseline models, achieving a 3% improvement in F1 score on both datasets. Additionally, the model provides interpretability through prototypes at different granularity levels while maintaining comparable performance to non-interpretable models. The learned prototypes also show potential for generating suggestions to enhance email text editing and improve the likelihood of effective email responses. This research contributes to enhancing sender-receiver communication and customer engagement in email interactions.

11:00-12:30 (East Foyer / Virtual)
**Query-aware Multi-modal based Ranking Relevance in Video Search**
*Chengcan Ye, Ting Peng, Tim Chang, Zhiyi Zhou and Feng Wang*
Relevance ranking system plays a crucial role in video search on streaming platforms. Most relevance ranking methods focus on text modality, incapable of fully exploiting cross-modal cues present in video. Recent multi-modal models have demonstrated promise in various vision-language tasks but provide limited help for downstream query-video relevance tasks due to the discrepency between relevance ranking-agnostic pre-training objectives and the real video search scenarios that demand comprehensive relevance modeling. To address these challenges, we propose a QUery-Aware pre-training model with multi-modaLITY (QUALITY) that incorporates hard-mined query information as alignment targets and utilizes video tag information for guidance. QUALITY is integrated into our relevance ranking model, which leverages multi-modal knowledge and improves ranking optimization method based on ordinal regression. Extensive experiments show our proposed model significantly enhances video search performance.

11:00-12:30 (East Foyer / Virtual)
**Relevance-assisted Generation for Robust Zero-shot Retrieval**
*Jihyuk Kim, Minsoo Kim, Joonsuk Park and Seung-won Hwang*
Zero-shot retrieval tasks such as the BEIR benchmark reveal out-of-domain generalization as a key weakness of high-performance dense retrievers. As a solution, domain adaptation for dense retrievers has been actively studied. A notable approach is synthesizing domain-specific data, by generating pseudo queries (PQ), for fine-tuning with domain-specific relevance between PQ and documents. Our contribution is showing that key biases can cause sampled PQ to be irrelevant, negatively contributing to generalization. We propose to preempt their generation, by dividing the generation into simpler subtasks, of generating relevance explanations and guiding the generation to avoid negative generalization. Experiment results show that our proposed approach is more robust to domain shifts, validated on challenging BEIR zero-shot retrieval tasks.

11:00-12:30 (East Foyer / Virtual)
**Retrieval-Enhanced Dual Encoder Training for Product Matching**
*Justin Chiu*
Product matching is the task of matching a seller-listed item to an appropriate product. It is a critical task for an e-commerce platform, and the approach needs to be efficient to run in a large-scale setting. A dual encoder approach has been a common practice for product matching recently, due to its high performance and computation efficiency. In this paper, we propose a two-stage training for the dual encoder model. Stage 1 trained a dual encoder to identify the more informative training data. Stage 2 then train on the more informative data to get a better dual encoder model. This technique is a learned approach for building training data. We evaluate the retrieval-enhanced training on two different datasets: a publicly available Large-Scale Product Matching dataset and a real-world e-commerce dataset containing 47 million products. Experiment results show that our approach improved by 2% F1 on the public dataset and 9% F1 on the real-world e-commerce dataset.

11:00-12:30 (East Foyer / Virtual)
**Retrieve and Copy: Scaling ASR Personalization to Large Catalogs**
*Sai Muralidhar Jayanthi, Devang Kulshreshtha, Saket Dingliwal, Srikanth Ronanki and Sravan Bodapati*
Personalization of automatic speech recognition (ASR) models is a widely studied topic because of its many practical applications. Most recently, attention-based contextual biasing techniques are used to improve the recognition of rare words and/or domain specific entities. However, due to performance constraints, the biasing is often limited to a few thousand entities, restricting real-world usability. To address this, we first propose a "Retrieve and Copy" mechanism to improve latency while retaining the accuracy even when scaled to a large catalog. We also propose a training strategy to overcome the degradation in recall at such scale due to an increased number of confusing entities. Overall, our approach achieves up to 6% more Word Error Rate reduction (WERR) and 3.6% absolute improvement in F1 when compared to a strong baseline. Our method also allows for large catalog sizes of up to 20K without significantly affecting WER and F1-scores, while achieving at least 20% inference speedup per acoustic frame.

11:00-12:30 (East Foyer / Virtual)
**SAMP: A Model Inference Toolkit of Post-Training Quantization for Text Processing via Self-Adaptive Mixed-Precision**
*Rong Tian, Zijing Zhao, Weijie Liu, Haoyan Liu, Weiquan Mao, Zhe Zhao and Kan Zhou*
The latest industrial inference engines, such as FasterTransformer and TurboTransformers, have verified that half-precision floating point (FP16) and 8-bit integer (INT8) quantization can greatly improve model inference speed. However, the existing INT8 quantization methods are too complicated, and improper usage will lead to model performance damage greatly. In this paper, we develop a toolkit for users to easily quantize their models for inference, in which Self-Adaptive Mixed-Precision (SAMP) is proposed to automatically control quantization rate by a mixed-precision architecture to balance model accuracy and efficiency. Experimental results show that our SAMP toolkit has a higher speedup than PyTorch and FasterTransformer while ensuring the required accuracy. In addition, SAMP is based on a modular design, decoupling the tokenizer, embedding, encoder and target layers, which allows users to handle various downstream tasks and can be seamlessly integrated into PyTorch.

11:00-12:30 (East Foyer / Virtual)
**Scaling Neural ITN for Numbers and Temporal Expressions in Tamil: Findings for an Agglutinative Low-resource Language**
*Bhavuk Singhal, Sindhuja Gopalan, Amrith Krishna and Malolan Chetlur*

ITN involves rewriting the verbalised form of text from spoken transcripts to its corresponding written form. The task inherently expects challenges in identifying ITN entries due to spelling variations in words arising out of dialects, transcription errors etc. Additionally, in Tamil, word boundaries between adjacent words in a sentence often get obscured due to Punarchi, i.e. phonetic transformation of these boundaries. Being morphologically rich, the words in Tamil show a high degree of agglutination due to inflection and clitics. The combination of such factors leads to a high degree of surface-form variations, making scalability with pure rule-based approaches difficult. Instead, we experiment with fine-tuning three pre-trained neural LMs, consisting of a seq2seq model (s2s), a non-autoregressive text editor (NAR) and a sequence tagger + rules combination (tagger). While the tagger approach works best in a fully-supervised setting, s2s performs the best (98.05 F-Score) when augmented with additional data, via bootstrapping and data augmentation (DA&B). S2S reports a cumulative percentage improvement of 20.1 %, and statistically significant gains for all our models with DA&B. Compared to a fully supervised setup, bootstrapping alone reports a percentage improvement as high as 14.12 %, even with a small seed set of 324 ITN entries.

11:00-12:30 (East Foyer / Virtual)
**Speakerly: A Voice-based Writing Assistant for Text Composition**
*Dhruv Kumar, Vipul Raheja, Alice Kaiser-Schatzlein, Robyn Perry, Apurva Joshi, Justin Hugues-Nuger, Samuel Lou and Navid Chowdhury*
We present Speakerly, a new real-time voice-based writing assistance system that helps users with text composition across various use cases such as emails, instant messages, and notes. The user can interact with the system through instructions or dictation, and the system generates a well-formatted and coherent document. We describe the system architecture and detail how we address the various challenges while building and deploying such a system at scale. More specifically, our system uses a combination of small, task-specific models as well as pre-trained language models for fast and effective text composition while supporting a variety of input modes for better usability.

11:00-12:30 (East Foyer / Virtual)
**STEER: Semantic Turn Extension-Expansion Recognition for Voice Assistants**
*Leon Zhang, Jiarui Lu, Joel Ruben Antony Moniz, Aditya Kulkarni, Dhivya Piraviperumal, Tien Dung Tran, Nick Tzou and Hong Yu*
In the context of a voice assistant system, steering refers to the phenomenon in which a user issues a follow-up command attempting to direct or clarify a previous turn. We propose STEER, a steering detection model that predicts whether a follow-up turn is a user's attempt to steer the previous command. Constructing a training dataset for steering use cases poses challenges due to the cold-start problem. To overcome this, we developed heuristic rules to sample opt-in usage data, approximating positive and negative samples without any annotation. Our experimental results show promising performance in identifying steering intent, with over 95% accuracy on our sampled data. Moreover, STEER, in conjunction with our sampling strategy, aligns effectively with real-world steering scenarios, as evidenced by its strong zero-shot performance on a human-graded evaluation set. In addition to relying solely on user transcripts as input, we introduce STEER+, an enhanced version of the model. STEER+ utilizes a semantic parse tree to provide more context on out-of-vocabulary words, such as named entities that often occur at the sentence boundary. This further improves model performance, reducing error rate in domains where entities frequently appear, such as messaging. Lastly, we present a data analysis that highlights the improvement in user experience when voice assistants support steering use cases.

11:00-12:30 (East Foyer / Virtual)
**Text2Topic: Multi-Label Text Classification System for Efficient Topic Detection in User Generated Content with Zero-Shot Capabilities**
*Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi and Benjamin Wang*
Multi-label text classification is a critical task in the industry. It helps to extract structured information from large amount of textual data. We propose Text to Topic (Text2Topic), which achieves high multi-label classification performance by employing a Bi-Encoder Transformer architecture that utilizes concatenation, subtraction, and multiplication of embeddings on both text and topic. Text2Topic also supports zero-shot predictions, produces domain-specific text embeddings, and enables production-scale batch-inference with high throughput. The final model achieves accurate and comprehensive results compared to state-of-the-art baselines, including large language models (LLMs). In this study, a total of 239 topics are defined, and around 1.6 million text-topic pairs annotations (in which 200K are positive) are collected on approximately 120K texts from 3 main data sources on Booking.com. The data is collected with optimized smart sampling and partial labeling. The final Text2Topic model is deployed on a real-world stream processing platform, and it outperforms other models with 92.9% micro mAP, as well as a 75.8% macro mAP score. We summarize the modeling choices which are extensively tested through ablation studies, and share detailed in-production decision-making steps.

11:00-12:30 (East Foyer / Virtual)
**TMID: A Comprehensive Real-world Dataset for Trademark Infringement Detection in E-Commerce**
*Tongxin Hu, Zhuang Li, Xin Jin, Lizhen Qu and Xin Zhang*
Annually, e-commerce platforms incur substantial financial losses due to trademark infringements, making it crucial to identify and mitigate potential legal risks tied to merchant information registered to the platforms. However, the absence of high-quality datasets hampers research in this area. To address this gap, our study introduces TMID, a novel dataset to detect trademark infringement in merchant registrations. This is a real-world dataset sourced directly from Alipay, one of the world's largest e-commerce and digital payment platforms. As infringement detection is a legal reasoning task requiring an understanding of the contexts and legal rules, we offer a thorough collection of legal rules and merchant and trademark-related contextual information with annotations from legal experts. We ensure the data quality by performing an extensive statistical analysis. Furthermore, we conduct an empirical study on this dataset to highlight its value and the key challenges. Through this study, we aim to contribute valuable resources to advance research into legal compliance related to trademark infringement within the e-commerce sphere.

11:00-12:30 (East Foyer / Virtual)
**Too much of product information : Don't worry, let's look for evidence!**
*Aryan Jain, Jitenkumar Rana and Chetan Aggarwal*
Product question answering (PQA) aims to provide an instant response to customer questions posted on shopping message boards, social media, brand websites and retail stores. In this paper, we propose a distantly supervised solution to answer customer questions by using product information. Auto-answering questions using product information poses two main challenges:(i) labelled data is not readily available (ii)lengthy product information requires attending to various parts of the text to answer the question. To this end, we first propose a novel distant supervision based NLI model to prepare training data without any manual efforts. To deal with lengthy context, we factorize answer generation into two sub-problems. First, given product information, model extracts evidence spans relevant to question. Then, model leverages evidence spans to generate answer. Further, we propose two novelties in fine-tuning approach: (i) First, we jointly fine-tune model for both the tasks in end-to-end manner and showcase that it outperforms standard multi-task fine-tuning. (ii) Next, we introduce an auxiliary contrastive loss for evidence extraction. We show that combination of these two ideas achieves an absolute improvement of 6% in accuracy (human evaluation) over baselines.

11:00-12:30 (East Foyer / Virtual)

**Towards Effective Automatic Debt Collection with Persona Awareness**
*Tong Zhang, Junhong Liu, Chen Huang, Jia Liu, Hongru Liang, Zujie Wen and Wenqiang Lei*
Understanding debtor personas is crucial for collectors to empathize with debtors and develop more effective collection strategies. In this paper, we take the first step towards comprehensively investigating the significance of debtor personas and present a successful commercial practice on automatic debt collection agents. Specifically, we organize the debtor personas into a taxonomy and construct a persona-aware conversation dataset. Building upon it, we implement a simple yet effective persona-aware agent called PAD. After two-month online testing, PAD increases the recovery rate by 3.31% and collects an additional $\sim$100K RMB. Our commercial practice brings inspiration to the debt collection industry by providing an effective automatic solution.

11:00-12:30 (East Foyer / Virtual)
**Towards Safer Operations: An Expert-involved Dataset of High-Pressure Gas Incidents for Preventing Future Failures**
*Shumpei Inoue, Minh-Tien Nguyen, Hiroki Mizokuchi, Tuan-Anh D. Nguyen, Huu-Hiep Nguyen and Dung Le*
This paper introduces a new IncidentAI dataset for safety prevention. Different from prior corpora that usually contain a single task, our dataset comprises three tasks: named entity recognition, cause-effect extraction, and information retrieval. The dataset is annotated by domain experts who have at least six years of practical experience as high-pressure gas conservation managers. We validate the contribution of the dataset in the scenario of safety prevention. Preliminary results on the three tasks show that NLP techniques are beneficial for analyzing incident reports to prevent future failures. The dataset facilitates future research in NLP and incident management communities. The access to the dataset is also provided (The IncidentAI dataset is available at: https://github.com/Cinnamon/incident-ai-dataset).

11:00-12:30 (East Foyer / Virtual)
**Unveiling Identity Biases in Toxicity Detection : A Game-Focused Dataset and Reactivity Analysis Approach**
*Josiane Van Dorpe, Zachary Yang, Nicolas Grenon-Godbout and Grégoire Winterstein*
Identity biases arise commonly from annotated datasets, can be propagated in language models and can cause further harm to marginal groups. Existing bias benchmarking datasets are mainly focused on gender or racial biases and are made to pinpoint which class the model is biased towards. They also are not designed for the gaming industry, a concern for models built for toxicity detection in videogames' chat. We propose a dataset and a method to highlight oversensitive terms using reactivity analysis and the model's performance. We test our dataset against ToxBuster, a language model developed by Ubisoft fine-tuned for toxicity detection on multiplayer videogame's written chat, and Perspective API. We find that these toxicity models often automatically tag terms related to a community's identity as toxic, which prevents members of already marginalized groups to make their presence known or have a mature / normal conversation. Through this process, we have generated an interesting list of terms that trigger the models to varying degrees, along with insights on establishing a baseline through human annotations.

11:00-12:30 (East Foyer / Virtual)
**ViGPTQA - State-of-the-Art LLMs for Vietnamese Question Answering: System Overview, Core Models Training, and Evaluations**
*Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen and Xuan-Son Vu*
Large language models (LLMs) and their applications in low-resource languages (such as in Vietnamese) are limited due to lack of training data and benchmarking datasets. This paper introduces a practical real-world implementation of a question answering system for Vietnamese, called ViGPTQA, leveraging the power of LLM. Since there is no effective LLM in Vietnamese to date, we also propose, evaluate, and open-source an instruction-tuned LLM for Vietnamese, named ViGPT. ViGPT demonstrates exceptional performances, especially on real-world scenarios. We curate a new set of benchmark datasets that encompass both AI and human-generated data, providing a comprehensive evaluation framework for Vietnamese LLMs. By achieving state-of-the-art results and approaching other multilingual LLMs, our instruction-tuned LLM underscores the need for dedicated Vietnamese-specific LLMs. Our open-source model supports customized and privacy-fulfilled Vietnamese language processing systems.

11:00-12:30 (East Foyer / Virtual)
**VKIE: The Application of Key Information Extraction on Video Text**
*Siyu An, Ye Liu, Haoyuan Peng and Di Yin*
Extracting structured information from videos is critical for numerous downstream applications in the industry. In this paper, we define a significant task of extracting hierarchical key information from visual texts on videos. To fulfill this task, we decouple it into four subtasks and introduce two implementation solutions called PipVKIE and UniVKIE. PipVKIE sequentially completes the four subtasks in continuous stages, while UniVKIE is improved by unifying all the subtasks into one backbone. Both PipVKIE and UniVKIE leverage multimodal information from vision, text, and coordinates for feature representation. Extensive experiments on one well-defined dataset demonstrate that our solutions can achieve remarkable performance and efficient inference speed.

11:00-12:30 (East Foyer / Virtual)
**Welcome to the Real World: Efficient, Incremental and Scalable Key Point Analysis**
*Lilach Eden, Yoav Kantor, Matan Orbach, Yoav Katz, Noam Slonim and Roy Bar-Haim*
Key Point Analysis (KPA) is an emerging summarization framework, which extracts the main points from a collection of opinions, and quantifies their prevalence. It has been successfully applied to diverse types of data, including arguments, user reviews and survey responses. Despite the growing academic interest in KPA, little attention has been given to the practical challenges of implementing a KPA system in production. This work presents a deployed KPA system, which regularly serves multiple teams in our organization. We discuss the main challenges we faced while building a real-world KPA system, as well as the architecture and algorithmic improvements we developed to address these challenges. Specifically, we focus on efficient matching of sentences to key points, incremental processing, scalability and resiliency. The value of our contributions is demonstrated in an extensive set of experiments, over five existing and novel datasets. Finally, we describe several use cases of the deployed system, which illustrate its practical value.

11:00-12:30 (East Foyer / Virtual)
**WordArt Designer: User-Driven Artistic Typography Synthesis using Large Language Models**
*Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Jingdong Sun, Wangmeng Xiang, Xianhui Lin, Xiaoyang Kang, Zengke Jin, Yusen Hu, Bin Luo, Yifeng Geng and Xuansong Xie*
This paper introduces WordArt Designer, a user-driven framework for artistic typography synthesis, relying on the Large Language Model (LLM). The system incorporates four key modules: the LLM Engine, SemTypo, StyTypo, and TexTypo modules. 1) The LLM Engine, empowered by the LLM (e.g. GPT-3.5), interprets user inputs and generates actionable prompts for the other modules, thereby transforming abstract concepts into tangible designs. 2) The SemTypo module optimizes font designs using semantic concepts, striking a balance between artistic transformation and readability. 3) Building on the semantic layout provided by the SemTypo module, the StyTypo module creates smooth, refined images. 4) The TexTypo module further enhances the design's aesthetics through texture rendering, enabling the generation of inventive textured fonts. Notably, WordArt Designer highlights the fusion of generative AI with artistic typography. Experience its capabilities on ModelScope: https://www.modelscope.cn/studios/WordArt/WordArt.

## Findings 7

11:00-12:30 (East Foyer / Virtual)

11:00-12:30 (East Foyer / Virtual)
**Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning**
*Liangming Pan, Alon Albalak, Xinyi Wang and William Yang Wang*
Large Language Models (LLMs) have shown human-like reasoning abilities but still struggle with complex logical problems. This paper introduces a novel framework, Logic-LM, which integrates LLMs with symbolic solvers to improve logical problem-solving. Our method first utilizes LLMs to translate a natural language problem into a symbolic formulation. Afterward, a deterministic symbolic solver performs inference on the formulated problem. We also introduce a self-refinement module, which utilizes the symbolic solver's error messages to revise symbolic formalizations. We demonstrate Logic-LM's effectiveness on five logical reasoning datasets: ProofWriter, PrOntoQA, FOLIO, LogicalDeduction, and AR-LSAT. On average, Logic-LM achieves a significant performance boost of 39.2% over using LLM alone with standard prompting and 18.4% over LLM with chain-of-thought prompting. Our findings suggest that Logic-LM, by combining LLMs with symbolic logic, offers a promising avenue for faithful logical reasoning.

11:00-12:30 (East Foyer / Virtual)
**HANSEN: Human and AI Spoken Text Benchmark for Authorship Analysis**
*Nafis Irtiza Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti and Dongwon Lee*
*Authorship Analysis*, also known as stylometry, has been an essential aspect of Natural Language Processing (NLP) for a long time. Likewise, the recent advancement of Large Language Models (LLMs) has made authorship analysis increasingly crucial for distinguishing between human-written and AI-generated texts. However, these authorship analysis tasks have primarily been focused on *written texts*, not considering *spoken texts*. Thus, we introduce the largest benchmark for spoken texts - HANSEN(*H*uman *AN*d ai *S*poken t*E*xt be*N*chmark). HANSEN encompasses meticulous curation of existing speech datasets accompanied by transcripts, alongside the creation of novel AI-generated spoken text datasets. Together, it comprises 17 human datasets, and AI-generated spoken texts created using 3 prominent LLMs: ChatGPT, PaLM2, and Vicuna13B. To evaluate and demonstrate the utility of HANSEN, we perform Authorship Attribution (AA) & Author Verification (AV) on human-spoken datasets and conducted Human vs. AI text detection using state-of-the-art (SOTA) models. While SOTA methods, such as, character n-gram or Transformer-based model, exhibit similar AA & AV performance in human-spoken datasets compared to written ones, there is much room for improvement in AI-generated spoken text detection. The HANSEN benchmark is available at: https://huggingface.co/datasets/HANSEN-REPO/HANSEN

11:00-12:30 (East Foyer / Virtual)
**GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets**
*Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar and Stefan Dietze*
Named Entity Recognition (NER) models play a crucial role in various NLP tasks, including information extraction (IE) and text understanding. In academic writing, references to machine learning models and datasets are fundamental components of various computer science publications and necessitate accurate models for identification. Despite the advancements in NER, existing ground truth datasets do not treat fine-grained types like ML model and model architecture as separate entity types, and consequently, baseline models cannot recognize them as such. In this paper, we release a corpus of 100 manually annotated full-text scientific publications and a first baseline model for 10 entity types centered around ML models and datasets. In order to provide a nuanced understanding of how ML models and datasets are mentioned and utilized, our dataset also contains annotations for informal mentions like "our BERT-based model" or "an image CNN". You can find the ground truth dataset and code to replicate model training at https://data.gesis.org/gsap/gsap-ner.

11:00-12:30 (East Foyer / Virtual)
**Quantifying the Dialect Gap and its Correlates Across Languages**
*Anjali Kantharuban, Ivan Vulić and Anna Korhonen*
Historically, researchers and consumers have noticed a decrease in quality when applying NLP tools to minority variants of languages (i.e. Puerto Rican Spanish or Swiss German), but studies exploring this have been limited to a select few languages. Additionally, past studies have mainly been conducted in a monolingual context, so cross-linguistic trends have not been identified and tied to external factors. In this work, we conduct a comprehensive evaluation of the most influential, state-of-the-art large language models (LLMs) across two high-use applications, machine translation and automatic speech recognition, to assess their functionality on the regional dialects of several high- and low-resource languages. Additionally, we analyze how the regional dialect gap is correlated with economic, social, and linguistic factors. The impact of training data, including related factors like dataset size and its construction procedure, is shown to be significant but not consistent across models or languages, meaning a one-size-fits-all approach cannot be taken in solving the dialect gap. This work will lay the foundation for furthering the field of dialectal NLP by laying out evident disparities and identifying possible pathways for addressing them through mindful data collection.

11:00-12:30 (East Foyer / Virtual)
**Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches**
*Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel and Aida Nematzadeh*
People rely heavily on context to enrich meaning beyond what is literally said, enabling concise but effective communication. To interact successfully and naturally with people, user-facing artificial intelligence systems will require similar skills in pragmatics: relying on various types of context — from shared linguistic goals and conventions, to the visual and embodied world — to use language effectively. We survey existing grounded settings and pragmatic modeling approaches and analyze how the task goals, environmental contexts, and communicative affordances in each work enrich linguistic meaning. We present recommendations for future grounded task design to naturally elicit pragmatic phenomena, and suggest directions that focus on a broader range of communicative contexts and affordances.

11:00-12:30 (East Foyer / Virtual)
**Getting MoRE out of Mixture of Language Model Reasoning Experts**
*Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer and Jordan Lee Boyd-Graber*
While recent large language models (LLMs) improve on various question answering (QA) datasets, it remains difficult for a single model to generalize across question types that require distinct reasoning abilities. We provide empirical evidence that state-of-the-art LLMs suffer from poor generalizability on reasoning types beyond those seen in the prompt. To remedy this, we propose a Mixture-of-Reasoning-Experts (MORE) framework that ensembles diverse specialized language models. We specialize the backbone language model with prompts optimized for different reasoning categories, including factual, multihop, mathematical, and commonsense reasoning. Our key insight is to leverage agreement among the specialized experts to select the best answer for each question, or to abstain from answering. This gives MORE higher accuracy than any single specialized model on a collection of 12 QA datasets from four reasoning types. Beyond generalizability, the interpretable design of MORE improves selective question answering results compared to baselines without incorporating inter-expert agree-

ment. This framework is also more interpretable and useful to human consumers of QA outputs. Our human study confirms that presenting expert predictions and the answer selection process helps annotators more accurately calibrate when to trust the system's output. We release all code and data to facilitate future work.

11:00-12:30 (East Foyer / Virtual)
**CLMSM: A Multi-Task Learning Framework for Pre-training on Procedural Text**
*Abhilash Nandy, Manav Nitin Kapadnis, Pawan Goyal and Niloy Ganguly*
In this paper, we propose ***CLMSM***, a domain-specific, continual pre-training framework, that learns from a large set of procedural recipes. ***CLMSM*** uses a Multi-Task Learning Framework to optimize two objectives - a) Contrastive Learning using hard triplets to learn fine-grained differences across entities in the procedures, and b) a novel Mask-Step Modelling objective to learn step-wise context of a procedure. We test the performance of ***CLMSM*** on the downstream tasks of tracking entities and aligning actions between two procedures on three datasets, one of which is an open-domain dataset not conforming with the pre-training dataset. We show that ***CLMSM*** not only outperforms baselines on recipes (in-domain) but is also able to generalize to open-domain procedural NLP tasks.

11:00-12:30 (East Foyer / Virtual)
**MEGClass: Extremely Weakly Supervised Text Classification via Mutually-Enhancing Text Granularities**
*Priyanka Kargupta, Tanay Komarlu, Susik Yoon, Xuan Wang and Jiawei Han*
Text classification is essential for organizing unstructured text. Traditional methods rely on human annotations or, more recently, a set of class seed words for supervision, which can be costly, particularly for specialized or emerging domains. To address this, using class surface names alone as extremely weak supervision has been proposed. However, existing approaches treat different levels of text granularity (documents, sentences, or words) independently, disregarding inter-granularity class disagreements and the context identifiable exclusively through joint extraction. In order to tackle these issues, we introduce MEGClass, an extremely weakly-supervised text classification method that leverages Mutually-Enhancing Text Granularities. MEGClass utilizes coarse- and fine-grained context signals obtained by jointly considering a document's most class-indicative words and sentences. This approach enables the learning of a contextualized document representation that captures the most discriminative class indicators. By preserving the heterogeneity of potential classes, MEGClass can select the most informative class-indicative documents as iterative feedback to enhance the initial word-based class representations and ultimately fine-tune a pre-trained text classifier. Extensive experiments on seven benchmark datasets demonstrate that MEGClass outperforms other weakly and extremely weakly supervised methods.

11:00-12:30 (East Foyer / Virtual)
**Drilling Down into the Discourse Structure with LLMs for Long Document Question Answering**
*Inderjeet Jayakumar Nair, Shwetha Somasundaram, Apoorv Saxena and Koustava Goswami*
We address the task of evidence retrieval for long document question answering, which involves locating relevant paragraphs within a document to answer a question. We aim to assess the applicability of large language models (LLMs) in the task of zero-shot long document evidence retrieval, owing to their unprecedented performance across various NLP tasks. However, currently the LLMs can consume limited context lengths as input, thus providing document chunks as inputs might overlook the global context while missing out on capturing the inter-segment dependencies. Moreover, directly feeding the large input sets can incur significant computational costs, particularly when processing the entire document (and potentially incurring monetary expenses with enterprise APIs like OpenAI's GPT variants). To address these challenges, we propose a suite of techniques that exploit the discourse structure commonly found in documents. By utilizing this structure, we create a condensed representation of the document, enabling a more comprehensive understanding and analysis of relationships between different parts. We retain $91.6\%$ of the best zero-shot approach's performance, while processing only $26\%$ of the total tokens used by the best approach in the information seeking evidence retrieval setup. We also show how our approach can be combined with *self-ask* reasoning agent to achieve best zero-shot performance in complex multi-hop question answering, just $\approx 4\%$ short of zero-shot performance using gold evidence.

11:00-12:30 (East Foyer / Virtual)
**SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages**
*Maharaj Brahma, Kaushal Kumar Maurya and Maunendra Sankar Desarkar*
In this work, we focus on the task of machine translation (MT) from extremely low-resource language (ELRLs) to English. The unavailability of parallel data, lack of representation from large multilingual pre-trained models, and limited monolingual data hinder the development of MT systems for ELRLs. However, many ELRLs often share lexical similarities with high-resource languages (HRLs) due to factors such as dialectical variations, geographical proximity, and language structure. We utilize this property to improve cross-lingual signals from closely related HRL to enable MT for ELRLs. Specifically, we propose a novel unsupervised approach, *SelectNoise*, based on *selective candidate extraction* and *noise injection* to generate noisy HRLs training data. The noise injection acts as a regularizer, and the model trained with noisy data learns to handle lexical variations such as spelling, grammar, and vocabulary changes, leading to improved cross-lingual transfer to ELRLs. The selective candidates are extracted using BPE merge operations and edit operations, and noise injection is performed using greedy, top-p, and top-k sampling strategies. We evaluate the proposed model on 12 ELRLs from the FLORES-200 benchmark in a zero-shot setting across two language families. The proposed model outperformed all the strong baselines, demonstrating its efficacy. It has comparable performance with the supervised noise injection model. Our code and model are publicly available.

11:00-12:30 (East Foyer / Virtual)
**The PEACE-Reviews dataset: Modeling Cognitive Appraisals in Emotion Text Analysis**
*Gerard Christopher Yeo and Kokil Jaidka*
Cognitive appraisal plays a pivotal role in deciphering emotions. Recent studies have delved into its significance, yet the interplay between various forms of cognitive appraisal and specific emotions, such as joy and anger, remains an area of exploration in consumption contexts. Our research introduces the PEACE-Reviews dataset, a unique compilation of annotated autobiographical accounts where individuals detail their emotional and appraisal experiences during interactions with personally significant products or services. Focusing on the inherent variability in consumer experiences, this dataset offers an in-depth analysis of participants' psychological traits, their evaluative feedback on purchases, and the resultant emotions. Notably, the PEACE-Reviews dataset encompasses emotion, cognition, individual traits, and demographic data. We also introduce preliminary models that predict certain features based on the autobiographical narratives.

11:00-12:30 (East Foyer / Virtual)
**ClozEx: A Task toward Generation of English Cloze Explanation**
*Zizheng Zhang, Masato Mita and Mamoru Komachi*
Providing explanations for cloze questions in language assessment (LA) has been recognized as a valuable approach to enhancing the language proficiency of learners. However, there is a noticeable absence of dedicated tasks and datasets specifically designed for generating language learner explanations. In response to this gap, this paper introduces a novel task ClozEx of generating explanations for cloze questions in LA, with a particular focus on English as a Second Language (ESL) learners. To support this task, we present a meticulously curated dataset comprising cloze questions paired with corresponding explanations. This dataset aims to assess language proficiency and facilitates language

learning by offering informative and accurate explanations. To tackle the task, we fine-tuned various baseline models with our training data, including encoder-decoder and decoder-only architectures. We also explored whether large language models (LLMs) are able to generate good explanations without fine-tuning, just using pre-defined prompts. The evaluation results demonstrate that encoder-decoder models have the potential to deliver fluent and valid explanations when trained on our dataset.

11:00-12:30 (East Foyer / Virtual)
**Large Language Models Meet Harry Potter: A Dataset for Aligning Dialogue Agents with Characters**
*Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang and Jia Li*
In recent years, Dialogue-style Large Language Models (LLMs) such as ChatGPT and GPT4 have demonstrated immense potential in constructing open-domain dialogue agents. However, aligning these agents with specific characters or individuals remains a considerable challenge due to the complexities of character representation and the lack of comprehensive annotations. In this paper, we introduce the Harry Potter Dialogue (HPD) dataset, designed to advance the study of dialogue agents and character alignment. The dataset encompasses all dialogue sessions (in both English and Chinese) from the Harry Potter series and is annotated with vital background information, including dialogue scenes, speakers, character relationships, and attributes. These extensive annotations may empower LLMs to unlock character-driven dialogue capabilities. Furthermore, it can serve as a universal benchmark for evaluating how well can a LLM aligning with a specific character. We benchmark LLMs on HPD using both fine-tuning and in-context learning settings. Evaluation results reveal that although there is substantial room for improvement in generating high-quality, character-aligned responses, the proposed dataset is valuable in guiding models toward responses that better align with the character of Harry Potter.

11:00-12:30 (East Foyer / Virtual)
**NASH: A Simple Unified Framework of Structured Pruning for Accelerating Encoder-Decoder Language Models**
*Jongwoo Ko, Seungjoon Park, Yujin Kim, Sumyeong Ahn, Du-Seong Chang, Euijai Ahn and Se-Young Yun*
Structured pruning methods have proven effective in reducing the model size and accelerating inference speed in various network architectures such as Transformers. Despite the versatility of encoder-decoder models in numerous NLP tasks, the structured pruning methods on such models are relatively less explored compared to encoder-only models. In this study, we investigate the behavior of the structured pruning of the encoder-decoder models in the decoupled pruning perspective of the encoder and decoder component, respectively. Our findings highlight two insights: (1) the number of decoder layers is the dominant factor of inference speed, and (2) low sparsity in the pruned encoder network enhances generation quality. Motivated by these findings, we propose a simple and effective framework, NASH, that narrows the encoder and shortens the decoder networks of encoder-decoder models. Extensive experiments on diverse generation and inference tasks validate the effectiveness of our method in both speedup and output quality.

11:00-12:30 (East Foyer / Virtual)
**Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization**
*Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You and Lidong Bing*
With the recent undeniable advancement in reasoning abilities in large language models (LLMs) like ChatGPT and GPT-4, there is a growing trend for using LLMs on various tasks. One area where LLMs can be employed is as an alternative evaluation metric for complex generative tasks, which generally demands expensive human judges to complement the traditional automatic metrics for various evaluation dimensions such as fluency and consistency. In this work, we conduct extensive analysis to investigate the stability and reliability of LLMs as automatic evaluators for abstractive summarization. We found that while ChatGPT and GPT-4 outperform the commonly used automatic metrics, they are not ready as human replacements due to significant limitations. That is, LLM evaluators rate each candidate system inconsistently and are dimension-dependent. They also struggle to compare candidates with close performance and become more unreliable with higher-quality summaries by obtaining a lower correlation with humans. In other words, with better abstractive summarization systems being introduced at a fast pace, LLMs may result in misleading and unreliable evaluations.

11:00-12:30 (East Foyer / Virtual)
**Universal Domain Adaptation for Robust Handling of Distributional Shifts in NLP**
*Hyuhng Joon Kim, Hyunsoo Cho, Sang-Woo Lee, Junyeob Kim, Choonghyun Park, Sang-goo Lee, Kang Min Yoo and Taeuk Kim*
When deploying machine learning systems to the wild, it is highly desirable for them to effectively leverage prior knowledge to the unfamiliar domain while also firing alarms to anomalous inputs. In order to address these requirements, Universal Domain Adaptation (UniDA) has emerged as a novel research area in computer vision, focusing on achieving both adaptation ability and robustness (i.e., the ability to detect out-of-distribution samples). While UniDA has led significant progress in computer vision, its application on language input still needs to be explored despite its feasibility. In this paper, we propose a comprehensive benchmark for natural language that offers thorough viewpoints of the model's generalizability and robustness. Our benchmark encompasses multiple datasets with varying difficulty levels and characteristics, including temporal shifts and diverse domains. On top of our testbed, we validate existing UniDA methods from computer vision and state-of-the-art domain adaptation techniques from NLP literature, yielding valuable findings: We observe that UniDA methods originally designed for image input can be effectively transferred to the natural language domain while also underscoring the effect of adaptation difficulty in determining the model's performance.

11:00-12:30 (East Foyer / Virtual)
**GTA: Gated Toxicity Avoidance for LM Performance Preservation**
*Heegyu Kim and Hyunsouk Cho*
Caution: This paper includes offensive words that could potentially cause unpleasantness. The fast-paced evolution of generative language models such as GPT-4 has demonstrated outstanding results in various NLP generation tasks. However, due to the potential generation of offensive words related to race or gender, various Controllable Text Generation (CTG) methods have been proposed to mitigate the occurrence of harmful words. However, existing CTG methods not only reduce toxicity but also negatively impact several aspects of the language model's generation performance, including topic consistency, grammar, and perplexity. This paper explores the limitations of previous methods and introduces a novel solution in the form of a simple Gated Toxicity Avoidance (GTA) that can be applied to any CTG method. We also evaluate the effectiveness of the proposed GTA by comparing it with state-of-the-art CTG methods across various datasets. Our findings reveal that gated toxicity avoidance efficiently achieves comparable levels of toxicity reduction to the original CTG methods while preserving the generation performance of the language model.

11:00-12:30 (East Foyer / Virtual)
**Logic Unveils Truth, While Disguise Obscures It: Transition Logic Augmented Response Selection for Multi-Turn Dialogue**
*Tingchen Fu, Xueliang Zhao, Lemao Liu and Rui Yan*
Multi-turn response selection aims to retrieve a response for a dialogue context from a candidate pool and negative sampling is the key to its retrieval performance. However, previous methods of negative samples tend to yield false negatives due to the one-to-many property in open-domain dialogue, which is detrimental to the optimization process. To deal with the problem, we propose a sequential variational ladder auto-encoder to capture the diverse one-to-many transition pattern of multiple characteristics in open-domain dialogue. The learned transition logic thus assists in identifying potential positives in disguise. Meanwhile, we propose a TRIGGER framework to adjust negative sampling

in the training process such that the scope of false negatives dynamically updates according to the model capacity. Extensive experiments on two benchmarks verify the effectiveness of our approach.

11:00-12:30 (East Foyer / Virtual)
**Generative Emotion Cause Triplet Extraction in Conversations with Commonsense Knowledge**
*Fanfan Wang, Jianfei Yu and Rui Xia*
Emotion Cause Triplet Extraction in Conversations (ECTEC) aims to simultaneously extract emotion utterances, emotion categories, and cause utterances from conversations. However, existing studies mainly decompose the ECTEC task into multiple subtasks and solve them in a pipeline manner. Moreover, since conversations tend to contain many informal and implicit expressions, it often requires external knowledge and reasoning-based inference to accurately identify emotional and causal clues implicitly mentioned in the context, which are ignored by previous work. To address these limitations, in this paper, we propose a commonSense knowledge-enHanced generAtive fRameworK named SHARK, which formulates the ECTEC task as an index generation problem and generates the emotion-cause-category triplets in an end-to-end manner with a sequence-to-sequence model. Furthermore, we propose to incorporate both retrieved and generated commonsense knowledge into the generative model via a dual-view gate mechanism and a graph attention layer. Experimental results show that our SHARK model consistently outperforms several competitive systems on two benchmark datasets. Our source codes are publicly released at https://github.com/NUSTM/SHARK.

11:00-12:30 (East Foyer / Virtual)
**Modeling Highlighting of Metaphors in Multitask Contrastive Learning Paradigms**
*Meghdut Sengupta, Milad Alshomary, Ingrid Scharlau and Henning Wachsmuth*
Metaphorical language, such as "spending time together", projects meaning from a source domain (here, *money*) to a target domain (*time*). Thereby, it highlights certain aspects of the target domain, such as the *effort* behind the time investment. Highlighting aspects with metaphors (while hiding others) bridges the two domains and is the core of metaphorical meaning construction. For metaphor interpretation, linguistic theories stress that identifying the highlighted aspects is important for a better understanding of metaphors. However, metaphor research in NLP has not yet dealt with the phenomenon of highlighting. In this paper, we introduce the task of identifying the main aspect highlighted in a metaphorical sentence. Given the inherent interaction of source domains and highlighted aspects, we propose two multitask approaches - a joint learning approach and a continual learning approach - based on a finetuned contrastive learning model to jointly predict highlighted aspects and source domains. We further investigate whether (predicted) information about a source domain leads to better performance in predicting the highlighted aspects, and vice versa. Our experiments on an existing corpus suggest that, with the corresponding information, the performance to predict the other improves in terms of model accuracy in predicting highlighted aspects and source domains notably compared to the single-task baselines.

11:00-12:30 (East Foyer / Virtual)
**Enhancing Neural Machine Translation with Semantic Units**
*Langlin Huang, Shuhao Gu, Zhang Zhuocheng and Yang Feng*
Conventional neural machine translation (NMT) models typically use subwords and words as the basic units for model input and comprehension. However, complete words and phrases composed of several tokens are often the fundamental units for expressing semantics, referred to as semantic units. To address this issue, we propose a method Semantic Units for Machine Translation (SU4MT) which models the integral meanings of semantic units within a sentence, and then leverages them to provide a new perspective for understanding the sentence. Specifically, we first propose Word Pair Encoding (WPE), a phrase extraction method to help identify the boundaries of semantic units. Next, we design an Attentive Semantic Fusion (ASF) layer to integrate the semantics of multiple subwords into a single vector: the semantic unit representation. Lastly, the semantic-unit-level sentence representation is concatenated to the token-level one, and they are combined as the input of encoder. Experimental results demonstrate that our method effectively models and leverages semantic-unit-level information and outperforms the strong baselines.

11:00-12:30 (East Foyer / Virtual)
**Survival of the Most Influential Prompts: Efficient Black-Box Prompt Search via Clustering and Pruning**
*Han Zhou, Xingchen Wan, Ivan Vulić and Anna Korhonen*
Prompt-based learning has been an effective paradigm for large pretrained language models (LLM), enabling few-shot or even zero-shot learning. Black-box prompt search has received growing interest recently for its distinctive properties of gradient-free optimization, proven particularly useful and powerful for model-as-a-service usage. However, the discrete nature and the complexity of combinatorial optimization hinder the efficiency of modern black-box approaches. Despite extensive research on search algorithms, the crucial aspect of search space design and optimization has been largely overlooked. In this paper, we first conduct a sensitivity analysis by prompting LLM, revealing that only a small number of tokens exert a disproportionate amount of influence on LLM predictions. Leveraging this insight, we propose the Clustering and Pruning for Efficient Black-box Prompt Search (ClaPS), a simple black-box search method that first clusters and prunes the search space to focus exclusively on influential prompt tokens. By employing even simple search methods within the pruned search space, ClaPS achieves state-of-the-art performance across various tasks and LLMs, surpassing the performance of complex approaches while significantly reducing search costs. Our findings underscore the critical role of search space design and optimization in enhancing both the usefulness and the efficiency of black-box prompt-based learning.

11:00-12:30 (East Foyer / Virtual)
**Harnessing Dataset Cartography for Improved Compositional Generalization in Transformers**
*Osman Batur İnce, Tanin Zeraati, Semih Yagcioglu, Yadollah Yaghoobzadeh, Erkut Erdem and Aykut Erdem*
Neural networks have revolutionized language modeling and excelled in various downstream tasks. However, the extent to which these models achieve compositional generalization comparable to human cognitive abilities remains a topic of debate. While existing approaches in the field have mainly focused on novel architectures and alternative learning paradigms, we introduce a pioneering method harnessing the power of dataset cartography (Swayamdipta et al., 2020). By strategically identifying a subset of compositional generalization data using this approach, we achieve a remarkable improvement in model accuracy, yielding enhancements of up to 10% on CFQ and COGS datasets. Notably, our technique incorporates dataset cartography as a curriculum learning criterion, eliminating the need for hyperparameter tuning while consistently achieving superior performance. Our findings highlight the untapped potential of dataset cartography in unleashing the full capabilities of compositional generalization within Transformer models.

11:00-12:30 (East Foyer / Virtual)
**Cache me if you Can: an Online Cost-aware Teacher-Student framework to Reduce the Calls to Large Language Models**
*Ilias Stogiannidis, Stavros Vassos, Prodromos Malakasiotis and Ion Androutsopoulos*
Prompting Large Language Models (LLMs) performs impressively in zero- and few-shot settings. Hence, small and medium-sized enterprises (SMEs) that cannot afford the cost of creating large task-specific training datasets, but also the cost of pretraining their own LLMs, are increasingly turning to third-party services that allow them to prompt LLMs. However, such services currently require a payment per call, which becomes a significant operating expense (OpEx). Furthermore, customer inputs are often very similar over time, hence SMEs end-up

prompting LLMs with very similar instances. We propose a framework that allows reducing the calls to LLMs by caching previous LLM responses and using them to train a local inexpensive model on the SME side. The framework includes criteria for deciding when to trust the local model or call the LLM, and a methodology to tune the criteria and measure the tradeoff between performance and cost. For experimental purposes, we instantiate our framework with two LLMs, GPT-3.5 or GPT-4, and two inexpensive students, a $k$-NN classifier or a Multi-Layer Perceptron, using two common business tasks, intent recognition and sentiment analysis. Experimental results indicate that significant OpEx savings can be obtained with only slightly lower performance.

11:00-12:30 (East Foyer / Virtual)
**Probing LLMs for Joint Encoding of Linguistic Categories**
*Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger and Ekaterina Shutova*
Large Language Models (LLMs) exhibit impressive performance on a range of NLP tasks, due to the general-purpose linguistic knowledge acquired during pretraining. Existing model interpretability research (Tenney et al., 2019) suggests that a linguistic hierarchy emerges in the LLM layers, with lower layers better suited to solving syntactic tasks and higher layers employed for semantic processing. Yet, little is known about how encodings of different linguistic phenomena interact within the models and to what extent processing of linguistically-related categories relies on the same, shared model representations. In this paper, we propose a framework for testing the joint encoding of linguistic categories in LLMs. Focusing on syntax, we find evidence of joint encoding both at the same (related part-of-speech (POS) classes) and different (POS classes and related syntactic dependency relations) levels of linguistic hierarchy. Our cross-lingual experiments show that the same patterns hold across languages in multilingual LLMs.

11:00-12:30 (East Foyer / Virtual)
**Frequency Balanced Datasets Lead to Better Language Models**
*Rodolfo Joel Zevallos, Mireia Farrús and Núria Bel*
This paper reports on the experiments aimed to improve our understanding of the role of the amount of data required for training attention-based transformer language models. Specifically, we investigate the impact of reducing the immense amounts of required pre-training data through sampling strategies that identify and reduce high-frequency tokens as different studies have indicated that the existence of very high-frequency tokens in pre-training data might bias learning, causing undesired effects. In this light, we describe our sampling algorithm that iteratively assesses token frequencies and removes sentences that contain still high-frequency tokens, eventually delivering a balanced, linguistically correct dataset. We evaluate the results in terms of model perplexity and fine-tuning linguistic probing tasks, NLP downstream tasks as well as more semantic SuperGlue tasks. The results show that pre-training with the resulting balanced dataset allows reducing up to three times the pre-training data.

11:00-12:30 (East Foyer / Virtual)
**Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language**
*Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie and Maarten Sap*
Counterspeech, i.e., responses to counteract potential harms of hateful speech, has become an increasingly popular solution to address online hate speech without censorship. However, properly countering hateful language requires countering and dispelling the underlying inaccurate stereotypes implied by such language. In this work, we draw from psychology and philosophy literature to craft six psychologically inspired strategies to challenge the underlying stereotypical implications of hateful language. We first examine the convincingness of each of these strategies through a user study, and then compare their usages in both human- and machine-generated counterspeech datasets. Our results show that human-written counterspeech uses countering strategies that are more specific to the implied stereotype (e.g., counter examples to the stereotype, external factors about the stereotype's origins), whereas machine-generated counterspeech uses less specific strategies (e.g., generally denouncing the hatefulness of speech). Furthermore, machine generated counterspeech often employs strategies that humans deem less convincing compared to human-produced counterspeech. Our findings point to the importance of accounting for the underlying stereotypical implications of speech when generating counterspeech and for better machine reasoning about anti-stereotypical examples.

11:00-12:30 (East Foyer / Virtual)
**Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs**
*Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal and Monojit Choudhury*
In this position paper, we argue that instead of morally aligning LLMs to specific set of ethical principles, we should infuse generic ethical reasoning capabilities into them so that they can handle value pluralism at a global scale. When provided with an ethical policy, an LLM should be capable of making decisions that are ethically consistent to the policy. We develop a framework that integrates moral dilemmas with moral principles pertaining to different foramlisms of normative ethics, and at different levels of abstractions. Initial experiments with GPT-x models shows that while GPT-4 is a nearly perfect ethical reasoner, the models still have bias towards the moral values of Western and English speaking societies.

11:00-12:30 (East Foyer / Virtual)
**Aligning Predictive Uncertainty with Clarification Questions in Grounded Dialog**
*Kata Naszadi, Putra Manggala and Christof Monz*
Asking for clarification is fundamental to effective collaboration. An interactive artificial agent must know when to ask a human instructor for more information in order to ascertain their goals. Previous work bases the timing of questions on supervised models learned from interactions between humans. Instead of a supervised classification task, we wish to ground the need for questions in the acting agent's predictive uncertainty. In this work, we investigate if ambiguous linguistic instructions can be aligned with uncertainty in neural models. We train an agent using the T5 encoder-decoder architecture to solve the Minecraft Collaborative Building Task and identify uncertainty metrics that achieve better distributional separation between clear and ambiguous instructions. We further show that well-calibrated prediction probabilities benefit the detection of ambiguous instructions. Lastly, we provide a novel empirical analysis on the relationship between uncertainty and dialog history length and highlight an important property that poses a difficulty for detection.

11:00-12:30 (East Foyer / Virtual)
**LLM aided semi-supervision for efficient Extractive Dialog Summarization**
*Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna and Issam H. Laradji*
Generating high-quality summaries for chat dialogs often requires large labeled datasets. We propose a method to efficiently use unlabeled data for extractive summarization of customer-agent dialogs. In our method, we frame summarization as a question-answering problem and use state-of-the-art large language models (LLMs) to generate pseudo-labels for a dialog. We then use these pseudo-labels to fine-tune a chat summarization model, effectively transferring knowledge from the large LLM into a smaller specialized model. We demonstrate our method on the TWEETSUMM dataset, and show that using 10% of the original labelled data set we can achieve 65.9/57.0/61.0 ROUGE-1/-2/-L, whereas the current state-of-the-art trained on the entire training data set obtains 65.16/55.81/64.37 ROUGE-1/-2/-L. In other words, in the worst case (i.e., ROUGE-L) we still effectively retain 94.7% of the performance while using only 10% of the data.

11:00-12:30 (East Foyer / Virtual)
**Test-Time Self-Adaptive Small Language Models for Question Answering**
*Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang and Jong C. Park*
Recent instruction-finetuned large language models (LMs) have achieved notable performances in various tasks, such as question-answering (QA). However, despite their ability to memorize a vast amount of general knowledge across diverse tasks, they might be suboptimal on specific tasks due to their limited capacity to transfer and adapt knowledge to target tasks. Moreover, further finetuning LMs with labeled datasets is often infeasible due to their absence, but it is also questionable if we can transfer smaller LMs having limited knowledge only with unlabeled test data. In this work, we show and investigate the capabilities of smaller self-adaptive LMs, only with unlabeled test data. In particular, we first stochastically generate multiple answers, and then ensemble them while filtering out low-quality samples to mitigate noise from inaccurate labels. Our proposed self-adaption strategy demonstrates significant performance improvements on benchmark QA datasets with higher robustness across diverse prompts, enabling LMs to stay stable. Code is available at: https://github.com/starsuzi/T-SAS.

11:00-12:30 (East Foyer / Virtual)
**A Benchmark for Semi-Inductive Link Prediction in Knowledge Graphs**
*Adrian Kochsiek and Rainer Gemulla*
Semi-inductive link prediction (LP) in knowledge graphs (KG) is the task of predicting facts for new, previously unseen entities based on context information. Although new entities can be integrated by retraining the model from scratch in principle, such an approach is infeasible for large-scale KGs, where retraining is expensive and new entities may arise frequently. In this paper, we propose and describe a large-scale benchmark to evaluate semi-inductive LP models. The benchmark is based on and extends Wikidata5M: It provides transductive, k-shot, and 0-shot LP tasks, each varying the available information from (i) only KG structure, to (ii) including textual mentions, and (iii) detailed descriptions of the entities. We report on a small study of recent approaches and found that semi-inductive LP performance is far from transductive performance on long-tail entities throughout all experiments. The benchmark provides a test bed for further research into integrating context and textual information in semi-inductive LP models.

11:00-12:30 (East Foyer / Virtual)
**Comparing the Evaluation and Production of Loophole Behavior in Humans and Large Language Models**
*Sonia Krishna Murthy, Kiera Maria Parece, Sophie Bridgers, Peng Qian and Tomer Ullman*
In law, lore, and everyday life, loopholes are commonplace. When people exploit a loophole, they understand the intended meaning or goal of another person, but choose to go with a different interpretation. Past and current AI research has shown that artificial intelligence engages in what seems superficially like the exploitation of loopholes, but this is likely anthropomorphization. It remains unclear to what extent current models, especially Large Language Models (LLMs), capture the pragmatic understanding required for engaging in loopholes. We examined the performance of LLMs on two metrics developed for studying loophole behavior in humans: evaluation (ratings of trouble, upset, and humor), and generation (coming up with new loopholes in a given context). We conducted a fine-grained comparison of state-of-the-art LLMs to humans, and find that while many of the models rate loophole behaviors as resulting in less trouble and upset than outright non-compliance (in line with adults), they struggle to recognize the humor in the creative exploitation of loopholes in the way that humans do. Furthermore, only two of the models, GPT 3 and 3.5, are capable of generating loopholes of their own, with GPT3.5 performing closest to the human baseline.

11:00-12:30 (East Foyer / Virtual)
**Are Language Models Worse than Humans at Following Prompts? It's Complicated**
*Albert Webson, Alyssa Marie Loo, Qinan Yu and Ellie Pavlick*
Prompts have been the center of progress in advancing language models' zero-shot and few-shot performance. However, recent work finds that models can perform surprisingly well when given intentionally irrelevant or misleading prompts. Such results may be interpreted as evidence that model behavior is not "human like'. In this study, we challenge a central assumption in such work: that humans would perform badly when given pathological instructions. We find that humans are able to reliably ignore irrelevant instructions and thus, like models, perform well on the underlying task despite an apparent lack of signal regarding the task they are being asked to do. However, when given deliberately misleading instructions, humans follow the instructions faithfully, whereas models do not. Our findings caution that future research should not idealize human behaviors as a monolith and should not train or evaluate models to mimic assumptions about these behaviors without first validating humans' behaviors empirically.

11:00-12:30 (East Foyer / Virtual)
**Automatic Pronunciation Assessment - A Review**
*Yassine El Kheir, Ahmed Ali and Shammur Absar Chowdhury*
Pronunciation assessment and its application in computer-aided pronunciation training (CAPT) have seen impressive progress in recent years. With the rapid growth in language processing and deep learning over the past few years, there is a need for an updated review. In this paper, we review methods employed in pronunciation assessment for both phonemic and prosodic. We categorize the main challenges observed in prominent research trends, and highlight existing limitations, and available resources. This is followed by a discussion of the remaining challenges and possible directions for future work.

11:00-12:30 (East Foyer / Virtual)
**Learning to Follow Object-Centric Image Editing Instructions Faithfully**
*Tuhin Chakrabarty, Kanishk Singh, Arkadiy Saakyan and Smaranda Muresan*
Natural language instructions are a powerful interface for editing the outputs of text-to-image diffusion models. However, several challenges need to be addressed: 1) underspecification (the need to model the implicit meaning of instructions) 2) grounding (the need to localize where the edit has to be performed), 3) faithfulness (the need to preserve the elements of the image not affected by the edit instruction). Current approaches focusing on image editing with natural language instructions rely on automatically generated paired data, which, as shown in our investigation, is noisy and sometimes nonsensical, exacerbating the above issues. Building on recent advances in segmentation, Chain-of-Thought prompting, and visual question answering, we significantly improve the quality of the paired data. In addition, we enhance the supervision signal by highlighting parts of the image that need to be changed by the instruction. The model fine-tuned on the improved data is capable of performing fine-grained object-centric edits better than state-of-the-art baselines, mitigating the problems outlined above, as shown by automatic and human evaluations. Moreover, our model is capable of generalizing to domains unseen during training, such as visual metaphors.

11:00-12:30 (East Foyer / Virtual)
**Connecting the Dots: What Graph-Based Text Representations Work Best for Text Classification using Graph Neural Networks?**
*Margarita Bugueño and Gerard de Melo*
Given the success of Graph Neural Networks (GNNs) for structure-aware machine learning, many studies have explored their use for text classification, but mostly in specific domains with limited data characteristics. Moreover, some strategies prior to GNNs relied on graph mining and classical machine learning, making it difficult to assess their effectiveness in modern settings. This work extensively investigates graph

representation methods for text classification, identifying practical implications and open challenges. We compare different graph construction schemes using a variety of GNN architectures and setups across five datasets, encompassing short and long documents as well as unbalanced scenarios in diverse domains. Two Transformer-based large language models are also included to complement the study. The results show that i) although the effectiveness of graphs depends on the textual input features and domain, simple graph constructions perform better the longer the documents are, ii) graph representations are especially beneficial for longer documents, outperforming Transformer-based models, iii) graph methods are particularly efficient for solving the task.

11:00-12:30 (East Foyer / Virtual)
**Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship?**
*Sathvik Nair and Philip Resnik*
An important assumption that comes with using LLMs on psycholinguistic data has gone unverified. LLM-based predictions are based on subword tokenization, not decomposition of words into morphemes. Does that matter? We carefully test this by comparing surprisal estimates using orthographic, morphological, and BPE tokenization against reading time data. Our results replicate previous findings and provide evidence that *in the aggregate*, predictions using BPE tokenization do not suffer relative to morphological and orthographic segmentation. However, a finer-grained analysis points to potential issues with relying on BPE-based tokenization, as well as providing promising results involving morphologically-aware surprisal estimates and suggesting a new method for evaluating morphological prediction.

11:00-12:30 (East Foyer / Virtual)
**Robustness of Named-Entity Replacements for In-Context Learning**
*Saeed Goodarzi, Nikhil Kagita, Dennis Minn, Shufan Wang, Roberto Dessi, Shubham Toshniwal, Adina Williams, Jack Lanchantin and Koustuv Sinha*
A key feature of modern large language models (LLMs) is their ability to perform in-context learning, a prompting technique where query-answer demonstrations are shown before the final query. This allows for generalization to novel distributions at inference time where the LLM can learn new rules without parameter updates. However, the choice of demonstrations and their relationship to a particular query can have a profound impact on model accuracy, raising concerns about the true in-context generalization capabilities (Zhao et al., 2021). In this work, we explore the robustness of the in-context learning paradigm by focusing on entities. In particular, we seek to understand the robustness of LLM in-context learning with respect to named entity replacements. We discover a significant variance in downstream performance based on the choice of the named entities, across three popular reasoning tasks and two popular LLMs. Specifically, model accuracy on the test sets can fluctuate between -2.7 to +8.0 points depending on the choice of named entity replacements. Our analysis exposes the sensitivity of LLM in-context learning with respect to named entities, and offers a simple recipe to improve test performance by hyper-parameter tuning the named entities for a given dataset. Code and datasets for reproducing the results are publicly available.

11:00-12:30 (East Foyer / Virtual)
**Co$^2$PT: Mitigating Bias in Pre-trained Language Models through Counterfactual Contrastive Prompt Tuning**
*Xiangjue Dong, Ziwei Zhu, Zhuoer Wang, Maria Teleki and James Caverlee*
Pre-trained Language Models are widely used in many important real-world applications. However, recent studies show that these models can encode social biases from large pre-training corpora and even amplify biases in downstream applications. To address this challenge, we propose Co$^2$PT, an efficient and effective *debias-while-prompt tuning* method for mitigating biases via counterfactual contrastive prompt tuning on downstream tasks. Our experiments conducted on three extrinsic bias benchmarks demonstrate the effectiveness of Co$^2$PT on bias mitigation during the prompt tuning process and its adaptability to existing upstream debiased language models. These findings indicate the strength of Co$^2$PT and provide promising avenues for further enhancement in bias mitigation on downstream tasks.

11:00-12:30 (East Foyer / Virtual)
**Learning Easily Updated General Purpose Text Representations with Adaptable Task-Specific Prefix**
*Kuan-Hao Huang, Liang Tan, Rui Hou, Sinong Wang, Amjad Almahairi and Ruty Rinott*
Many real-world applications require making multiple predictions from the same text. Fine-tuning a large pre-trained language model for each downstream task causes computational burdens in the inference time due to several times of forward passes. To amortize the computational cost, freezing the language model and building lightweight models for downstream tasks based on fixed text representations are common solutions. Accordingly, how to learn fixed but general text representations that can generalize well to unseen downstream tasks becomes a challenge. Previous works have shown that the generalizability of representations can be improved by fine-tuning the pre-trained language model with some source tasks in a multi-tasking way. In this work, we propose a prefix-based method to learn the fixed text representations with source tasks. We learn a task-specific prefix for each source task independently and combine them to get the final representations. Our experimental results show that prefix-based training performs better than multi-tasking training and can update the text representations at a smaller computational cost than multi-tasking training.

11:00-12:30 (East Foyer / Virtual)
**Dimensions of Online Conflict: Towards Modeling Agonism**
*Matt Canute, Mali Jin, hannah holtzclaw, Alberto Lusoli, Philippa R Adams, Mugdha Pandya, Maite Taboada, Diana Maynard and Wendy Hui Kyong Chun*
Agonism plays a vital role in democratic dialogue by fostering diverse perspectives and robust discussions. Within the realm of online conflict there is another type: hateful antagonism, which undermines constructive dialogue. Detecting conflict online is central to platform moderation and monetization. It is also vital for democratic dialogue, but only when it takes the form of agonism. To model these two types of conflict, we collected Twitter conversations related to trending controversial topics. We introduce a comprehensive annotation schema for labelling different dimensions of conflict in the conversations, such as the source of conflict, the target, and the rhetorical strategies deployed. Using this schema, we annotated approximately 4,000 conversations with multiple labels. We then train both logistic regression and transformer-based models on the dataset, incorporating context from the conversation, including the number of participants and the structure of the interactions. Results show that contextual labels are helpful in identifying conflict and make the models robust to variations in topic. Our research contributes a conceptualization of different dimensions of conflict, a richly annotated dataset, and promising results that can contribute to content moderation.

11:00-12:30 (East Foyer / Virtual)
**Error Detection for Text-to-SQL Semantic Parsing**
*Shijie Chen, Ziru Chen, Huan Sun and Yu Su*
Despite remarkable progress in text-to-SQL semantic parsing in recent years, the performance of existing parsers is still far from perfect. Specifically, modern text-to-SQL parsers based on deep learning are often over-confident, thus casting doubt on their trustworthiness when deployed for real use. In this paper, we propose a parser-independent error detection model for text-to-SQL semantic parsing. Using a language model of code as its bedrock, we enhance our error detection model with graph neural networks that learn structural features of both natural language questions and SQL queries. We train our model on realistic parsing errors collected from a cross-domain setting, which

leads to stronger generalization ability. Experiments with three strong text-to-SQL parsers featuring different decoding mechanisms show that our approach outperforms parser-dependent uncertainty metrics. Our model could also effectively improve the performance and usability of text-to-SQL semantic parsers regardless of their architectures.

11:00-12:30 (East Foyer / Virtual)
**Long-Horizon Dialogue Understanding for Role Identification in the Game of Avalon with Large Language Models**
*Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Sharon Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Michael Lewis and Katia P. Sycara*
Deception and persuasion play a critical role in long-horizon dialogues between multiple parties, especially when the interests, goals, and motivations of the participants are not aligned. Such complex tasks pose challenges for current Large Language Models (LLM) as deception and persuasion can easily mislead them, especially in long-horizon multi-party dialogues. To this end, we explore the game of Avalon: The Resistance, a social deduction game in which players must determine each other's hidden identities to complete their team's objective. We introduce an online testbed and a dataset containing 20 carefully collected and labeled games among human players that exhibit long-horizon deception in a cooperative-competitive setting. We discuss the capabilities of LLMs to utilize deceptive long-horizon conversations between six human players to determine each player's goal and motivation. Particularly, we discuss the multimodal integration of the chat between the players and the game's state that grounds the conversation, providing further insights into the true player identities. We find that even current state-of-the-art LLMs do not reach human performance, making our dataset a compelling benchmark to investigate the decision-making and language-processing capabilities of LLMs. Our dataset and online testbed can be found at our project website: https://sstepput.github.io/Avalon-NLU/

11:00-12:30 (East Foyer / Virtual)
**GPT Deciphering Fedspeak: Quantifying Dissent Among Hawks and Doves**
*Denis Peskoff, Adam Visokay, Sander V Schulhoff, Benjamin Wachspress, Alan Blinder and Brandon M. Stewart*
Markets and policymakers around the world hang on the consequential monetary policy decisions made by the Federal Open Market Committee (FOMC). Publicly available textual documentation of their meetings provides insight into members' attitudes about the economy. We use GPT-4 to quantify dissent among members on the topic of inflation. We find that transcripts and minutes reflect the diversity of member views about the macroeconomic outlook in a way that is lost or omitted from the public statements. In fact, diverging opinions that shed light upon the committee's "true" attitudes are almost entirely omitted from the final statements. Hence, we argue that forecasting FOMC sentiment based solely on statements will not sufficiently reflect dissent among the hawks and doves.

11:00-12:30 (East Foyer / Virtual)
**Beyond Layout Embedding: Layout Attention with Gaussian Biases for Structured Document Understanding**
*Xi Zhu, Xue Han, Shuyuan Peng, Shuo Lei, Chao Deng and Junlan Feng*
Effectively encoding layout information is a central problem in structured document understanding. Most existing methods rely heavily on millions of trainable parameters to learn the layout features of each word from Cartesian coordinates. However, two unresolved questions remain: (1) Is the Cartesian coordinate system the optimal choice for layout modeling? (2) Are massive learnable parameters truly necessary for layout representation? In this paper, we address these questions by proposing Layout Attention with Gaussian Biases (LAGaBi): Firstly, we find that polar coordinates provide a superior choice over Cartesian coordinates as they offer a measurement of both distance and angle between word pairs, capturing relative positions more effectively. Furthermore, by feeding the distances and angles into 2-D Gaussian kernels, we model intuitive inductive layout biases, i.e., the words closer within a document should receive more attention, which will act as the attention biases to revise the textual attention distribution. LAGaBi is model-agnostic and language-independent, which can be applied to a range of transformer-based models, such as the text pre-training models from the BERT series and the LayoutLM series that incorporate visual features. Experimental results on three widely used benchmarks demonstrate that, despite reducing the number of layout parameters from millions to 48, LAGaBi achieves competitive or even superior performance.

11:00-12:30 (East Foyer / Virtual)
**$R^3$ Prompting: Review, Rephrase and Resolve for Chain-of-Thought Reasoning in Large Language Models under Noisy Context**
*Qingyuan Tian, Hanlun Zhu, Lei Wang, Yang Li and Yunshi Lan*
With the help of Chain-of-Thought (CoT) prompting, Large Language Models (LLMs) have achieved remarkable performance on various reasoning tasks. However, most of them have been evaluated under noise-free context and the dilemma for LLMs to produce inaccurate results under the noisy context has not been fully investigated. Existing studies utilize trigger sentences to encourage LLMs to concentrate on the relevant information but the trigger has limited effect on final answer prediction. Inspired by interactive CoT method, where intermediate reasoning steps are promoted by multiple rounds of interaction between users and LLMs, we propose a novel prompting method, namely $R^3$ prompting, for CoT reasoning under noisy context. Specifically, $R^3$ prompting interacts with LLMs to perform key sentence extraction, variable declaration and answer prediction, which corresponds to a thought process of reviewing, rephrasing and resolving. The responses generated at the last interaction will perform as hints to guide toward the responses of the next interaction. Our experiments show that $R^3$ prompting significantly outperforms existing CoT prompting methods on five reasoning tasks under noisy context. With GPT-3.5-turbo, we observe 3.7% accuracy improvement on average on the reasoning tasks under noisy context compared to the most competitive prompting baseline. More analyses and ablation studies show the robustness and generalization of $R^3$ prompting method in solving reasoning tasks in LLMs under noisy context.

11:00-12:30 (East Foyer / Virtual)
**TRAMS: Training-free Memory Selection for Long-range Language Modeling**
*Haofei Yu, Cunxiang Wang, Yue Zhang and Wei Bi*
The Transformer architecture is crucial for numerous AI models, but it still faces challenges in long-range language modeling. Though several specific transformer architectures have been designed to tackle issues of long-range dependencies, existing methods like Transformer-XL are plagued by a high percentage of ineffective memories. In this study, we present a plug-and-play strategy, known as TRAining-free Memory Selection (TRAMS), that selects tokens participating in attention calculation based on one simple metric. This strategy allows us to keep tokens that are likely to have a high attention score with the current queries and ignore the other ones. We have tested our approach on the word-level benchmark (WikiText-103) and the character-level benchmark (enwik8), and the results indicate an improvement without having additional training or adding additional parameters.

11:00-12:30 (East Foyer / Virtual)
**Large Language Models Are Better Adversaries: Exploring Generative Clean-Label Backdoor Attacks Against Text Classifiers**
*Wencong You, Zayd Hammoudeh and Daniel Lowd*
Backdoor attacks manipulate model predictions by inserting innocuous triggers into training and test data. We focus on more realistic and more challenging clean-label attacks where the adversarial training examples are correctly labeled. Our attack, LLMBkd, leverages language models to automatically insert diverse style-based triggers into texts. We also propose a poison selection technique to improve the effec-

tiveness of both LLMBkd as well as existing textual backdoor attacks. Lastly, we describe REACT, a baseline defense to mitigate backdoor attacks via antidote training examples. Our evaluations demonstrate LLMBkd's effectiveness and efficiency, where we consistently achieve high attack success rates across a wide range of styles with little effort and no model training.

11:00-12:30 (East Foyer / Virtual)
**A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection**
*Shiping Yang, Renliang Sun and Xiaojun Wan*
Large Language Models (LLMs) have shown their ability to collaborate effectively with humans in real-world scenarios. However, LLMs are apt to generate hallucinations, i.e., makeup incorrect text and unverified information, which can cause significant damage when deployed for mission-critical tasks. In this paper, we propose a self-check approach based on reverse validation to detect factual errors automatically in a zero-resource fashion. To facilitate future studies and assess different methods, we construct a hallucination detection benchmark named PHD, which is generated by ChatGPT and annotated by human annotators. Contrasting previous studies of zero-resource hallucination detection, our method and benchmark concentrate on passage-level detection instead of sentence-level. We empirically evaluate our method and existing zero-resource detection methods on two datasets. The experimental results demonstrate that the proposed method considerably outperforms the baselines while costing fewer tokens and less time. Furthermore, we manually analyze some hallucination cases that LLM failed to capture, revealing the shared limitation of zero-resource methods.

11:00-12:30 (East Foyer / Virtual)
**DiFair: A Benchmark for Disentangled Assessment of Gender Knowledge and Bias**
*Mahdi Zakizadeh, Kaveh Eskandari Miandoab and Mohammad Taher Pilehvar*
Numerous debiasing techniques have been proposed to mitigate the gender bias that is prevalent in pretrained language models. These are often evaluated on datasets that check the extent to which the model is gender-neutral in its predictions. Importantly, this evaluation protocol overlooks the possible adverse impact of bias mitigation on useful gender knowledge. To fill this gap, we propose **DiFair**, a manually curated dataset based on masked language modeling objectives. **DiFair** allows us to introduce a unified metric, *gender invariance score*, that not only quantifies a model's biased behavior, but also checks if useful gender knowledge is preserved. We use **DiFair** as a benchmark for a number of widely-used pretained language models and debiasing techniques. Experimental results corroborate previous findings on the existing gender biases, while also demonstrating that although debiasing techniques ameliorate the issue of gender bias, this improvement usually comes at the price of lowering useful gender knowledge of the model.

11:00-12:30 (East Foyer / Virtual)
**Improving Seq2Seq Grammatical Error Correction via Decoding Interventions**
*Houquan Zhou, Yumeng Liu, Zhenghua Li, Min Zhang, Bo Zhang, Chen Li, Ji Zhang and Fei Huang*
The sequence-to-sequence (Seq2Seq) approach has recently been widely used in grammatical error correction (GEC) and shows promising performance. However, the Seq2Seq GEC approach still suffers from two issues. First, a Seq2Seq GEC model can only be trained on parallel data, which, in GEC task, is often noisy and limited in quantity. Second, the decoder of a Seq2Seq GEC model lacks an explicit awareness of the correctness of the token being generated. In this paper, we propose a unified decoding intervention framework that employs an external critic to assess the appropriateness of the token to be generated incrementally, and then dynamically influence the choice of the next token. We discover and investigate two types of critics: a pre-trained left-to-right language model critic and an incremental target-side grammatical error detector critic. Through extensive experiments on English and Chinese datasets, our framework consistently outperforms strong baselines and achieves results competitive with state-of-the-art methods.

11:00-12:30 (East Foyer / Virtual)
**FFAEval: Evaluating Dialogue System via Free-For-All Ranking**
*Zeyao Ma, Zijun Yao, Jing Zhang, Jifan Yu, Xiaohan Zhang, Juanzi Li and Jie Tang*
Evaluating open-domain dialogue systems is currently an open question. Automatic evaluation metrics have shown poor correlation with human assessment in dialogue generation tasks. Human evaluation, which involves annotators for multi-dimension scoring, is trustworthy but time-consuming. In this work, we propose FFAEval, a reliable and efficient human evaluation framework using Free-For-All ranking approach. By sharing the dialogue history, the framework enables annotators to converse with multiple dialogue systems simultaneously in a single-blind, multi-turn manner. The subsequent free-for-all allows annotators to select the most favourable model in each turn from among all the participating dialogue systems. The final performance of each model is represented by calculating the TrueSkill score derived from the free-for-all competition. Our empirical study on English and Chinese dialogue systems demonstrates that FFAEval achieves a strong correlation with score-based human assessment compared to existing evaluation methods. We further prove the efficiency and stability of our framework in additional experiments. The source code and data are available on Github.

11:00-12:30 (East Foyer / Virtual)
**Impact of Co-occurrence on Factual Knowledge of Large Language Models**
*Cheongwoong Kang and Jaesik Choi*
Large language models (LLMs) often make factually incorrect responses despite their success in various applications. In this paper, we hypothesize that relying heavily on simple co-occurrence statistics of the pre-training corpora is one of the main factors that cause factual errors. Our results reveal that LLMs are vulnerable to the co-occurrence bias, defined as preferring frequently co-occurred words over the correct answer. Consequently, LLMs struggle to recall facts whose subject and object rarely co-occur in the pre-training dataset although they are seen during finetuning. We show that co-occurrence bias remains despite scaling up model sizes or finetuning. Therefore, we suggest finetuning on a debiased dataset to mitigate the bias by filtering out biased samples whose subject-object co-occurrence count is high. Although debiased finetuning allows LLMs to memorize rare facts in the training set, it is not effective in recalling rare facts unseen during finetuning. Further research in mitigation will help build reliable language models by preventing potential errors. The code is available at https://github.com/CheongWoong/impact_of_cooccurrence.

11:00-12:30 (East Foyer / Virtual)
**PCMID: Multi-Intent Detection through Supervised Prototypical Contrastive Learning**
*Yurun Song, Junchen Zhao, Spencer B. Koehler, Amir Abdullah and Ian Harris*
Intent detection is a major task in Natural Language Understanding (NLU) and is the component of dialogue systems for interpreting users' intentions based on their utterances. Many works have explored detecting intents by assuming that each utterance represents only a single intent. Such systems have achieved very good results; however, intent detection is a far more challenging task in typical real-world scenarios, where each user utterance can be highly complex and express multiple intents. Therefore, in this paper, we propose PCMID, a novel Multi-Intent Detection framework enabled by Prototypical Contrastive Learning under a supervised setting. The PCMID model can learn multiple semantic representations of a given user utterance under the context of different intent labels in an optimized semantic space. Our experiments show that PCMID achieves the current state-of-the-art performance on both multiple public benchmark datasets and a private real-world dataset for the multi-intent detection task.

11:00-12:30 (East Foyer / Virtual)
**Beneath Surface Similarity: Large Language Models Make Reasonable Scientific Analogies after Structure Abduction**
*Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao and Deqing Yang*
The vital role of analogical reasoning in human cognition allows us to grasp novel concepts by linking them with familiar ones through shared relational structures. Despite the attention previous research has given to word analogies, this work suggests that Large Language Models (LLMs) often overlook the structures that underpin these analogies, raising questions about the efficacy of word analogies as a measure of analogical reasoning skills akin to human cognition. In response to this, our paper introduces a task of analogical structure abduction, grounded in cognitive psychology, designed to abduce structures that form an analogy between two systems. In support of this task, we establish a benchmark called SCAR, containing 400 scientific analogies from 13 distinct fields, tailored for evaluating analogical reasoning with structure abduction. The empirical evidence underlines the continued challenges faced by LLMs, including ChatGPT and GPT-4, in mastering this task, signifying the need for future exploration to enhance their abilities.

11:00-12:30 (East Foyer / Virtual)
**Incorporating Syntactic Knowledge into Pre-trained Language Model using Optimization for Overcoming Catastrophic Forgetting**
*Ran Iwamoto, Issei Yoshida, Hiroshi Kanayama, Takuya Ohko and Masayasu Muraoka*
Syntactic knowledge is invaluable information for many tasks which handle complex or long sentences, but typical pre-trained language models do not contain sufficient syntactic knowledge. Thus it results in failures in downstream tasks that require syntactic knowledge. In this paper, we explore additional training to incorporate syntactic knowledge to a language model. We designed four pre-training tasks that learn different syntactic perspectives. For adding new syntactic knowledge and keeping a good balance between the original and additional knowledge, we addressed the problem of catastrophic forgetting that prevents the model from keeping semantic information when the model learns additional syntactic knowledge. We demonstrated that additional syntactic training produced consistent performance gains while clearly avoiding catastrophic forgetting.

11:00-12:30 (East Foyer / Virtual)
**A Novel Contrastive Learning Method for Clickbait Detection on RoCliCo: A Romanian Clickbait Corpus of News Articles**
*Daria Mihaela Broscoteanu and Radu Tudor Ionescu*
To increase revenue, news websites often resort to using deceptive news titles, luring users into clicking on the title and reading the full news. Clickbait detection is the task that aims to automatically detect this form of false advertisement and avoid wasting the precious time of online users. Despite the importance of the task, to the best of our knowledge, there is no publicly available clickbait corpus for the Romanian language. To this end, we introduce a novel Romanian Clickbait Corpus (RoCliCo) comprising 8,313 news samples which are manually annotated with clickbait and non-clickbait labels. Furthermore, we conduct experiments with four machine learning methods, ranging from handcrafted models to recurrent and transformer-based neural networks, to establish a line-up of competitive baselines. We also carry out experiments with a weighted voting ensemble. Among the considered baselines, we propose a novel BERT-based contrastive learning model that learns to encode news titles and contents into a deep metric space such that titles and contents of non-clickbait news have high cosine similarity, while titles and contents of clickbait news have low cosine similarity. Our data set and code to reproduce the baselines are publicly available for download at https://github.com/dariabroscoteanu/RoCliCo.

11:00-12:30 (East Foyer / Virtual)
**Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search**
*Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen and Hongjin Qian*
Precisely understanding users' contextual search intent has been an important challenge for conversational search. As conversational search sessions are much more diverse and long-tailed, existing methods trained on limited data still show unsatisfactory effectiveness and robustness to handle real conversational search scenarios. Recently, large language models (LLMs) have demonstrated amazing capabilities for text generation and conversation understanding. In this work, we present a simple yet effective prompting framework, called LLM4CS, to leverage LLMs as a text-based search intent interpreter to help conversational search. Under this framework, we explore three prompting methods to generate multiple query rewrites and hypothetical responses, and propose to aggregate them into an integrated representation that can robustly represent the user's real contextual search intent. Extensive automatic evaluations and human evaluations on three widely used conversational search benchmarks, including CAsT-19, CAsT-20, and CAsT-21, demonstrate the remarkable performance of our simple LLM4CS framework compared with existing methods and even using human rewrites. Our findings provide important evidence to better understand and leverage LLMs for conversational search.

11:00-12:30 (East Foyer / Virtual)
**Conversational Recommender System and Large Language Model Are Made for Each Other in E-commerce Pre-sales Dialogue**
*Yuanxing Liu, Weinan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li and Wanxiang Che*
E-commerce pre-sales dialogue aims to understand and elicit user needs and preferences for the items they are seeking so as to provide appropriate recommendations. Conversational recommender systems (CRSs) learn user representation and provide accurate recommendations based on dialogue context, but rely on external knowledge. Large language models (LLMs) generate responses that mimic pre-sales dialogues after fine-tuning, but lack domain-specific knowledge for accurate recommendations. Intuitively, the strengths of LLM and CRS in E-commerce pre-sales dialogues are complementary, yet no previous work has explored this. This paper investigates the effectiveness of combining LLM and CRS in E-commerce pre-sales dialogues, proposing two collaboration methods: CRS assisting LLM and LLM assisting CRS. We conduct extensive experiments on a real-world dataset of E-commerce pre-sales dialogues. We analyze the impact of two collaborative approaches with two CRSs and two LLMs on four tasks of E-commerce pre-sales dialogue. We find that collaborations between CRS and LLM can be very effective in some cases.

11:00-12:30 (East Foyer / Virtual)
**A Causal View of Entity Bias in (Large) Language Models**
*Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou and Muhao Chen*
Entity bias widely affects pretrained (large) language models, causing them to rely on (biased) parametric knowledge to make unfaithful predictions. Although causality-inspired methods have shown great potential to mitigate entity bias, it is hard to precisely estimate the parameters of underlying causal models in practice. The rise of black-box LLMs also makes the situation even worse, because of their inaccessible parameters and uncalibrated logits. To address these problems, we propose a specific structured causal model (SCM) whose parameters are comparatively easier to estimate. Building upon this SCM, we propose causal intervention techniques to mitigate entity bias for both white-box and black-box settings. The proposed causal intervention perturbs the original entity with neighboring entities. This intervention reduces specific biasing information pertaining to the original entity while still preserving sufficient semantic information from similar entities. Under the white-box setting, our training-time intervention improves OOD performance of PLMs on relation extraction (RE) and machine reading comprehension (MRC) by 5.7 points and by 9.1 points, respectively. Under the black-box setting, our in-context intervention effectively reduces the entity-based knowledge conflicts of GPT-3.5, achieving up to 20.5 points of improvement of exact match accuracy on MRC and up to 17.6 points of reduction in memorization ratio on RE.

11:00-12:30 (East Foyer / Virtual)
**xDial-Eval: A Multilingual Open-Domain Dialogue Evaluation Benchmark**
*Chen Zhang, Luis Fernando D'Haro, Chengguang Tang, Ke Shi, Guohua Tang and Haizhou Li*
Recent advancements in reference-free learned metrics for open-domain dialogue evaluation have been driven by the progress in pre-trained language models and the availability of dialogue data with high-quality human annotations. However, current studies predominantly concentrate on English dialogues, and the generalization of these metrics to other languages has not been fully examined. This is largely due to the absence of a multilingual dialogue evaluation benchmark. To address the issue, we introduce xDial-Eval, built on top of open-source English dialogue evaluation datasets. xDial-Eval includes 12 turn-level and 6 dialogue-level English datasets, comprising 14930 annotated turns and 8691 annotated dialogues respectively. The English dialogue data are extended to nine other languages with commercial machine translation systems. On xDial-Eval, we conduct comprehensive analyses of previous BERT-based metrics and the recently-emerged large language models. Lastly, we establish strong self-supervised and multilingual baselines. In terms of average Pearson correlations over all datasets and languages, the best baseline outperforms OpenAI's ChatGPT by absolute improvements of 6.5% and 4.6% at the turn and dialogue levels respectively, albeit with much fewer parameters. The data and code are publicly available at https://github.com/e0397123/xDial-Eval.

11:00-12:30 (East Foyer / Virtual)
**Intuitive Multilingual Audio-Visual Speech Recognition with a Single-Trained Model**
*Joanna Hong, Se Jin Park and Yong Man Ro*
We present a novel approach to multilingual audio-visual speech recognition tasks by introducing a single model on a multilingual dataset. Motivated by a human cognitive system where humans can intuitively distinguish different languages without any conscious effort or guidance, we propose a model that can capture which language is given as an input speech by distinguishing the inherent similarities and differences between languages. To do so, we design a prompt fine-tuning technique into the largely pre-trained audio-visual representation model so that the network can recognize the language class as well as the speech with the corresponding language. Our work contributes to developing robust and efficient multilingual audio-visual speech recognition systems, reducing the need for language-specific models.

11:00-12:30 (East Foyer / Virtual)
**Toxicity in Multilingual Machine Translation at Scale**
*Marta R. Costa-jussà, Eric Michael Smith, Christophe Ropers, Daniel Edward Licht, Jean Maillard, Javier Ferrando and Carlos Escolano*
Machine Translation systems can produce different types of errors, some of which are characterized as critical or catastrophic due to the specific negative impact that they can have on users. In this paper we focus on one type of critical error: added toxicity. We evaluate and analyze added toxicity when translating a large evaluation dataset (HOLISTICBIAS, over 472k sentences, covering 13 demographic axes) from English into 164 languages. An automatic toxicity evaluation shows that added toxicity across languages varies from 0% to 5%. The output languages with the most added toxicity tend to be low-resource ones, and the demographic axes with the most added toxicity include sexual orientation, gender and sex, and ability. We also perform human evaluation on a subset of 8 translation directions, confirming the prevalence of true added toxicity. We use a measurement of the amount of source contribution to the translation, where a low source contribution implies hallucination, to interpret what causes toxicity. Making use of the input attributions allows us to explain toxicity, because the source contributions significantly correlate with toxicity for 84% of languages studied. Given our findings, our recommendations to reduce added toxicity are to curate training data to avoid mistranslations, mitigate hallucination and check unstable translations.

11:00-12:30 (East Foyer / Virtual)
**Explain-then-translate: an analysis on improving program translation with self-generated explanations**
*Zilu Tang, Mayank Agarwal, Alexander G Shypula, Bailin Wang, Derry Wijaya, Jie Chen and Yoon Kim*
This work explores the use of self-generated natural language explanations as an intermediate step for code-to-code translation with language models. Across three types of explanations and 19 programming languages constructed from the MultiPL-E dataset, we find the explanations to be particularly effective in the zero-shot case, improving performance by 12% on average. Improvements with natural language explanations are particularly pronounced on difficult programs. We release our dataset, code, and canonical solutions in all 19 languages.

11:00-12:30 (East Foyer / Virtual)
**CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models**
*Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent and Byron C Wallace*
We propose CHiLL (Crafting High-Level Latents), an approach for natural-language specification of features for linear models. CHiLL prompts LLMs with expert-crafted queries to generate interpretable features from health records. The resulting noisy labels are then used to train a simple linear classifier. Generating features based on queries to an LLM can empower physicians to use their domain expertise to craft features that are clinically meaningful for a downstream task of interest, without having to manually extract these from raw EHR. We are motivated by a real-world risk prediction task, but as a reproducible proxy, we use MIMIC-III and MIMIC-CXR data and standard predictive tasks (e.g., 30-day readmission) to evaluate this approach. We find that linear models using automatically extracted features are comparably performant to models using reference features, and provide greater interpretability than linear models using "Bag-of-Words" features. We verify that learned feature weights align well with clinical expectations.

11:00-12:30 (East Foyer / Virtual)
**USB: A Unified Summarization Benchmark Across Tasks and Domains**
*Kundan Krishna, Prakhar Gupta, Sanjana Ramprasad, Byron C Wallace, Jeffrey P. Bigham and Zachary Chase Lipton*
While the NLP community has produced numerous summarization benchmarks, none provide the rich annotations required to simultaneously address many important problems related to control and reliability. We introduce a Wikipedia-derived benchmark, complemented by a rich set of crowd-sourced annotations, that supports 8 interrelated tasks: (i) extractive summarization; (ii) abstractive summarization; (iii) topic-based summarization; (iv) compressing selected sentences into a one-line summary; (v) surfacing evidence for a summary sentence; (vi) predicting the factual accuracy of a summary sentence; (vii) identifying unsubstantiated spans in a summary sentence; (viii) correcting factual errors in summaries. We compare various methods on this benchmark and discover that on multiple tasks, moderately-sized fine-tuned models consistently outperform much larger few-shot prompted language models. For factuality-related tasks, we also evaluate existing heuristics to create training data and find that training on them results in worse performance than training on $20\times$ less human-labeled data. Our articles draw from 6 domains, facilitating cross-domain analysis. On some tasks, the amount of training data matters more than the domain where it comes from, while for other tasks training specifically on data from the target domain, even if limited, is more beneficial.

11:00-12:30 (East Foyer / Virtual)
**MacLaSa: Multi-Aspect Controllable Text Generation via Efficient Sampling from Compact Latent Space**
*Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, Xueqi Cheng and Tat-Seng Chua*
Multi-aspect controllable text generation aims to generate fluent sentences that possess multiple desired attributes simultaneously. Traditional methods either require expensive iteration / searching within the discrete text space during the decoding stage, or train separate controllers for each aspect, resulting in a degradation of text quality due to the discrepancy between different aspects. To address these limitations, we introduce a novel approach for **M**ulti-**a**spect **c**ontrol, namely MacLaSa, that estimates compact **La**tent space for multiple aspects, and

performs efficient **Sa**mpling with a fast sampler. To eliminate the domain discrepancies between different aspects, we first utilize a variational autoencoder (VAE) network to map text sequences from various data sources into close latent representations. The estimated latent space enables the formulation of joint energy-based models and the plugging in of arbitrary attribute discriminators to achieve multi-aspect control. Afterwards, we draw latent samples with a fast sampler based on ordinary differential equations and feed sampled examples to the VAE decoder to produce target text sequences. Experimental results demonstrate that MacLaSa outperforms strong baselines on both attribute relevance and textual quality while maintaining a high inference speed.

11:00-12:30 (East Foyer / Virtual)
**A Zero-Shot Language Agent for Computer Control with Structured Reflection**
*Tao Li, Gang Li, Zhiwei Deng, Bryan Wang and Yang Li*
Large language models (LLMs) have shown increasing capacity at planning and executing a high-level goal in a live computer environment (e.g. MiniWoB++). To perform a task, recent works often require a model to learn from trace examples of the task via either supervised learning or few/many-shot prompting. Without these trace examples, it remains a challenge how an agent can autonomously learn and improve its control on a computer, which limits the ability of an agent to perform a new task. We approach this problem with a zero-shot agent that requires no given expert traces. Our agent plans for executable actions on a partially observed environment, and iteratively progresses a task by identifying and learning from its mistakes via self-reflection and structured thought management. On the easy tasks of MiniWoB++, we show that our zero-shot agent often outperforms recent SoTAs, with more efficient reasoning. For tasks with more complexity, our reflective agent performs on par with prior best models, even though previous works had the advantages of accessing expert traces or additional screen information.

11:00-12:30 (East Foyer / Virtual)
**Prompting ChatGPT in MNER: Enhanced Multimodal Named Entity Recognition with Auxiliary Refined Knowledge**
*Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang and Gang Pan*
Multimodal Named Entity Recognition (MNER) on social media aims to enhance textual entity prediction by incorporating image-based clues. Existing studies mainly focus on maximizing the utilization of pertinent image information or incorporating external knowledge from explicit knowledge bases. However, these methods either neglect the necessity of providing the model with external knowledge, or encounter issues of high redundancy in the retrieved knowledge. In this paper, we present PGIM — a two-stage framework that aims to leverage ChatGPT as an implicit knowledge base and enable it to heuristically generate auxiliary knowledge for more efficient entity prediction. Specifically, PGIM contains a Multimodal Similar Example Awareness module that selects suitable examples from a small number of predefined artificial samples. These examples are then integrated into a formatted prompt template tailored to the MNER and guide ChatGPT to generate auxiliary refined knowledge. Finally, the acquired knowledge is integrated with the original text and fed into a downstream model for further processing. Extensive experiments show that PGIM outperforms state-of-the-art methods on two classic MNER datasets and exhibits a stronger robustness and generalization capability.

11:00-12:30 (East Foyer / Virtual)
**MenatQA: A New Dataset for Testing the Temporal Comprehension and Reasoning Abilities of Large Language Models**
*Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao and Kang Liu*
Large language models (LLMs) have shown nearly saturated performance on many natural language processing (NLP) tasks. As a result, it is natural for people to believe that LLMs have also mastered abilities such as time understanding and reasoning. However, research on the temporal sensitivity of LLMs has been insufficiently emphasized. To fill this gap, this paper constructs Multiple Sensitive Factors Time QA (MenatQA), which encompasses three temporal factors (scope factor, order factor, counterfactual factor) with total 2,853 samples for evaluating the time comprehension and reasoning abilities of LLMs. This paper tests current mainstream LLMs with different parameter sizes, ranging from billions to hundreds of billions. The results show most LLMs fall behind smaller temporal reasoning models with different degree on these factors. In specific, LLMs show a significant vulnerability to temporal biases and depend heavily on the temporal information provided in questions. Furthermore, this paper undertakes a preliminary investigation into potential improvement strategies by devising specific prompts and leveraging external tools. These approaches serve as valuable baselines or references for future research endeavors.

11:00-12:30 (East Foyer / Virtual)
**Tunable Soft Prompts are Messengers in Federated Learning**
*Chenhe Dong, Yuexiang Xie, Bolin Ding, Ying Shen and Yaliang Li*
Federated learning (FL) enables multiple participants to collaboratively train machine learning models using decentralized data sources, alleviating privacy concerns that arise from directly sharing local data. However, the lack of model privacy protection in FL becomes an unneglectable challenge, especially when people want to federally finetune models based on a proprietary large language model. In this study, we propose a novel FL training approach that accomplishes information exchange among participants via tunable soft prompts. These soft prompts, updated and transmitted between the server and clients, assume the role of the global model parameters and serve as messengers to deliver useful knowledge from the local data and global model. As the global model itself is not required to be shared and the local training is conducted based on an auxiliary model with fewer parameters than the global model, the proposed approach provides protection for the global model while reducing communication and computation costs in FL. Extensive experiments show the effectiveness of the proposed approach compared to several baselines. We have released the source code at https://github.com/alibaba/FederatedScope/tree/fedsp/federatedscope/nlp/fedsp.

11:00-12:30 (East Foyer / Virtual)
**ExplainCPE: A Free-text Explanation Benchmark of Chinese Pharmacist Examination**
*Dongfang Li, Jindi Yu, Baotian Hu, Zhenran Xu and Min Zhang*
In the field of Large Language Models (LLMs), researchers are increasingly exploring their effectiveness across a wide range of tasks. However, a critical area that requires further investigation is the interpretability of these models, particularly the ability to generate rational explanations for their decisions. Most existing explanation datasets are limited to the English language and the general domain, which leads to a scarcity of linguistic diversity and a lack of resources in specialized domains, such as medical. To mitigate this, we propose ExplainCPE, a challenging medical dataset consisting of over 7K problems from Chinese Pharmacist Examination, specifically tailored to assess the model-generated explanations. From the overall results, only GPT-4 passes the pharmacist examination with a 75.7% accuracy, while other models like ChatGPT fail. Further detailed analysis of LLM-generated explanations reveals the limitations of LLMs in understanding medical text and executing computational reasoning. With the increasing importance of AI safety and trustworthiness, ExplainCPE takes a step towards improving and evaluating the interpretability of LLMs in the medical domain.

11:00-12:30 (East Foyer / Virtual)
**Making Body Movement in Sign Language Corpus Accessible for Linguists and Machines with Three-Dimensional Normalization of MediaPipe**
*Victor Skobov and Mayumi Bono*
Linguists can access movement in the sign language video corpus through manual annotation or computational methods. The first relies on a predefinition of features, and the second requires technical knowledge. Methods like MediaPipe and OpenPose are now more often used in

sign language processing. MediaPipe detects a two-dimensional (2D) body pose in a single image with a limited approximation of the depth coordinate. Such 2D projection of a three-dimensional (3D) body pose limits the potential application of the resulting models outside the capturing camera settings and position. 2D pose data does not provide linguists with direct and human-readable access to the collected movement data. We propose our four main contributions: A novel 3D normalization method for MediaPipe's 2D pose, a novel human-readable way of representing the 3D normalized pose data, an analysis of Japanese Sign Language (JSL) sociolinguistic features using the proposed techniques, where we show how an individual signer can be identified based on unique personal movement patterns suggesting a potential threat to anonymity. Our method outperforms the common 2D normalization on a small, diverse JSL dataset. We demonstrate its benefit for deep learning approaches by significantly outperforming the pose-based state-of-the-art models on the open sign language recognition benchmark.

11:00-12:30 (East Foyer / Virtual)
**Locally Differentially Private Document Generation Using Zero Shot Prompting**
*Saiteja Utpala, Sara Hooker and Pin-Yu Chen*
Numerous studies have highlighted the privacy risks associated with pretrained large language models. In contrast, our research offers a unique perspective by demonstrating that pretrained large language models can effectively contribute to privacy preservation. We propose a locally differentially private mechanism called DP-Prompt, which leverages the power of pretrained large language models and zero-shot prompting to counter author de-anonymization attacks while minimizing the impact on downstream utility. When DP-Prompt is used with a powerful language model like ChatGPT (gpt-3.5), we observe a notable reduction in the success rate of de-anonymization attacks, showing that it surpasses existing approaches by a considerable margin despite its simpler design. For instance, in the case of the IMDB dataset, DP-Prompt (with ChatGPT) perfectly recovers the clean sentiment F1 score while achieving a 46% reduction in author identification F1 score against static attackers and a 26% reduction against adaptive attackers. We conduct extensive experiments across six open-source large language models, ranging up to 7 billion parameters, to analyze various effects of the privacy-utility tradeoff. Code is avaliable at https://github.com/SaitejaUtpala/dp_prompt

11:00-12:30 (East Foyer / Virtual)
**SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations**
*Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu and Xiangmin Xu*
Large language models (LLMs) have been widely applied in various fields due to their excellent capability for memorizing knowledge and chain of thought (CoT). When these language models are applied in the field of psychological counseling, they often rush to provide universal advice. However, when users seek psychological support, they need to gain empathy, trust, understanding and comfort, rather than just reasonable advice. To this end, we constructed a multi-turn empathetic conversation dataset of more than 2 million samples, in which the input is the multi-turn conversation context, and the target is empathetic responses that cover expressions such as questioning, comfort, recognition, listening, trust, emotional support, etc. Experiments have shown that the empathy ability of LLMs can be significantly enhanced when finetuning by using multi-turn dialogue history and responses that are closer to the expression of a psychological consultant.

11:00-12:30 (East Foyer / Virtual)
**NarrativeXL: a Large-scale Dataset for Long-Term Memory Models**
*Arsenii Kirillovich Moskvichev and Ky-Vinh Mai*
We propose a new large-scale (nearly a million questions) ultra-long-context (more than 50,000 words average document length) reading comprehension dataset. Using GPT 3.5, we summarized each scene in 1,500 hand-curated fiction books from Project Gutenberg, which resulted in approximately 150 scene-level summaries per book. After that, we created a number of reading comprehension questions based on these summaries, including three types of multiple-choice scene recognition questions, as well as free-form narrative reconstruction questions. With 990,595 total questions, our dataset is an order of magnitude larger than the closest alternatives. Crucially, most questions have a known "retention demand", indicating how long-term of a memory is needed to answer them, which should aid long-term memory performance evaluation. We validate our data in four small-scale experiments: one with human labelers, and three with existing language models. We show that our questions 1) adequately represent the source material 2) can be used to diagnose a model's memory capacity 3) are not trivial for modern language models even when the memory demand does not exceed those models' context lengths. Lastly, we provide our code which can be used to further expand the dataset with minimal human labor.

11:00-12:30 (East Foyer / Virtual)
**Retrieval-Augmented Parsing for Complex Graphs by Exploiting Structure and Uncertainty**
*Zi Lin, Quan Yuan, Panupong Pasupat, Jeremiah Zhe Liu and Jingbo Shang*
Retrieval augmentation enhances generative language models by retrieving informative exemplars relevant for output prediction. However, in realistic graph parsing problems where the output space is large and complex, classic retrieval methods based on input-sentence similarity can fail to identify the most informative exemplars that target graph elements the model is most struggling about, leading to suboptimal retrieval and compromised prediction under limited retrieval budget. In this work, we improve retrieval-augmented parsing for complex graph problems by exploiting two unique sources of information (1) structural similarity and (2) model uncertainty. We propose ***S**tructure-aware and **U**ncertainty-**G**uided **A**daptive **R**etrieval*(**SUGAR**) that first quantify the model uncertainty in graph prediction and identify its most uncertain subgraphs, and then retrieve exemplars based on their structural similarity with the identified uncertain subgraphs. On a suite of real-world parsing benchmarks with non-trivial graph structure (SMCalflow and E-commerce), SUGAR exhibits a strong advantage over its classic counterparts that do not leverage structure or model uncertainty.

11:00-12:30 (East Foyer / Virtual)
**VERVE: Template-based ReflectiVE Rewriting for MotiVational IntErviewing**
*Do June Min, Veronica Perez-Rosas, Ken Resnicow and Rada Mihalcea*
Reflective listening is a fundamental skill that counselors must acquire to achieve proficiency in motivational interviewing (MI). It involves responding in a manner that acknowledges and explores the meaning of what the client has expressed in the conversation. In this work, we introduce the task of counseling response rewriting, which transforms non-reflective statements into reflective responses. We introduce VERVE, a template-based rewriting system with paraphrase-augmented training and adaptive template updating. VERVE first creates a template by identifying and filtering out tokens that are not relevant to reflections and constructs a reflective response using the template. Paraphrase-augmented training allows the model to learn less-strict fillings of masked spans, and adaptive template updating helps discover effective templates for rewriting without significantly removing the original content. Using both automatic and human evaluations, we compare our method against text rewriting baselines and show that our framework is effective in turning non-reflective statements into more reflective responses while achieving a good content preservation-reflection style trade-off.

11:00-12:30 (East Foyer / Virtual)
**Investigating the Effectiveness of Multiple Expert Models Collaboration**
*Ikumi Ito, Takumi Ito, Jun Suzuki and Kentaro Inui*

This paper aims to investigate the effectiveness of several machine translation (MT) models and aggregation methods in a multi-domain setting under fair conditions and explore a direction for tackling multi-domain MT. We mainly compare the performance of the single model approach by jointly training all domains and the multi-expert models approach with a particular aggregation strategy. We conduct experiments on multiple domain datasets and demonstrate that a combination of smaller domain expert models can outperform a larger model trained for all domain data.

11:00-12:30 (East Foyer / Virtual)
**Topic-DPR: Topic-based Prompts for Dense Passage Retrieval**
*Qingfa Xiao, Shuangyin Li and Lei Chen*
Prompt-based learning's efficacy across numerous natural language processing tasks has led to its integration into dense passage retrieval. Prior research has mainly focused on enhancing the semantic understanding of pre-trained language models by optimizing a single vector as a continuous prompt. This approach, however, leads to a semantic space collapse; identical semantic information seeps into all representations, causing their distributions to converge in a restricted region. This hinders differentiation between relevant and irrelevant passages during dense retrieval. To tackle this issue, we present Topic-DPR, a dense passage retrieval model that uses topic-based prompts. Unlike the single prompt method, multiple topic-based prompts are established over a probabilistic simplex and optimized simultaneously through contrastive learning. This encourages representations to align with their topic distributions, improving space uniformity. Furthermore, we introduce a novel positive and negative sampling strategy, leveraging semi-structured data to boost dense retrieval efficiency. Experimental results from two datasets affirm that our method surpasses previous state-of-the-art retrieval techniques.

11:00-12:30 (East Foyer / Virtual)
**Contrastive Learning-based Sentence Encoders Implicitly Weight Informative Words**
*Hiroto Kurita, Goro Kobayashi, Sho Yokoi and Kentaro Inui*
The performance of sentence encoders can be significantly improved through the simple practice of fine-tuning using contrastive loss. A natural question arises: what characteristics do models acquire during contrastive learning? This paper theoretically and experimentally shows that contrastive-based sentence encoders implicitly weight words based on information-theoretic quantities; that is, more informative words receive greater weight, while others receive less. The theory states that, in the lower bound of the optimal value of the contrastive learning objective, the norm of word embedding reflects the information gain associated with the distribution of surrounding words. We also conduct comprehensive experiments using various models, multiple datasets, two methods to measure the implicit weighting of models (Integrated Gradients and SHAP), and two information-theoretic quantities (information gain and self-information). The results provide empirical evidence that contrastive fine-tuning emphasizes informative words.

11:00-12:30 (East Foyer / Virtual)
**ChatCoT: Tool-Augmented Chain-of-Thought Reasoning on Chat-based Large Language Models**
*Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Xin Zhao and Ji-Rong Wen*
Although large language models (LLMs) have achieved excellent performance in a variety of evaluation benchmarks, they still struggle in complex reasoning tasks which require specific knowledge and multi-hop reasoning. To improve the reasoning abilities, we propose **ChatCoT**, a tool-augmented chain-of-thought reasoning framework for chat-based LLMs (*e.g.,* ChatGPT). In ChatCoT, we model the chain-of-thought (CoT) reasoning as multi-turn conversations, to utilize tools in a more natural way through chatting. At each turn, LLMs can either interact with tools or perform the reasoning. Our approach can effectively leverage the multi-turn conversation ability of chat-based LLMs, and integrate the thought chain following and tools manipulation in a unified way. Specially, we initialize the early turns of the conversation by the knowledge about tools, tasks, and reasoning format, and propose an iterative *tool-augmented reasoning* step to perform step-by-step tool-augmented reasoning. The experiment results on two complex reasoning datasets (MATH and HotpotQA) have shown the effectiveness of ChatCoT on complex reasoning tasks, achieving a 7.9% relative improvement over the state-of-the-art baseline.

11:00-12:30 (East Foyer / Virtual)
**RefGPT: Dialogue Generation of GPT, by GPT, and for GPT**
*Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang and Hai Zhao*
Large Language Models (LLMs) have attained the impressive capability to resolve a wide range of NLP tasks by fine-tuning high-quality instruction data. However, collecting human-written data of high quality, especially multi-turn dialogues, is expensive and unattainable for most people. Though previous studies have used powerful LLMs to generate the dialogues automatically, they all suffer from generating untruthful dialogues because of the model hallucination. Therefore, we propose a method called RefGPT to generate enormous truthful and customized dialogues without worrying about factual errors caused by the model hallucination. RefGPT solves the model hallucination in dialogue generation by restricting the LLMs to leverage the given reference instead of reciting their own knowledge to generate dialogues. Additionally, RefGPT adds detailed controls on every utterance to enable high customization capability, which previous studies have ignored. On the basis of RefGPT, we also propose two high-quality dialogue datasets generated by GPT-4, namely **RefGPT-Fact** and **RefGPT-Code**. RefGPT-Fact is a dataset with 100k multi-turn dialogues based on factual knowledge and RefGPT-Code has 76k multi-turn dialogues covering a wide range of coding scenarios. Our code and datasets are released in https://github.com/mutonix/RefGPT.

11:00-12:30 (East Foyer / Virtual)
**SQLPrompt: In-Context Text-to-SQL with Minimal Labeled Data**
*Ruoxi Sun, Sercan Arik, Rajarishi Sinha, Hootan Nakhost, Hanjun Dai, Pengcheng Yin and Tomas Pfister*
Text-to-SQL aims to automate the process of generating SQL queries on a database from natural language text. In this work, we propose "SQLPrompt", tailored to improve the few-shot prompting capabilities of Text-to-SQL for Large Language Models (LLMs). Our methods include innovative prompt design, execution-based consistency decoding strategy which selects the SQL with the most consistent execution outcome among other SQL proposals, and a method that aims to improve performance by diversifying the SQL proposals during consistency selection with different prompt designs ("MixPrompt") and foundation models ("MixLLMs"). We show that *SQLPrompt* outperforms previous approaches for in-context learning with zero labeled data by a large margin, closing the gap with finetuning state-of-the-art with thousands of labeled data.

11:00-12:30 (East Foyer / Virtual)
**EffEval: A Comprehensive Evaluation of Efficiency for MT Evaluation Metrics**
*Daniil Larionov, Jens Grünwald, Christoph Leiter and Steffen Eger*
Efficiency is a key property to foster inclusiveness and reduce environmental costs, especially in an era of LLMs. In this work, we provide a comprehensive evaluation of efficiency for MT evaluation metrics. Our approach involves replacing computation-intensive transformers with lighter alternatives and employing linear and quadratic approximations for alignment algorithms on top of LLM representations. We evaluate six (reference-free and reference-based) metrics across three MT datasets and examine 16 lightweight transformers. In addition, we look into the training efficiency of metrics like COMET by utilizing adapters. Our results indicate that (a) TinyBERT provides the optimal balance between quality and efficiency, (b) CPU speed-ups are more substantial than those on GPU; (c) WMD approximations yield no efficiency gains while reducing quality and (d) adapters enhance training efficiency (regarding backward pass speed and memory requirements) as well

as, in some cases, metric quality. These findings can help to strike a balance between evaluation speed and quality, which is essential for effective NLG systems. Furthermore, our research contributes to the ongoing efforts to optimize NLG evaluation metrics with minimal impact on performance. To our knowledge, ours is the most comprehensive analysis of different aspects of efficiency for MT metrics conducted so far.

11:00-12:30 (East Foyer / Virtual)
**A Table-to-Text Framework with Heterogeneous Multidominance Attention and Self-Evaluated Multi-Pass Deliberation**
*Xi Chen, Xinjiang Lu, Haoran Xin, Wenjun Peng, Haoyang Duan, Feihu Jiang, Jingbo Zhou and Hui Xiong*
Though big progress in table-to-text works, effectively leveraging table structure signals, e.g., hierarchical structure, remains challenging. Besides, deliberating generated descriptions proves to be effective for table-to-text. However, determining the appropriate outcome when encountering multi-pass candidates is another challenge. To this end, we propose a novel table-to-text approach on top of Self-evaluated multi-pass Generation and Heterogenous Multidominance Attention, namely SG-HMA. Specifically, we formulate the table structure into a multidominance (MD) structure and devise a heterogenous multidominance attention (HMA) to comprehensively explore the complex interactions encoded in the hierarchical structure, which can further deliver rich signals for text generation with the help of pre-trained language models (PLMs). Afterward, a contrastive loss is introduced to align the generation objective with evaluation metrics, so the more faithful generated descriptions can be guaranteed. We conduct extensive experiments on three public datasets, demonstrating that SG-HMA outperforms several SOTA methods quantitatively and qualitatively.

11:00-12:30 (East Foyer / Virtual)
**Structure and Label Constrained Data Augmentation for Cross-domain Few-shot NER**
*Jingyi Zhang, Ying Zhang, Yufeng Chen and Jinan Xu*
Cross-domain few-shot named entity recognition (NER) is a challenging task that aims to recognize entities in target domains with limited labeled data by leveraging relevant knowledge from source domains. However, domain gaps limit the effect of knowledge transfer and harm the performance of NER models. In this paper, we analyze those domain gaps from two new perspectives, i.e., entity annotations and entity structures and leverage word-to-tag and word-to-word relations to model them, respectively. Moreover, we propose a novel method called Structure and Label Constrained Data Augmentation (SLC-DA) for Cross-domain Few-shot NER, which novelly design a label constrained pre-train task and a structure constrained optimization objectives in the data augmentation process to generate domain-specific augmented data to help NER models smoothly transition from source to target domains. We evaluate our approach on several standard datasets and achieve state-of-the-art or competitive results, demonstrating the effectiveness of our method in cross-domain few-shot NER.

11:00-12:30 (East Foyer / Virtual)
**Beyond Good Intentions: Reporting the Research Landscape of NLP for Social Good**
*Fernando Gonzalez Adauto, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan and Rada Mihalcea*
With the recent advances in natural language processing (NLP), a vast number of applications have emerged across various use cases. Among the plethora of NLP applications, many academic researchers are motivated to do work that has a positive social impact, in line with the recent initiatives of NLP for Social Good (NLP4SG). However, it is not always obvious to researchers how their research efforts are tackling today's big social problems. Thus, in this paper, we introduce NLP4SGPapers, a scientific dataset with three associated tasks that can help identify NLP4SG papers and characterize the NLP4SG landscape by: (1) identifying the papers that address a social problem, (2) mapping them to the corresponding UN Sustainable Development Goals (SDGs), and (3) identifying the task they are solving and the methods they are using. Using state-of-the-art NLP models, we address each of these tasks and use them on the entire ACL Anthology, resulting in a visualization workspace that gives researchers a comprehensive overview of the field of NLP4SG. Our website is available at https://nlp4sg.vercel.app . We released our data at https://huggingface.co/datasets/feradauto/NLP4SGPapers and code at https://github.com/feradauto/nlp4sg

11:00-12:30 (East Foyer / Virtual)
**MEAL: Stable and Active Learning for Few-Shot Prompting**
*Abdullatif Köksal, Timo Schick and Hinrich Schuetze*
Few-shot classification has made great strides due to foundation models that, through priming and prompting, are highly effective few-shot learners. However, this approach has high variance both across different sets of few shots (*data selection*) and across different finetuning runs (*run variability*). This is problematic not only because it impedes the fair comparison of different approaches, but especially because it makes few-shot learning too unreliable for many real-world applications. To alleviate these issues, we make two contributions for more stable and effective few-shot learning: First, we propose novel ensembling methods and show that they substantially reduce *run variability*. Second, we introduce a new active learning (AL) criterion for *data selection* and present the first AL-based approach specifically tailored towards prompt-based learning. In our experiments, we show that our combined method, MEAL (**M**ultiprompt finetuning and prediction **E**nsembling with **A**ctive **L**earning), improves overall performance of prompt-based finetuning by 2.3 points on five diverse tasks. We publicly share our code and data splits in https://github.com/akoksal/MEAL.

11:00-12:30 (East Foyer / Virtual)
**MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation**
*Joshua Tanner and Jacob Hoffman*
Recent approaches to word sense disambiguation (WSD) utilize encodings of the sense gloss (definition), in addition to the input context, to improve performance. In this work we demonstrate that this approach can be adapted for use in multiword expression (MWE) identification by training models which use gloss and context information to filter MWE candidates produced by a rule-based extraction pipeline. Our approach substantially improves precision, outperforming the state-of-the-art in MWE identification on the DiMSUM dataset by up to 1.9 F1 points and achieving competitive results on the PARSEME 1.1 English dataset. Our models also retain most of their WSD performance, showing that a single model can be used for both tasks. Finally, building on similar approaches using Bi-encoders for WSD, we introduce a novel Poly-encoder architecture which improves MWE identification performance.

11:00-12:30 (East Foyer / Virtual)
**Unsupervised Opinion Summarization Using Approximate Geodesics**
*Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed and Snigdha Chaturvedi*
Opinion summarization is the task of creating summaries capturing popular opinions from user reviews. In this paper, we introduce Geodesic Summarizer (GeoSumm), a novel system to perform unsupervised extractive opinion summarization. GeoSumm consists of an encoder-decoder based representation learning model that generates topical representations of texts. These representations capture the underlying semantics of the text as a distribution over learnable latent units. GeoSumm generates these topical representations by performing dictionary learning over pre-trained text representations at multiple layers of the decoder. We then use these topical representations to quantify the importance of review sentences using a novel approximate geodesic distance-based scoring mechanism. We use the importance scores to identify popular opinions in order to compose general and aspect-specific summaries. Our proposed model, GeoSumm, achieves strong performance on three opinion summarization datasets. We perform additional experiments to analyze the functioning of our model and showcase the generalization ability of GeoSumm across different domains.

11:00-12:30 (East Foyer / Virtual)
**On the Relation between Sensitivity and Accuracy in In-Context Learning**
*Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown and He He*
In-context learning (ICL) suffers from oversensitivity to the prompt, making it unreliable in real-world scenarios. We study the sensitivity of ICL with respect to multiple perturbation types. First, we find that label bias obscures the true sensitivity, and therefore prior work may have significantly underestimated ICL sensitivity. Second, we observe a strong negative correlation between ICL sensitivity and accuracy: predictions sensitive to perturbations are less likely to be correct. Motivated by these findings, we propose $SenSel$, a few-shot selective prediction method that abstains from sensitive predictions. Experiments on ten classification datasets show that $SenSel$ consistently outperforms two commonly used confidence-based and entropy-based baselines on abstention decisions.

11:00-12:30 (East Foyer / Virtual)
**Long-Form Speech Translation through Segmentation with Finite-State Decoding Constraints on Large Language Models**
*Arya D. McCarthy, Hao Zhang, Shankar Kumar, Felix Stahlberg and Ke Wu*
One challenge in speech translation is that plenty of spoken content is long-form, but short units are necessary for obtaining high-quality translations. To address this mismatch, we adapt large language models (LLMs) to split long ASR transcripts into segments that can be independently translated so as to maximize the overall translation quality. We overcome the tendency of hallucination in LLMs by incorporating finite-state constraints during decoding; these eliminate invalid outputs without requiring additional training. We discover that LLMs are adaptable to transcripts containing ASR errors through prompt-tuning or fine-tuning. Relative to a state-of-the-art automatic punctuation baseline, our best LLM improves the average BLEU by 2.9 points for English–German, English–Spanish, and English–Arabic TED talk translation in 9 test sets, just by improving segmentation.

11:00-12:30 (East Foyer / Virtual)
**Bidirectional Masked Self-attention and N-gram Span Attention for Constituency Parsing**
*Soohyeong Kim, Whanhee Cho, Minji Kim and Yong Suk Choi*
Attention mechanisms have become a crucial aspect of deep learning, particularly in natural language processing (NLP) tasks. However, in tasks such as constituency parsing, attention mechanisms can lack the directional information needed to form sentence spans. To address this issue, we propose a Bidirectional masked and N-gram span Attention (BNA) model, which is designed by modifying the attention mechanisms to capture the explicit dependencies between each word and enhance the representation of the output span vectors. The proposed model achieves state-of-the-art performance on the Penn Treebank and Chinese Penn Treebank datasets, with F1 scores of 96.47 and 94.15, respectively. Ablation studies and analysis show that our proposed BNA model effectively captures sentence structure by contextualizing each word in a sentence through bidirectional dependencies and enhancing span representation.

11:00-12:30 (East Foyer / Virtual)
**PAXQA: Generating Cross-lingual Question Answering Examples at Training Scale**
*Bryan Li and Chris Callison-Burch*
Existing question answering (QA) systems owe much of their success to large, high-quality training data. Such annotation efforts are costly, and the difficulty compounds in the cross-lingual setting. Therefore, prior cross-lingual QA work has focused on releasing evaluation datasets, and then applying zero-shot methods as baselines. This work proposes a synthetic data generation method for cross-lingual QA which leverages indirect supervision from existing parallel corpora. Our method termed PAXQA (Projecting annotations for cross-lingual (x) QA) decomposes cross-lingual QA into two stages. First, we apply a question generation (QG) model to the English side. Second, we apply annotation projection to translate both the questions and answers. To better translate questions, we propose a novel use of lexically-constrained machine translation, in which constrained entities are extracted from the parallel bitexts. We apply PAXQA to generate cross-lingual QA examples in 4 languages (662K examples total), and perform human evaluation on a subset to create validation and test splits. We then show that models fine-tuned on these datasets outperform prior synthetic data generation models over several extractive QA datasets. The largest performance gains are for directions with non-English questions and English contexts. Ablation studies show that our dataset generation method is relatively robust to noise from automatic word alignments, showing the sufficient quality of our generations. To facilitate follow-up work, we release our code and datasets at https://github.com/manestay/paxqa.

11:00-12:30 (East Foyer / Virtual)
**Data Augmentation Techniques for Machine Translation of Code-Switched Texts: A Comparative Study**
*Injy Hamed, Nizar Habash and Thang Vu*
Code-switching (CSW) text generation has been receiving increasing attention as a solution to address data scarcity. In light of this growing interest, we need more comprehensive studies comparing different augmentation approaches. In this work, we compare three popular approaches: lexical replacements, linguistic theories, and back-translation (BT), in the context of Egyptian Arabic-English CSW. We assess the effectiveness of the approaches on machine translation and the quality of augmentations through human evaluation. We show that BT and CSW predictive-based lexical replacement, being trained on CSW parallel data, perform best on both tasks. Linguistic theories and random lexical replacement prove to be effective in the lack of CSW parallel data, where both approaches achieve similar results.

11:00-12:30 (East Foyer / Virtual)
**Counterfactual Augmentation for Multimodal Learning Under Presentation Bias**
*Victoria Lin, Louis-Philippe Morency, Dimitrios Dimitriadis and Srinagesh Sharma*
In real-world machine learning systems, labels are often derived from user behaviors that the system wishes to encourage. Over time, new models must be trained as new training examples and features become available. However, feedback loops between users and models can bias future user behavior, inducing a *presentation bias* in the labels that compromises the ability to train new models. In this paper, we propose *counterfactual augmentation*, a novel causal method for correcting presentation bias using generated counterfactual labels. Our empirical evaluations demonstrate that counterfactual augmentation yields better downstream performance compared to both uncorrected models and existing bias-correction methods. Model analyses further indicate that the generated counterfactuals align closely with true counterfactuals in an oracle setting.

11:00-12:30 (East Foyer / Virtual)
**Aksharantar: Open Indic-language Transliteration datasets and models for the Next Billion Users**
*Yash Madhani, Sushane Parthan, Priyanka Vasant Bedekar, Gokul NC, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar and Mitesh M Khapra*
Transliteration is very important in the Indian language context due to the usage of multiple scripts and the widespread use of romanized inputs. However, few training and evaluation sets are publicly available. We introduce Aksharantar, the largest publicly available transliteration dataset for Indian languages created by mining from monolingual and parallel corpora, as well as collecting data from human annotators. The dataset contains 26 million transliteration pairs for 21 Indic languages from 3 language families using 12 scripts. Aksharantar is 21 times larger than existing datasets and is the first publicly available dataset for 7 languages and 1 language family. We also introduce a test set of 103k word pairs for 19 languages that enables a fine-grained analysis of transliteration models on native origin words, foreign words,

frequent words, and rare words. Using the training set, we trained IndicXlit, a multilingual transliteration model that improves accuracy by 15% on the Dakshina test set, and establishes strong baselines on the Aksharantar testset introduced in this work. The models, mining scripts, transliteration guidelines, and datasets are available at https://github.com/AI4Bharat/IndicXlit under open-source licenses.

11:00-12:30 (East Foyer / Virtual)
**Pseudointelligence: A Unifying Lens on Language Model Evaluation**
*Shikhar Murty, Orr Paradise and Pratyusha Sharma*
With large language models surpassing human performance on an increasing number of benchmarks, we must take a principled approach for targeted evaluation of model capabilities. Inspired by pseudorandomness, we propose pseudointelligence, which captures the maxim that "(perceived) intelligence lies in the eye of the beholder." That is, that claims of intelligence are meaningful only when their evaluator is taken into account. Concretely, we propose a complexity-theoretic framework of model evaluation cast as a dynamic interaction between a model and a learned evaluator. We demonstrate that this framework can be used to reason about two case studies in language model evaluation, as well as analyze existing evaluation methods.

11:00-12:30 (East Foyer / Virtual)
**INVITE: a Testbed of Automatically Generated Invalid Questions to Evaluate Large Language Models for Hallucinations**
*Anil Ramakrishna, Rahul Gupta, Jens Lehmann and Morteza Ziyadi*
Recent advancements in Large language models (LLMs) have enabled them to hold free form conversations over multiple turns, but they exhibit a tendency to make unfounded and incorrect statements, commonly known as hallucinations. In particular, LLMs hallucinate frequently when given invalid questions, i.e. ones with incorrect assumptions. The most common approach to evaluate LLMs on hallucinations is to test them on Question Answering (QA) test sets such as TruthfulQA. However, LLMs are increasingly pretrained on massive text corpora scraped from the Internet, which may inevitably expose these test sets to the model during training, leading eventually to an overestimation of model performances on these test sets. In this work, we present an alternative framework to address this risk and to foster further research towards making LLMs robust against invalid questions. We name our framework INVITE: a testbed of automatically generated INValId questions to evaluaTE large language models for hallucinations. In each instantiation, our framework is set up to create a fresh batch of invalid questions by distorting valid facts in which subjects or objects are replaced by similar entities. We evaluate several state of the art LLMs against a testset generated by our framework and highlight its capacity to trigger hallucinations in these models.

11:00-12:30 (East Foyer / Virtual)
**Query-based Image Captioning from Multi-context 360$cdegree$ Images**
*Koki Maeda, Shuhei Kurita, Taiki Miyanishi and Naoaki Okazaki*
A 360-degree image captures the entire scene without the limitations of a camera's field of view, which makes it difficult to describe all the contexts in a single caption. We propose a novel task called Query-based Image Captioning (QuIC) for 360-degree images, where a query (words or short phrases) specifies the context to describe. This task is more challenging than the conventional image captioning task, which describes salient objects in images, as it requires fine-grained scene understanding to select the contents consistent with user's intent based on the query. We construct a dataset for the new task that comprises 3,940 360-degree images and 18,459 pairs of queries and captions annotated manually. Experiments demonstrate that fine-tuning image captioning models further on our dataset can generate more diverse and controllable captions from multiple contexts of 360-degree images.

11:00-12:30 (East Foyer / Virtual)
**ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation**
*Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang and Jingbo Shang*
Despite remarkable advances that large language models have achieved in chatbots nowadays, maintaining a non-toxic user-AI interactive environment has become increasingly critical nowadays. However, previous efforts in toxicity detection have been mostly based on benchmarks derived from social media contents, leaving the unique challenges inherent to real-world user-AI interactions insufficiently explored. In this work, we introduce ToxicChat, a novel benchmark constructed based on real user queries from an open-source chatbot. This benchmark contains the rich, nuanced phenomena that can be tricky for current toxicity detection models to identify, revealing a significant domain difference when compared to social media contents. Our systematic evaluation of models trained on existing toxicity datasets has shown their shortcomings when applied to this unique domain of ToxicChat. Our work illuminates the potentially overlooked challenges of toxicity detection in real-world user-AI conversations. In the future, ToxicChat can be a valuable resource to drive further advancements toward building a safe and healthy environment for user-AI interactions.

11:00-12:30 (East Foyer / Virtual)
**DiffuVST: Narrating Fictional Scenes with Global-History-Guided Denoising Models**
*Shengguang Wu, Mei Yuan and Qi Su*
Recent advances in image and video creation, especially AI-based image synthesis, have led to the production of numerous visual scenes that exhibit a high level of abstractness and diversity. Consequently, Visual Storytelling (VST), a task that involves generating meaningful and coherent narratives from a collection of images, has become even more challenging and is increasingly desired beyond real-world imagery. While existing VST techniques, which typically use autoregressive decoders, have made significant progress, they suffer from low inference speed and are not well-suited for synthetic scenes. To this end, we propose a novel diffusion-based system DiffuVST, which models the generation of a series of visual descriptions as a single conditional denoising process. The stochastic and non-autoregressive nature of DiffuVST at inference time allows it to generate highly diverse narratives more efficiently. In addition, DiffuVST features a unique design with bi-directional text history guidance and multimodal adapter modules, which effectively improve inter-sentence coherence and image-to-text fidelity. Extensive experiments on the story generation task covering four fictional visual-story datasets demonstrate the superiority of DiffuVST over traditional autoregressive models in terms of both text quality and inference speed.

11:00-12:30 (East Foyer / Virtual)
**Can you Summarize my learnings? Towards Perspective-based Educational Dialogue Summarization**
*Raghav Jain, Tulika Saha, Jhagrut Lalwani and Sriparna Saha*
The steady increase in the utilization of Virtual Tutors (VT) over recent years has allowed for a more efficient, personalized, and interactive AI-based learning experiences. A vital aspect in these educational chatbots is summarizing the conversations between the VT and the students, as it is critical in consolidating learning points and monitoring progress. However, the approach to summarization should be tailored according to the perspective. Summarization from the VTs perspective should emphasize on its teaching efficiency and potential improvements. Conversely, student-oriented summaries should distill learning points, track progress, and suggest scope for improvements. Based on this hypothesis, in this work, we propose a new task of Multi-modal Perspective based Dialogue Summarization (MM-PerSumm), demonstrated in an educational setting. Towards this aim, we introduce a novel dataset, CIMA-Summ that summarizes educational dialogues from three unique perspectives: the Student, the Tutor, and a Generic viewpoint. In addition, we propose an Image and Perspective-guided Dialogue Summarization (IP-Summ) model which is a Seq2Seq language model incorporating (i) multi-modal learning from images and (ii) a perspective-based encoder that constructs a dialogue graph capturing the intentions and actions of both the VT and the student, enabling the

summarization of a dialogue from diverse perspectives. Lastly, we conduct detailed analyses of our model's performance, highlighting the aspects that could lead to optimal modeling of IP-Summ.

11:00-12:30 (East Foyer / Virtual)
**A Frustratingly Easy Plug-and-Play Detection-and-Reasoning Module for Chinese Spelling Check**
*Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou and Hai-Tao Zheng*
In recent years, Chinese Spelling Check (CSC) has been greatly improved by designing task-specific pre-training methods or introducing auxiliary tasks, which mostly solve this task in an end-to-end fashion. In this paper, we propose to decompose the CSC workflow into detection, reasoning, and searching subtasks so that the rich external knowledge about the Chinese language can be leveraged more directly and efficiently. Specifically, we design a plug-and-play detection-and-reasoning module that is compatible with existing SOTA non-autoregressive CSC models to further boost their performance. We find that the detection-and-reasoning module trained for one model can also benefit other models. We also study the primary interpretability provided by the task decomposition. Extensive experiments and detailed analyses demonstrate the effectiveness and competitiveness of the proposed module.

11:00-12:30 (East Foyer / Virtual)
**When and Why Does Bias Mitigation Work?**
*Abhilasha Ravichander, Joe Stacey and Marek Rei*
Neural models have been shown to exploit shallow surface features to perform language understanding tasks, rather than learning the deeper language understanding and reasoning skills that practitioners desire. Previous work has developed debiasing techniques to pressure models away from spurious features or artifacts in datasets, with the goal of having models instead learn useful, task-relevant representations. However, what do models actually learn as a result of such debiasing procedures? In this work, we evaluate three model debiasing strategies, and through a set of carefully designed tests we show how debiasing can actually increase the model's reliance on hidden biases, instead of learning robust features that help it solve a task. Further, we demonstrate how even debiasing models against all shallow features in a dataset may still not help models address NLP tasks. As a result, we suggest that debiasing existing models may not be sufficient for many language understanding tasks, and future work should consider new learning paradigms, to address complex challenges such as commonsense reasoning and inference.

11:00-12:30 (East Foyer / Virtual)
**DemaFormer: Damped Exponential Moving Average Transformer with Energy-Based Modeling for Temporal Language Grounding**
*Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T Nguyen, See-Kiong Ng and Anh Tuan Luu*
Temporal Language Grounding seeks to localize video moments that semantically correspond to a natural language query. Recent advances employ the attention mechanism to learn the relations between video moments and the text query. However, naive attention might not be able to appropriately capture such relations, resulting in ineffective distributions where target video moments are difficult to separate from the remaining ones. To resolve the issue, we propose an energy-based model framework to explicitly learn moment-query distributions. Moreover, we propose DemaFormer, a novel Transformer-based architecture that utilizes exponential moving average with a learnable damping factor to effectively encode moment-query inputs. Comprehensive experiments on four public temporal language grounding datasets showcase the superiority of our methods over the state-of-the-art baselines.

11:00-12:30 (East Foyer / Virtual)
**KG-GPT: A General Framework for Reasoning on Knowledge Graphs Using Large Language Models**
*Jiho Kim, Yeonsu Kwon, Yohan Jo and Edward Choi*
While large language models (LLMs) have made considerable advancements in understanding and generating unstructured text, their application in structured data remains underexplored. Particularly, using LLMs for complex reasoning tasks on knowledge graphs (KGs) remains largely untouched. To address this, we propose KG-GPT, a multi-purpose framework leveraging LLMs for tasks employing KGs. KG-GPT comprises three steps: Sentence Segmentation, Graph Retrieval, and Inference, each aimed at partitioning sentences, retrieving relevant graph components, and deriving logical conclusions, respectively. We evaluate KG-GPT using KG-based fact verification and KGQA benchmarks, with the model showing competitive and robust performance, even outperforming several fully-supervised models. Our work, therefore, marks a significant step in unifying structured and unstructured data processing within the realm of LLMs.

11:00-12:30 (East Foyer / Virtual)
**SWEET - Weakly Supervised Person Name Extraction for Fighting Human Trafficking**
*Javin Liu, Hao Yu, Vidya Sujaya, Pratheeksha Nair, Kellin Pelrine and Reihaneh Rabbany*
In this work, we propose a weak supervision pipeline SWEET: Supervise Weakly for Entity Extraction to fight Trafficking for extracting person names from noisy escort advertisements. Our method combines the simplicity of rule-matching (through antirules, i.e., negated rules) and the generalizability of large language models fine-tuned on benchmark, domain-specific and synthetic datasets, treating them as weak labels. One of the major challenges in this domain is limited labeled data. SWEET addresses this by obtaining multiple weak labels through labeling functions and effectively aggregating them. SWEET outperforms the previous supervised SOTA method for this task by 9% F1 score on domain data and better generalizes to common benchmark datasets. Furthermore, we also release HTGEN, a synthetically generated dataset of escort advertisements (built using ChatGPT) to facilitate further research within the community.

11:00-12:30 (East Foyer / Virtual)
**Systematic Assessment of Factual Knowledge in Large Language Models**
*Linhao Luo, Trang Vu, Dinh Phung and Reza Haf*
Previous studies have relied on existing question-answering benchmarks to evaluate the knowledge stored in large language models (LLMs). However, this approach has limitations regarding factual knowledge coverage, as it mostly focuses on generic domains which may overlap with the pretraining data. This paper proposes a framework to systematically assess the factual knowledge of LLMs by leveraging knowledge graphs (KGs). Our framework automatically generates a set of questions and expected answers from the facts stored in a given KG, and then evaluates the accuracy of LLMs in answering these questions. We systematically evaluate the state-of-the-art LLMs with KGs in generic and specific domains. The experiment shows that ChatGPT is consistently the top performer across all domains. We also find that LLMs performance depends on the instruction finetuning, domain and question complexity and is prone to adversarial context.

11:00-12:30 (East Foyer / Virtual)
**Affective and Dynamic Beam Search for Story Generation**
*Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen and Snigdha Chaturvedi*
Storytelling's captivating potential makes it a fascinating research area, with implications for entertainment, education, therapy, and cognitive studies. In this paper, we propose Affective Story Generator (AffGen) for generating interesting narratives. AffGen introduces 'intriguing twists' in narratives by employing two novel techniques—Dynamic Beam Sizing and Affective Reranking. Dynamic Beam Sizing encourages less predictable, more captivating word choices using a contextual multi-arm bandit model. Affective Reranking prioritizes sentence candidates based on affect intensity. Our empirical evaluations, both automatic and human, demonstrate AffGen's superior performance

over existing baselines in generating affectively charged and interesting narratives. Our ablation study and analysis provide insights into the strengths and weaknesses of AffGen.

11:00-12:30 (East Foyer / Virtual)
**Learning to love diligent trolls: Accounting for rater effects in the dialogue safety task**
*Michael John Ilagan*
Chatbots have the risk of generating offensive utterances, which must be avoided. Post-deployment, one way for a chatbot to continuously improve is to source utterance/label pairs from feedback by live users. However, among users are trolls, who provide training examples with incorrect labels. To de-troll training data, previous work removed training examples that have high user-aggregated cross-validation (CV) error. However, CV is expensive; and in a coordinated attack, CV may be overwhelmed by trolls in number and in consistency among themselves. In the present work, I address both limitations by proposing a solution inspired by methodology in automated essay scoring (AES): have multiple users rate each utterance, then perform latent class analysis (LCA) to infer correct labels. As it does not require GPU computations, LCA is inexpensive. In experiments, I found that the AES-like solution can infer training labels with high accuracy when trolls are consistent, even when trolls are the majority.

11:00-12:30 (East Foyer / Virtual)
**STEER: Unified Style Transfer with Expert Reinforcement**
*Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck and Yejin Choi*
While text style transfer has many applications across natural language processing, the core premise of transferring from a single source style is unrealistic in a real-world setting. In this work, we focus on arbitrary style transfer: rewriting a text from an arbitrary, unknown style to a target style. We propose STEER: Unified Style Transfer with Expert Reinforcement, a unified frame-work developed to overcome the challenge of limited parallel data for style transfer. STEER involves automatically generating a corpus of style-transfer pairs using a product of experts during decoding. The generated offline data is then used to pre-train an initial policy before switching to online, off-policy reinforcement learning for further improvements via fine-grained reward signals. STEER is unified and can transfer to multiple target styles from an arbitrary, unknown source style, making it particularly flexible and efficient. Experimental results on a challenging dataset with text from a diverse set of styles demonstrate state-of-the-art results compared to competitive baselines. Remarkably, STEER outperforms the 175B parameter instruction-tuned GPT-3 on overall style transfer quality, despite being 226 times smaller in size. We also show STEER is robust, maintaining its style transfer capabilities on out-of-domain data, and surpassing nearly all baselines across various styles. The success of our method highlights the potential of RL algorithms when augmented with controllable decoding to overcome the challenge of limited data supervision.

11:00-12:30 (East Foyer / Virtual)
**Improving Pacing in Long-Form Story Planning**
*Yichen Wang, Kevin Yang, Xiaoming Liu and Dan Klein*
Existing LLM-based systems for writing long-form stories or story outlines frequently suffer from unnatural pacing, whether glossing over important events or over-elaborating on insignificant details, resulting in a jarring experience for the reader. We propose a **CONC**rete **O**utline **C**on**T**rol (CONCOCT) system to improve pacing when automatically generating story outlines. We first train a *concreteness evaluator* to judge which of two events is more concrete (low-level-detailed). This evaluator can then be used to control pacing in hierarchical outline generation; in this work, we explore a *vaguest-first* expansion procedure that aims for uniform pacing. We further use the evaluator to filter new outline items based on predicted concreteness. Compared to a baseline hierarchical outline generator, humans judge CONCOCT's pacing to be more consistent over 57% of the time across multiple outline lengths; the gains also translate to downstream stories. All code, data, and models are open-sourced.

11:00-12:30 (East Foyer / Virtual)
**A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction**
*Ruihao Shui, Yixin Cao, Xiang Wang and Tat-Seng Chua*
Large language models (LLMs) have demonstrated great potential for domain-specific applications, such as the law domain. However, recent disputes over GPT-4's law evaluation raise questions concerning their performance in real-world legal tasks. To systematically investigate their competency in the law, we design practical baseline solutions based on LLMs and test on the task of legal judgment prediction. In our solutions, LLMs can work alone to answer open questions or coordinate with an information retrieval (IR) system to learn from similar cases or solve simplified multi-choice questions. We show that similar cases and multi-choice options, namely label candidates, included in prompts can help LLMs recall domain knowledge that is critical for expertise legal reasoning. We additionally present an intriguing paradox wherein an IR system surpasses the performance of LLM+IR due to limited gains acquired by weaker LLMs from powerful IR systems. In such case, the role of LLMs becomes redundant. Our evaluation pipeline can be easily extended into other tasks to facilitate evaluations in other domains. Code is available at https://github.com/srhthu/LM-CompEval-Legal

11:00-12:30 (East Foyer / Virtual)
**Few-shot Unified Question Answering: Tuning Models or Prompts?**
*Srijan Bansal, Semih Yavuz, Bo Pang, Meghana Moorthy Bhat and Yingbo Zhou*
Question-answering (QA) tasks often investigate specific question types, knowledge domains, or reasoning skills, leading to specialized models catering to specific categories of QA tasks. While recent research has explored the idea of unified QA models, such models are usually explored for high-resource scenarios and require re-training to extend their capabilities. To overcome these drawbacks, the paper explores the potential of two paradigms of tuning, model, and prompts, for unified QA under a low-resource setting. The paper provides an exhaustive analysis of their applicability using 16 QA datasets, revealing that prompt tuning can perform as well as model tuning in a few-shot setting with a good initialization. The study also shows that parameter-sharing results in superior few-shot performance, simple knowledge transfer techniques for prompt initialization can be effective, and prompt tuning achieves a significant performance boost from pre-training in a low-resource regime. The research offers insights into the advantages and limitations of prompt tuning for unified QA in a few-shot setting, contributing to the development of effective and efficient systems in low-resource scenarios.

11:00-12:30 (East Foyer / Virtual)
**Improving Zero-shot Reader by Reducing Distractions from Irrelevant Documents in Open-Domain Question Answering**
*Sukmin Cho, Jeongyeon Seo, Soyeong Jeong and Jong C. Park*
Large language models (LLMs) enable zero-shot approaches in open-domain question answering (ODQA), yet with limited advancements as the reader is compared to the retriever. This study aims at the feasibility of a zero-shot reader that addresses the challenges of computational cost and the need for labeled data. We find that LLMs are distracted due to irrelevant documents in the retrieved set and the overconfidence of the generated answers when they are exploited as zero-shot readers. To tackle these problems, we mitigate the impact of such documents via Distraction-aware Answer Selection (DAS) with a negation-based instruction and score adjustment for proper answer selection. Experimental results show that our approach successfully handles distraction across diverse scenarios, enhancing the performance of zero-shot readers. Furthermore, unlike supervised readers struggling with unseen data, zero-shot readers demonstrate outstanding transferability without any

training.

11:00-12:30 (East Foyer / Virtual)
**The Law and NLP: Bridging Disciplinary Disconnects**
*Robert Mahari, Dominik Stammbach, Elliott Ash and Alex Pentland*
Legal practice is intrinsically rooted in the fabric of language, yet legal practitioners and scholars have been slow to adopt tools from natural language processing (NLP). At the same time, the legal system is experiencing an access to justice crisis, which could be partially alleviated with NLP. In this position paper, we argue that the slow uptake of NLP in legal practice is exacerbated by a disconnect between the needs of the legal community and the focus of NLP researchers. In a review of recent trends in the legal NLP literature, we find limited overlap between the legal NLP community and legal academia. Our interpretation is that some of the most popular legal NLP tasks fail to address the needs of legal practitioners. We discuss examples of legal NLP tasks that promise to bridge disciplinary disconnects and highlight interesting areas for legal NLP research that remain underexplored.

11:00-12:30 (East Foyer / Virtual)
**That was the last straw, we need more: Are Translation Systems Sensitive to Disambiguating Context?**
*Jaechan Lee, Alisa Liu, Orevaoghene Ahia, Hila Gonen and Noah A. Smith*
The translation of ambiguous text presents a challenge for translation systems, as it requires using the surrounding context to disambiguate the intended meaning as much as possible. While prior work has studied ambiguities that result from different grammatical features of the source and target language, we study semantic ambiguities that exist in the source (English in this work) itself. In particular, we focus on idioms that are open to both literal and figurative interpretations (e.g., goose egg), and collect TIDE, a dataset of 512 pairs of English sentences containing idioms with disambiguating context such that one is literal (it laid a goose egg) and another is figurative (they scored a goose egg, as in a score of zero). In experiments, we compare MT-specific models and language models for (i) their preference when given an ambiguous subsentence, (ii) their sensitivity to disambiguating context, and (iii) the performance disparity between figurative and literal source sentences. We find that current MT models consistently translate English idioms literally, even when the context suggests a figurative interpretation. On the other hand, LMs are far more context-aware, although there remain disparities across target languages. Our findings underline the potential of LMs as a strong backbone for context-aware translation.

11:00-12:30 (East Foyer / Virtual)
**Eyes Show the Way: Modelling Gaze Behaviour for Hallucination Detection**
*Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra and Pushpak Bhattacharyya*
Detecting hallucinations in natural language processing (NLP) is a critical undertaking that demands a deep understanding of both the semantic and pragmatic aspects of languages. Cognitive approaches that leverage users' behavioural signals, such as gaze, have demonstrated effectiveness in addressing NLP tasks with similar linguistic complexities. However, their potential in the context of hallucination detection remains largely unexplored. In this paper, we propose a novel cognitive approach for hallucination detection that leverages gaze signals from humans. We first collect and introduce an eye tracking corpus (IITB-HGC: IITB-Hallucination Gaze corpus) consisting of 500 instances, annotated by five annotators for hallucination detection. Our analysis reveals that humans selectively attend to relevant parts of the text based on distributional similarity, similar to the attention bias phenomenon in psychology. We identify two attention strategies employed by humans: global attention, which focuses on the most informative sentence, and local attention, which focuses on important words within a sentence. Leveraging these insights, we propose a novel cognitive framework for hallucination detection that incorporates these attention biases. Experimental evaluations on the FactCC dataset demonstrate the efficacy of our approach, obtaining a balanced accuracy of 87.1%. Our study highlights the potential of gaze-based approaches in addressing the task of hallucination detection and sheds light on the cognitive processes employed by humans in identifying inconsistencies.

11:00-12:30 (East Foyer / Virtual)
**Probing Representations for Document-level Event Extraction**
*Barry Wang, Xinya Du and Claire Cardie*
The probing classifiers framework has been employed for interpreting deep neural network models for a variety of natural language processing (NLP) applications. Studies, however, have largely focused on sentencelevel NLP tasks. This work is the first to apply the probing paradigm to representations learned for document-level information extraction (IE). We designed eight embedding probes to analyze surface, semantic, and event-understanding capabilities relevant to document-level event extraction. We apply them to the representations acquired by learning models from three different LLM-based document-level IE approaches on a standard dataset. We found that trained encoders from these models yield embeddings that can modestly improve argument detections and labeling but only slightly enhance event-level tasks, albeit trade-offs in information helpful for coherence and event-type prediction. We further found that encoder models struggle with document length and cross-sentence discourse.

11:00-12:30 (East Foyer / Virtual)
**Addressing the Length Bias Challenge in Document-Level Neural Machine Translation**
*Zhang Zhuocheng, Shuhao Gu, Min Zhang and Yang Feng*
Document-level neural machine translation (DNMT) has shown promising results by incorporating context information through increased maximum lengths of source and target sentences. However, this approach also introduces a length bias problem, whereby DNMT suffers from significant translation quality degradation when decoding sentences that are much shorter or longer than the maximum sentence length during training, i.e., the length bias problem. To prevent the model from neglecting shorter sentences, we sample the training data to ensure a more uniform distribution across different sentence lengths while progressively increasing the maximum sentence length during training. Additionally, we introduce a length-normalized attention mechanism to aid the model in focusing on target information, mitigating the issue of attention divergence when processing longer sentences. Furthermore, during the decoding stage of DNMT, we propose a sliding decoding strategy that limits the length of target sentences to not exceed the maximum length encountered during training. The experimental results indicate that our method can achieve state-of-the-art results on several open datasets, and further analysis shows that our method can significantly alleviate the length bias problem.

11:00-12:30 (East Foyer / Virtual)
**Knowledge Corpus Error in Question Answering**
*Yejoon Lee, Philhoon Oh and James Thorne*
Recent works in open-domain question answering (QA) have explored generating context passages from large language models (LLMs), replacing the traditional retrieval step in the QA pipeline. However, it is not well understood why generated passages can be more effective than retrieved ones. This study revisits the conventional formulation of QA and introduces the concept of *knowledge corpus error*. This error arises when the knowledge corpus used for retrieval is only a subset of the entire string space, potentially excluding more helpful passages that exist outside the corpus. LLMs may mitigate this shortcoming by generating passages in a larger space. We come up with an experiment of paraphrasing human-annotated gold context using LLMs to observe knowledge corpus error empirically. Our results across three QA benchmarks reveal an increased performance (10% - 13%) when using paraphrased passage, indicating a signal for the existence of knowledge corpus error.

11:00-12:30 (East Foyer / Virtual)
**Automatic Evaluate Dialogue Appropriateness by Using Dialogue Act**
*Bao Chen, Yuanjie Wang, Zeming Liu and Yuhang Guo*
Evaluation of dialogue systems requires assessing various aspects, among which appropriateness holds significance as a core element of communicative language competence. However, current evaluations heavily rely on human judgments, which are time-consuming, labor-intensive, prone to biases, and lacking objectivity. In this paper, we introduce Dialogue Act Appropriateness (DAA), a novel method that utilizes the underlying patterns of dialogue act transitions to evaluate the appropriateness of chatbot responses. We learn transition patterns from human-human dialogue corpora, evaluating chatbot appropriateness by measuring the similarity of their transition patterns to those observed in human-human dialogues. To validate DAA, we annotate a test dataset by manually evaluating the appropriateness of dialogues from multiple chatbot systems. The experimental results demonstrate a strong correlation between our evaluation metric and human ratings, establishing the reliability of DAA as a measure of dialogue appropriateness.

11:00-12:30 (East Foyer / Virtual)
**Better Together: Enhancing Generative Knowledge Graph Completion with Language Models and Neighborhood Information**
*Alla Chepurova, Aydar Bulatov, Yuri Kuratov and Mikhail Burtsev*
Real-world Knowledge Graphs (KGs) often suffer from incompleteness, which limits their potential performance. Knowledge Graph Completion (KGC) techniques aim to address this issue. However, traditional KGC methods are computationally intensive and impractical for large-scale KGs, necessitating the learning of dense node embeddings and computing pairwise distances. Generative transformer-based language models (e.g., T5 and recent KGT5) offer a promising solution as they can predict the tail nodes directly. In this study, we propose to include node neighborhoods as additional information to improve KGC methods based on language models. We examine the effects of this imputation and show that, on both inductive and transductive Wikidata subsets, our method outperforms KGT5 and conventional KGC approaches. We also provide an extensive analysis of the impact of neighborhood on model prediction and show its importance. Furthermore, we point the way to significantly improve KGC through more effective neighborhood selection.

11:00-12:30 (East Foyer / Virtual)
**LLMDet: A Third Party Large Language Models Generated Text Detection Tool**
*Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng and Tat-Seng Chua*
Generated texts from large language models (LLMs) are remarkably close to high-quality human-authored text, raising concerns about their potential misuse in spreading false information and academic misconduct. Consequently, there is an urgent need for a highly practical detection tool capable of accurately identifying the source of a given text. However, existing detection tools typically rely on access to LLMs and can only differentiate between machine-generated and human-authored text, failing to meet the requirements of fine-grained tracing, intermediary judgment, and rapid detection. Therefore, we propose LLMDet, a model-specific, secure, efficient, and extendable detection tool, that can source text from specific LLMs, such as GPT-2, OPT, LLaMA, and others. In LLMDet, we record the next-token probabilities of salient n-grams as features to calculate proxy perplexity for each LLM. By jointly analyzing the proxy perplexities of LLMs, we can determine the source of the generated text. Experimental results show that LLMDet yields impressive detection performance while ensuring speed and security, achieving 98.54% precision and about $\times 5.0$ faster for recognizing human-authored text. Additionally, LLMDet can effortlessly extend its detection capabilities to a new open-source model. We will provide an open-source tool at `https://github.com/TrustedLLM/LLMDet`.

11:00-12:30 (East Foyer / Virtual)
**Pit One Against Many: Leveraging Attention-head Embeddings for Parameter-efficient Multi-head Attention**
*Huiyin Xue and Nikolaos Aletras*
Scaling pre-trained language models has resulted in large performance gains in various natural language processing tasks but comes with a large cost in memory requirements. Inspired by the position embeddings in transformers, we aim to simplify and reduce the memory footprint of the multi-head attention (MHA) mechanism. We propose an alternative module that uses only a single shared projection matrix and multiple head embeddings (MHE), i.e. one per head. We empirically demonstrate that our MHE attention is substantially more memory efficient compared to alternative attention mechanisms while achieving high predictive performance retention ratio to vanilla MHA on several downstream tasks. MHE attention only requires a negligible fraction of additional parameters ($3nd$, where $n$ is the number of attention heads and $d$ the size of the head embeddings) compared to a single-head attention, while MHA requires $(3n^2 - 3n)d^2 - 3nd$ additional parameters.

11:00-12:30 (East Foyer / Virtual)
**Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning**
*Duarte Miguel Alves, Nuno M Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo and Andre Martins*
Large language models (LLMs) are a promising avenue for machine translation (MT). However, current LLM-based MT systems are brittle: their effectiveness highly depends on the choice of few-shot examples and they often require extra post-processing due to overgeneration. Alternatives such as finetuning on translation instructions are computationally expensive and may weaken in-context learning capabilities, due to overspecialization. In this paper, we provide a closer look at this problem. We start by showing that adapter-based finetuning with LoRA matches the performance of traditional finetuning while reducing the number of training parameters by a factor of 50. This method also outperforms few-shot prompting and eliminates the need for post-processing or in-context examples. However, we show that finetuning generally degrades few-shot performance, hindering adaptation capabilities. Finally, to obtain the best of both worlds, we propose a simple approach that incorporates few-shot examples during finetuning. Experiments on 10 language pairs show that our proposed approach recovers the original few-shot capabilities while keeping the added benefits of finetuning.

11:00-12:30 (East Foyer / Virtual)
**FaLA: Fast Linear Adaptation for Replacing Backbone Models on Edge Devices**
*Shuo Huang, Lizhen Qu, Xingliang Yuan and Chunyang Chen*
In this work, we study the language model backbone replacement problem for personalized downstream tasks in a non-stationary on-device scenario. In real world, company may periodically update the knowledge and architectures of backbones to keep the competitive in the market, meanwhile, to accommodate the users' own preference, models are personalized to fit users' own distribution locally. Traditional full model tuning or transfer learning for such replacements often incur considerable local device training costs and necessitate extensive backpropagation within deep transformer layers. Addressing this issue, we propose a novel, lightweight tuning method for personalized NLP classification tasks post-backbone replacement. Our approach leverages a personalized matrix calculated from documents corresponding to users' old and new backbones. This matrix facilitates top-layer parameter tuning, drastically reducing backpropagation computation. To further mitigate training costs associated with matrix linear optimization, we employ correlation clustering to curate a few examples from personalized cluster sets for individuals. Our method achieves over 1000 times computation reduction in Flops for backpropagation and brings the user-specific initialization for personal matrix yielding significant performance boost compared with popular transfer learning methods.

11:00-12:30 (East Foyer / Virtual)
**DiffuSeq-v2: Bridging Discrete and Continuous Text Spaces for Accelerated Seq2Seq Diffusion Models**
*Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu and Lingpeng Kong*
Diffusion models have gained prominence in generating high-quality sequences of text. Nevertheless, current approaches predominantly represent discrete text within a continuous diffusion space, which incurs substantial computational overhead during training and results in slower sampling speeds. In this paper, we introduce a soft absorbing state that facilitates the diffusion model in learning to reconstruct discrete mutations based on the underlying Gaussian space, thereby enhancing its capacity to recover conditional signals. During the sampling phase, we employ state-of-the-art ODE solvers within the continuous space to expedite the sampling process. Comprehensive experimental evaluations reveal that our proposed method effectively accelerates the training convergence by 4x and generates samples of similar quality 800x faster, rendering it significantly closer to practical application.

11:00-12:30 (East Foyer / Virtual)
**EMO-KNOW: A Large Scale Dataset on Emotion-Cause**
*Mia Huong Nguyen, Yasith Samaradivakara, Prasanth Sasikumar, Chitralekha Gupta and Suranga Chandima Nanayakkara*
Emotion-Cause analysis has attracted the attention of researchers in recent years. However, most existing datasets are limited in size and number of emotion categories. They often focus on extracting parts of the document that contain the emotion cause and fail to provide more abstractive, generalizable root cause. To bridge this gap, we introduce a large-scale dataset of emotion causes, derived from 9.8 million cleaned tweets over 15 years. We describe our curation process, which includes a comprehensive pipeline for data gathering, cleaning, labeling, and validation, ensuring the dataset's reliability and richness. We extract emotion labels and provide abstractive summarization of the events causing emotions. The final dataset comprises over 700,000 tweets with corresponding emotion-cause pairs spanning 48 emotion classes, validated by human evaluators. The novelty of our dataset stems from its broad spectrum of emotion classes and the abstractive emotion cause that facilitates the development of an emotion-cause knowledge graph for nuanced reasoning. Our dataset will enable the design of emotion-aware systems that account for the diverse emotional responses of different people for the same event.

11:00-12:30 (East Foyer / Virtual)
**Synthesize, if you do not have: Effective Synthetic Dataset Creation Strategies for Self-Supervised Opinion Summarization in E-commerce**
*Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, Nikesh Garera and Pushpak Bhattacharyya*
In e-commerce, opinion summarization is the process of condensing the opinions presented in product reviews. However, the absence of large amounts of supervised datasets presents challenges in generating both aspect-specific and general opinion summaries. Existing approaches have attempted to address these challenges through synthetic dataset creation (SDC). However, general opinion summarization models struggle to generate summaries faithful to the input reviews whereas aspect-specific opinion summarization models are limited due to their reliance on human-specified aspects and seed words. To address this, we propose SDC strategies tailored for general and aspect-specific opinion summarization. We experimented on three e-commerce test sets: Oposum+, Amazon, and Flipkart. For general opinion summarization, pre-trained language model (PLM) fine-tuned on our general synthetic dataset surpass the SOTA on average by 2.3 R1 points. Faithfulness evaluation metrics and human evaluations indicate that our model-generated summaries are more faithful to the input compared to others. For aspect-specific opinion summarization, PLM fine-tuned on our aspect-specific synthetic dataset surpass SOTA by $\sim$ 1 R1 point without the aid of any human-specified aspects or seed words.

11:00-12:30 (East Foyer / Virtual)
**In-Context Learning Creates Task Vectors**
*Roee Hendel, Mor Geva and Amir Globerson*
In-context learning (ICL) in Large Language Models (LLMs) has emerged as a powerful new learning paradigm. However, its underlying mechanism is still not well understood. In particular, it is challenging to map it to the "standard' machine learning framework, where one uses a training set $S$ to find a best-fitting function $f(x)$ in some hypothesis class. Here we make progress on this problem by showing that the functions learned by ICL often have a very simple structure: they correspond to the transformer LLM whose only inputs are the query $x$ and a single "task vector' calculated from the training set. Thus, ICL can be seen as compressing $S$ into a single task vector $\theta(S)$ and then using this task vector to modulate the transformer to produce the output. We support the above claim via comprehensive experiments across a range of models and tasks.

11:00-12:30 (East Foyer / Virtual)
**Simple Hardware-Efficient PCFGs with Independent Left and Right Productions**
*Wei Liu, Songlin Yang, Yoon Kim and Kewei Tu*
Scaling dense PCFGs to thousands of nonterminals via low-rank parameterizations of the rule probability tensor has been shown to be beneficial for unsupervised parsing. However, PCFGs scaled this way still perform poorly as a language model, and even underperform similarly-sized HMMs. This work introduces *SimplePCFG*, a simple PCFG formalism with independent left and right productions. Despite imposing a stronger independence assumption than the low-rank approach, we find that this formalism scales more effectively both as a language model and as an unsupervised parser. We further introduce *FlashInside*, a hardware IO-aware implementation of the inside algorithm for efficiently scaling simple PCFGs. Through extensive experiments on multiple grammar induction benchmarks, we validate the effectiveness of simple PCFGs over low-rank baselines.

11:00-12:30 (East Foyer / Virtual)
**CONTRASTE: Supervised Contrastive Pre-training With Aspect-based Prompts For Aspect Sentiment Triplet Extraction**
*Rajdeep Mukherjee, Nithish Kannen, Saurabh Kumar Pandey and Pawan Goyal*
Existing works on Aspect Sentiment Triplet Extraction (ASTE) explicitly focus on developing more efficient fine-tuning techniques for the task. Instead, our motivation is to come up with a generic approach that can improve the downstream performances of multiple ABSA tasks simultaneously. Towards this, we present CONTRASTE, a novel pre-training strategy using CONTRastive learning to enhance the ASTE performance. While we primarily focus on ASTE, we also demonstrate the advantage of our proposed technique on other ABSA tasks such as ACOS, TASD, and AESC. Given a sentence and its associated (aspect, opinion, sentiment) triplets, first, we design aspect-based prompts with corresponding sentiments masked. We then (pre)train an encoder-decoder model by applying contrastive learning on the decoder-generated aspect-aware sentiment representations of the masked terms. For fine-tuning the model weights thus obtained, we then propose a novel multi-task approach where the base encoder-decoder model is combined with two complementary modules, a tagging-based Opinion Term Detector, and a regression-based Triplet Count Estimator. Exhaustive experiments on four benchmark datasets and a detailed ablation study establish the importance of each of our proposed components as we achieve new state-of-the-art ASTE results.

11:00-12:30 (East Foyer / Virtual)
**On the Impact of Cross-Domain Data on German Language Models**
*Amin Dada, Aokun Chen, Cheng Peng, Kaleb E Smith, Ahmad Idrissi-Yaghir, Constantin Marc Seibold, Jianning Li, Lars Heiliger, Christoph*

*M. Friedrich, Daniel Truhn, Jan Egger, Jiang Bian, Jens Kleesiek and Yonghui Wu*
Traditionally, large language models have been either trained on general web crawls or domain-specific data. However, recent successes of generative large language models, have shed light on the benefits of cross-domain datasets. To examine the significance of prioritizing data diversity over quality, we present a German dataset comprising texts from five domains, along with another dataset aimed at containing high-quality data. Through training a series of models ranging between 122M and 750M parameters on both datasets, we conduct a comprehensive benchmark on multiple downstream tasks. Our findings demonstrate that the models trained on the cross-domain dataset outperform those trained on quality data alone, leading to improvements up to 4.45% over the previous state-of-the-art.

11:00-12:30 (East Foyer / Virtual)
**TSTR: Target Similarity Tuning Meets the Real World**
*Anirudh Khatry, Sumit Gulwani, Priyanshu Gupta, Vu Le, Mukul Singh, Ananya Singha and Gust Verbruggen*
Target similarity tuning (TST) is a method of selecting relevant examples in natural language (NL) to code generation through large language models (LLMs) to improve performance. Its goal is to adapt a sentence embedding model to have the similarity between two NL inputs match the similarity between their associated code outputs. In this paper, we propose different methods to apply and improve TST in the real world. First, we replace the sentence transformer with embeddings from a larger model, which reduces sensitivity to the language distribution and thus provides more flexibility in synthetic generation of examples, and we train a tiny model that transforms these embeddings to a space where embedding similarity matches code similarity, which allows the model to remain a black box and only requires a few matrix multiplications at inference time. Second, we how to efficiently select a smaller number of training examples to train the TST model. Third, we introduce a ranking-based evaluation for TST that does not require end-to-end code generation experiments, which can be expensive to perform.

11:00-12:30 (East Foyer / Virtual)
**Towards Formality-Aware Neural Machine Translation by Leveraging Context Information**
*Dohee Kim, Yujin Baek, Soyoung Yang and Jaegul Choo*
Formality is one of the most important linguistic properties to determine the naturalness of translation. Although a target-side context contains formality-related tokens, the sparsity within the context makes it difficult for context-aware neural machine translation (NMT) models to properly discern them. In this paper, we introduce a novel training method to explicitly inform the NMT model by pinpointing key informative tokens using a formality classifier. Given a target context, the formality classifier guides the model to concentrate on the formality-related tokens within the context. Additionally, we modify the standard cross-entropy loss, especially toward the formality-related tokens obtained from the classifier. Experimental results show that our approaches not only improve overall translation quality but also reflect the appropriate formality from the target context.

11:00-12:30 (East Foyer / Virtual)
**AniEE: A Dataset of Animal Experimental Literature for Event Extraction**
*Dohee Kim, Ra Yoo, Soyoung Yang, Hee Yang and Jaegul Choo*
Event extraction (EE), as a crucial information extraction (IE) task, aims to identify event triggers and their associated arguments from unstructured text, subsequently classifying them into pre-defined types and roles. In the biomedical domain, EE is widely used to extract complex structures representing biological events from literature. Due to the complicated semantics and specialized domain knowledge, it is challenging to construct biomedical event extraction datasets. Additionally, most existing biomedical EE datasets primarily focus on cell experiments or the overall experimental procedures. Therefore, we introduce AniEE, an event extraction dataset concentrated on the animal experiment stage. We establish a novel animal experiment customized entity and event scheme in collaboration with domain experts. We then create an expert-annotated high-quality dataset containing discontinuous entities and nested events and evaluate our dataset on the recent outstanding NER and EE models.

11:00-12:30 (East Foyer / Virtual)
**Data Pruning for Efficient Model Pruning in Neural Machine Translation**
*Abdul Hameed Azeemi, Ihsan Ayyub Qazi and Agha Ali Raza*
Model pruning methods reduce memory requirements and inference time of large-scale pre-trained language models after deployment. However, the actual pruning procedure is computationally intensive, involving repeated training and pruning until the required sparsity is achieved. This paper combines data pruning with movement pruning for Neural Machine Translation (NMT) to enable efficient fine-pruning. We design a dataset pruning strategy by leveraging cross-entropy scores of individual training instances. We conduct pruning experiments on the task of machine translation from Romanian-to-English and Turkish-to-English, and demonstrate that selecting hard-to-learn examples (top-k) based on training cross-entropy scores outperforms other dataset pruning methods. We empirically demonstrate that data pruning reduces the overall steps required for convergence and the training time of movement pruning. Finally, we perform a series of experiments to tease apart the role of training data during movement pruning and uncover new insights to understand the interplay between data and model pruning in the context of NMT.

11:00-12:30 (East Foyer / Virtual)
**Pretraining Without Attention**
*Junxiong Wang, Jing Nathan Yan, Albert Gu and Alexander M Rush*
Transformers have been essential to pretraining success in NLP. While other architectures have been used, downstream accuracy is either significantly worse, or requires attention layers to match standard benchmarks such as GLUE. This work explores pretraining without attention by using recent advances in sequence routing based on state-space models (SSMs). Our proposed model, Bidirectional Gated SSM (BiGS), combines SSM layers with a multiplicative gating architecture that has been effective in simplified sequence modeling architectures. The model learns static layers that do not consider pair-wise interactions. Even so, BiGS is able to match BERT pretraining accuracy on GLUE and can be extended to long-form pretraining of 4096 tokens without approximation. Analysis shows that while the models have similar average accuracy, the approach has different inductive biases than BERT and scales more efficiently to longer sequences.

11:00-12:30 (East Foyer / Virtual)
**Dual Contrastive Learning Framework for Incremental Text Classification**
*Yigong Wang, Zhuoyi Wang, Yu Lin, Jinghui Guo, Sadaf MD Halim and Latifur Khan*
Incremental learning plays a pivotal role in the context of online knowledge discovery, as it encourages large models (LM) to learn and refresh knowledge continuously. Many approaches have been proposed to simultaneously preserve knowledge from previous tasks while learning new concepts in online NLP applications. In this paper, we primarily focus on learning a more generalized embedding space that could be better transferred to various downstream sequence tasks. The key idea is to learn from both task-agnostic and task-specific embedding aspects so that the inherent challenge of catastrophic forgetting that arises in incremental learning scenarios can be addressed with a more generalized solution. We propose a dual contrastive learning (DCL) based framework to foster the transferability of representations across different tasks, it consists of two key components: firstly, we utilize global contrastive learning that intertwines a task-agnostic strategy for promoting a generalized embedding space; secondly, considering the domain shift from unseen distributions can compromise the quality

of learned embeddings. We further incorporate a task-specific attention mechanism to enhance the adaptability of task-specific weight for various emerging tasks and ultimately reduce errors in generic representations. Experiments over various text datasets demonstrate that our work achieves superior performance and outperforms the current state-of-the-art methods.

11:00-12:30 (East Foyer / Virtual)
**Re-Temp: Relation-Aware Temporal Representation Learning for Temporal Knowledge Graph Completion**
*Kunze Wang, Caren Han and Josiah Poon*
Temporal Knowledge Graph Completion (TKGC) under the extrapolation setting aims to predict the missing entity from a fact in the future, posing a challenge that aligns more closely with real-world prediction problems. Existing research mostly encodes entities and relations using sequential graph neural networks applied to recent snapshots. However, these approaches tend to overlook the ability to skip irrelevant snapshots according to entity-related relations in the query and disregard the importance of explicit temporal information. To address this, we propose our model, Re-Temp (Relation-Aware Temporal Representation Learning), which leverages explicit temporal embedding as input and incorporates skip information flow after each timestamp to skip unnecessary information for prediction. Additionally, we introduce a two-phase forward propagation method to prevent information leakage. Through the evaluation on six TKGC (extrapolation) datasets, we demonstrate that our model outperforms all eight recent state-of-the-art models by a significant margin.

11:00-12:30 (East Foyer / Virtual)
**Weakly-supervised Deep Cognate Detection Framework for Low-Resourced Languages Using Morphological Knowledge of Closely-Related Languages**
*Koustava Goswami, Priya Rani, Theodorus Fransen and John Philip McCrae*
Exploiting cognates for transfer learning in under-resourced languages is an exciting opportunity for language understanding tasks, including unsupervised machine translation, named entity recognition and information retrieval. Previous approaches mainly focused on supervised cognate detection tasks based on orthographic, phonetic or state-of-the-art contextual language models, which under-perform for most under-resourced languages. This paper proposes a novel language-agnostic weakly-supervised deep cognate detection framework for under-resourced languages using morphological knowledge from closely related languages. We train an encoder to gain morphological knowledge of a language and transfer the knowledge to perform unsupervised and weakly-supervised cognate detection tasks with and without the pivot language for the closely-related languages. While unsupervised, it overcomes the need for hand-crafted annotation of cognates. We performed experiments on different published cognate detection datasets across language families and observed not only significant improvement over the state-of-the-art but also our method outperformed the state-of-the-art supervised and unsupervised methods. Our model can be extended to a wide range of languages from any language family as it overcomes the requirement of the annotation of the cognate pairs for training.

11:00-12:30 (East Foyer / Virtual)
**Code Search Debiasing: Improve Search Results beyond Overall Ranking Performance**
*Sheng Zhang, Hui Li, Yanlin Wang, Zhao Wei, Yong Xu, Juhong Wang and Rongrong Ji*
Code search engine is an essential tool in software development. Many code search methods have sprung up, focusing on the overall ranking performance of code search. In this paper, we study code search from another perspective by analyzing the bias of code search models. Biased code search engines provide poor user experience, even though they show promising overall performance. Due to different development conventions (e.g., prefer long queries or abbreviations), some programmers will find the engine useful, while others may find it hard to get desirable search results. To mitigate biases, we develop a general debiasing framework that employs reranking to calibrate search results. It can be easily plugged into existing engines and handle new code search biases discovered in the future. Experiments show that our framework can effectively reduce biases. Meanwhile, the overall ranking performance of code search gets improved after debiasing. Our implementation is available at: https://github.com/KDEGroup/CodeSearchDebiasing.

11:00-12:30 (East Foyer / Virtual)
**Unlocking the Heterogeneous Landscape of Big Data NLP with DUUI**
*Alexander Leonhardt, Giuseppe Abrami, Daniel Baumartz and Alexander Mehler*
Automatic analysis of large corpora is a complex task, especially in terms of time efficiency. This complexity is increased by the fact that flexible, extensible text analysis requires the continuous integration of ever new tools. Since there are no adequate frameworks for these purposes in the field of NLP, and especially in the context of UIMA, that are not outdated or unusable for security reasons, we present a new approach to address the latter task: Docker Unified UIMA Interface (DUUI), a scalable, flexible, lightweight, and feature-rich framework for automatic distributed analysis of text corpora that leverages Big Data experience and virtualization with Docker. We evaluate DUUI's communication approach against a state-of-the-art approach and demonstrate its outstanding behavior in terms of time efficiency, enabling the analysis of big text data.

11:00-12:30 (East Foyer / Virtual)
**CR-COPEC: Causal Rationale of Corporate Performance Changes to learn from Financial Reports**
*Ye Eun Chun, Sunjae Kwon, Kyunghwan Sohn, Nakwon Sung, Junyoup Lee, Byoung Ki Seo, Kevin Compher, Seung-won Hwang and Jaesik Choi*
In this paper, we introduce CR-COPEC called Causal Rationale of Corporate Performance Changes from financial reports. This is a comprehensive large-scale domain-adaptation causal sentence dataset to detect financial performance changes of corporate. CR-COPEC contributes to two major achievements. First, it detects causal rationale from 10-K annual reports of the U.S. companies, which contain experts' causal analysis following accounting standards in a formal manner. This dataset can be widely used by both individual investors and analysts as material information resources for investing and decision-making without tremendous effort to read through all the documents. Second, it carefully considers different characteristics which affect the financial performance of companies in twelve industries. As a result, CR-COPEC can distinguish causal sentences in various industries by taking unique narratives in each industry into consideration. We also provide an extensive analysis of how well CR-COPEC dataset is constructed and suited for classifying target sentences as causal ones with respect to industry characteristics.

11:00-12:30 (East Foyer / Virtual)
**BERT Goes Off-Topic: Investigating the Domain Transfer Challenge using Genre Classification**
*Dmitri Roussinov and Serge Sharoff*
While performance of many text classification tasks has been recently improved due to Pretrained Language Models (PLMs), in this paper we show that they still suffer from a performance gap when the underlying distribution of topics changes. For example, a genre classifier trained on political topics often fails when tested on documents in the same genre, but about sport or medicine. In this work, we quantify this phenomenon empirically with a large corpus and a large set of topics. Thus, we verify that domain transfer remains challenging both for classic PLMs, such as BERT, and for modern large models (LLMs), such as GPT. We develop a data augmentation approach by generating texts in any desired genre and on any desired topic, even when there are no documents in the training corpus that are both in that particular genre and on that particular topic. When we augment the training dataset with the topically-controlled synthetic texts, F1 improves up to 50% for some topics, approaching on-topic training, while showing no or next to no improvement for other topics. While our empirical results

focus on genre classification, our methodology is applicable to other classification tasks such as gender, authorship, or sentiment classification.

## Lunch
12:30-14:00 - Location: Unknown

## Session 11: Plenary - Keynote Speaker: Christopher D. Manning
14:00-15:00 - Location: East & Central

## Coffee Break
15:00-15:30 - Location: West Foyer

## Session 12: Plenary - Best Paper Awards
15:30-16:15 - Location: East & Central

## Session 13: Plenary - Closing Session
16:15-17:00 - Location: East & Central

# 12 Local Guide

## Conference Venue

EMNLP 2023 will be held at "Resorts World Convention Centre"
Located at 8 Sentosa Gateway, Singapore 098269

## About Singapore

Singapore is a small island nation in Southeast Asia, known for its economic prowess and cultural diversity. Nestled at the crossroads of major trade routes, it has evolved into a global financial hub and a melting pot of various ethnicities. The population of over 5.7 million comprises predominantly Chinese, Malay, Indian, and other communities, fostering a rich cultural tapestry.
The city-state boasts a stable political environment, operating as a parliamentary republic with a focus on education, innovation, and technology. English, Malay, Mandarin Chinese, and Tamil are the official languages, reflecting the nation's multicultural identity. Singapore's economy is highly developed, driven by finance, electronics, manufacturing, and tourism.
For such a small island, measuring just half the size of London, it's impressive how Singapore packs in so many iconic landmarks and attractions. Around every corner you'll find something new to explore, a new adventure to have, and a new selfie to take. We've put together a list of the top 10 tourist attractions in Singapore. Although they barely scratch the surface, it's a great start.

- **Marina Bay Sands** - Singapore's Marina Bay Sands is an architectural masterpiece. It's home to many tourist attractions, housing two exhibition centers, two theaters, over 40 restaurants, a museum, a three-story large art gallery, the world's most expensive standalone casino, two shopping malls, and the world's longest elevated pool. The Marina Bay Sands SkyPark is the world's largest public cantilevered platform with a height of 200 meters and a 150-meter-long infinity pool. It stands on top of the three towers and offers a breathtaking panoramic view of the city.

  The two observation decks, The Sands SkyPark Observation Deck and the Sands SkyPark Infinity Pool offer unparalleled views, plus unmissable photo opportunities of the city skyline, the Singapore River, Gardens by the Bay, and the Singapore Strait.

- **Gardens by the Bay** - Gardens by the Bay is a 250-acre garden spanning three waterfront parks in Singapore's Marina Bay area. It features several attractions, including a 22-meter tall cloud forest dome, both indoor and outdoor waterfalls, a 150-meter long hillside garden with 35 terraces, and

over 200,000 plants from 100+ species. Gardens by the Bay is an iconic attraction in Singapore, visited by approximately 1.5 million local and international visitors each year.

Hosting diverse attractions, the Gardens offer unique experiences for people of all ages and interests. These include the Flower Dome for plant lovers, Cloud Forest for nature lovers, the Supertree Grove for adventurers, and the Heritage Gardens for history buffs.

- **Sentosa Island** - Sentosa Island is an island resort off mainland Singapore. The island's attractions include beaches, theme parks, and Singapore's first casino. Sentosa Island is part of the Southern Islands of Singapore.

  The contrasts of Sentosa Island are striking, from its pristine beaches to its exhilarating activities. The island's filled with historical landmarks and cultural treasures, offering something for everyone.

  On Sentosa Island you'll find:

  - Universal Studios
  - Adventure Cove Waterpark
  - Resorts World Sentosa
  - Palawan Beach
  - Tanjong Beach Club
  - Skypark Sentosa by AJ Hackett
  - Many, many more attractions

- **Universal Studios Singapore** - Without a doubt, Universal Studios Singapore is a must-visit attraction for your itinerary, regardless of how long you're visiting. There's something for everyone, with rides for kids and adults. There's the Transformers ride, Shrek 4D Adventure, and Madagascar: A Crate Adventure, to name a few. Some are nice and peaceful, while others are white-knuckle thrill rides.

  You'll find plenty of shops, cafes, restaurants, and kiosks all offering refreshments. If you need to calm down from the rides or escape the endless sun, head inside one of these air-conditioned oases and catch your breath.

  Aside from the roller coaster and rides, you'll also find live shows and meet & greets, plus seven themed zones to explore: Hollywood, New York, Sci-Fi City, Ancient Egypt, The Lost World, Far Far Away, and Madagascar.

- **Changi Experience Studio, Changi Airport** - The Changi Experience Studio (CES) is one of Singapore's largest attractions. It houses 18 unique attractions, including the Butterfly Garden and Rain Vortex. This indoor playground features state-of-the-art technology and combines physical and digital interactions to provide a unique, unforgettable experience.

  The CES works like a living museum that showcases Singapore's heritage and culture, as well as the future of air travel. It aims to inspire the curiosity of visitors about the country's history, culture, and future through interactive exhibits and multimedia shows.

  You may have never considered an airport to be a major attraction, but Changi Airport is unlike any other. Waterfalls, art exhibits, high canopy walks, a variety of mazes, and a giant slide. It's certainly not your standard airport.

- **Bird Paradise** - Bird Paradise at the Mandai Wildlife Reserve is your chance to peek into a mesmerizing world full of colorful birds. Home to over 3,500 birds, the park is a must-visit for all

animal lovers. You'll find perfectly pink flamingos and striking scarlet macaws, as well as some more unique species, such as Shoebills, Southern Cassowaty's and Andean Cock-of-the-rock.

The park has ten different zones to discover, including several vast walk-through aviaries and an impressive penguin habitat. There's also the opportunity to attend presentations where you'll get to see some of the world's most successful winged predators in flight or have fun feeding the pelicans.

For nature lovers, old and young, Bird Paradise is a fine addition to your itinerary.

- **Orchard Road** - Orchard Road is Singapore's most famous shopping belt and a top tourist attraction. This mega-shopping destination has become a global symbol of Singapore's multiculturalism, with shoppers from India, China, and Southeast Asia contributing to the mix of people shopping here.

  The shopping district is busiest in the evenings and weekends, when locals and tourists alike flock to the shops for shopping, dining, and entertainment. Treat yourself to luxury brands, high-street fashion, and cutting-edge electronics as you explore the futuristic malls and shopping complexes.

- **Singapore Flyer** - One of the world's largest observation wheels, standing 165 meters tall with 28 fully air-conditioned capsules, the Singapore Flyer provides breathtaking views of Singapore's skyline.

  The Flyer provides panoramic views from Singapore's central business districts, Marina Bay, and East Coast Park, providing ample photo opportunities during the ride.

  It's an ideal attraction for couples, families, and groups of friends to enjoy together. It offers a unique experience, especially at night when the skyline is beautifully lit up with vibrant colors, plus there's a chance you'll catch a light show or two while you're making the rotation.

- **Chinatown** - Just a short walk from Singapore's central business district, Chinatown is an iconic neighborhood that offers a glimpse into Singapore's rich Chinese heritage. From stunning cultural architecture to mouth-watering food, Chinatown is an essential part of the Singapore experience.

  Whether you are looking for a gourmet family meal or a simple bowl of wonton noodles, Chinatown provides the ultimate dining experience. Whatever your budget, Chinatown has a variety of restaurants offering delicious, authentic dishes.

  Chinatown's also a cultural hub, with the Chinatown Heritage Center showcasing Chinese culture and heritage through traditional arts and performances. You'll also find an incredible selection of art galleries, traditional street markets, and exquisite temples.

- **Singapore Botanic Gardens** - Singapore Botanic Gardens is a nature reserve in the heart of the city, and a place everyone should visit at least once in their lifetime. Not only a beautiful place to relax, it's also an educational and scientific research center that houses over 5,000 plant species.

  Home to a variety of tropical plants, flowers, and trees, it's the perfect place to learn about plants, insects, animals, and even geology. With walking trails leading you through the gardens, you can spot squirrels, butterflies, and even exotic birds.

  The gardens are also a perfect spot to switch off and reconnect with nature. Unwind and relax with a picnic, or stretch your legs on the walking trails or jogging paths. Whatever your energy levels, the Botanic Gardens have you covered.

# Useful Information

### Electricity
In Singapore, the standard electrical voltage is 230V, 50Hz, and the outlets accommodate Type G plugs, identical to UK standards. Please ensure your devices are compatible or bring an appropriate adapter.

### Driving in Singapore
Traffic in Singapore drives on the left.

### Insurance
The Conference Organising Committee or its agents will not be responsible for any medical expenses, loss or accidents incurred during the conference. Delegates are strongly advised to arrange their own personal insurance to cover medical and other expenses including accident or loss. Where a delegate has to cancel for medical reasons, the normal cancellation policy will apply.

### Language
The official languages of Singapore are Malay, Mandarin, Tamil, and English, with English being the most widely used for official and business purposes.

### Money
The official currency of Singapore is the Singapore Dollar (SGD).

### Smoking
Smoking is prohibited in many indoor and public places in Singapore, and designated smoking areas are provided to regulate and confine smoking activities.

### Time
Singapore operates on Singapore Time (SGT), which is UTC+8.

### Weather
Singapore has a tropical climate characterized by high temperatures, high humidity, and significant rainfall throughout the year, with no distinct seasons.

# Visa & Passport

### Do I need a visa?

Information for participants entering Singapore for EMNLP 2023 Traveling to Singapore All travelers are welcome to Singapore regardless of COVID-19 vaccination status. From 13 February 2023, all travelers can enter Singapore with no entry approvals, pre-departure tests, on-arrival tests, quarantine, and COVID-19 travel insurance required. To enter Singapore, travelers need to ensure the following:

- A minimum of 6-months passport validity
- Visa, if applicable
- A submitted SG Arrival Card, up to 3 days before arrival
- An international certificate for vaccination for Yellow Fever, if applicable.

For more information on entry to Singapore, you may visit here.

# Travel to the Conference Venue

- **By Car**
  - On Sentosa Gateway, keep to the left lane and drive down the slope leading to Resorts World Sentosa Car Park.
  - Take a light turn right and drive into tunnel for "Cars/Taxis".
  - Follow signage to "B1 West" and park in green zone. Look for "Lift to Hard Rock Hotel" signage.
  - Take the lift to Hard Rock Hotel lobby.
  - Note that there can be additional charges for entrace to Sentosa Island. Please refer to: https://www.sentosa.com.sg/en/getting-to-sentosa/.
- **By Bus**
  - Take buses 65, 80, 93, 188, 855, 10, 30, 97, 100, 131, 143, 145, 166 and alight at VivoCity.
  - Board bus RWS8 from bus stop 14141 at VivoCity or bus stop 14121 at Merrill Lynch, Harbour Front. Alight at Resorts World Sentosa drop-off point.
  - Enter via the Forum, you will see the Resorts World Convention Centre on the right.
- **By MRT**
  - Take North-East line or Circle line to HarbourFront station.
  - Take Exit E to VivoCity and proceed to level 3 to board the Sentosa Express.
  - Alight 1 stop later at Waterfront station and walk straight until you reach the Lake of Dreams.
  - Take the escalator down and enter via The Forum, you will see the Resorts World Convention Centre on the right.
  - If you are coming from Harbourfront MRT Station, please take the Sentosa Express located Level VivoCity (Lobby L) and alight at Waterfront Station

Please note shuttles are available at the following hotels. If you are staying at a hotel outside of Resorts World check with the hotel for shuttles to the gate at Resorts World.

**RWS Shuttle Route**

For those staying at the Genting Hotel Jurong, a short, 25 minutes drive from Resorts World Sentosa. Hotel guests can use the FREE SHUTTLE between the hotel and the resort.
Please check with the hotel for shuttle times.

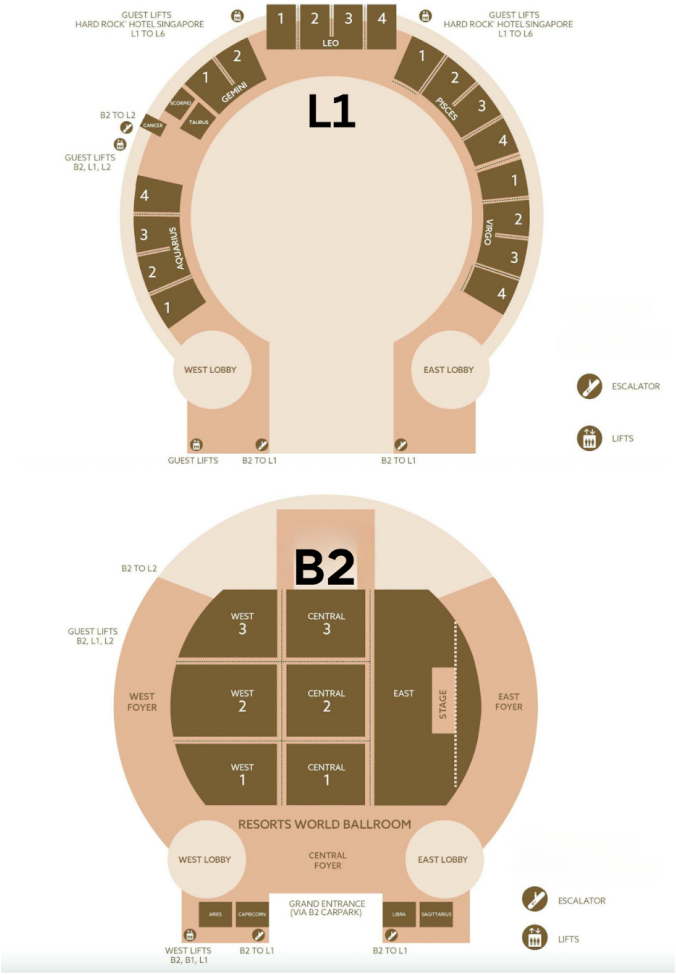
# 13 Venue Map

## Resorts World Map

EMNLP 2023 will be held at “Resorts World Convention Centre”
Located at 8 Sentosa Gateway, Singapore 098269

Asia’s ultimate premium lifestyle destination, Resorts World Sentosa Singapore, is home to six unique hotels, the Asian flagship spa of a world-renowned spa brand, an exclusive casino, as well as four world-class attractions: Universal Studios Singapore™, S.E.A. Aquarium™, Dolphin Island™, and Adventure Cove Waterpark™. With the most number of Michelin stars in one destination, Resorts World Sentosa Singapore offers dining experiences that are truly superior to anyone else in town.

# Venue Layout



## RESORTS WORLD SENTOSA
## PDF Map 圣淘沙名胜世界PDF地图

**ATTRACTIONS 娱乐景点**
1. Universal Studios Singapore 新加坡环球影城
2. S.E.A. Aquarium S.E.A. 海洋馆
3. Adventure Cove Waterpark 水上探险乐园
4. Dolphin Island 海豚岛

**ENTERTAINMENT 娱乐活动**
1. CASINO 赌场
2. The Coliseum 硬石坊

**SHOPPING 购物乐趣**
1. Bvlgari 宝曼斯
2. Beauty Love
3. Chopard 萧邦
4. Coach 蔻驰
5. Cartier 卡地亚
6. Furla 芙拉
7. Hershey's Chocolate World Singapore 好时
8. Hugo Boss
9. IWC 万国表
10. Jaeger-LeCoultre 积家
11. Lego Certified Store 乐高认证店
12. LongChamp
13. Michael Kors 迈克高仕
14. Montblanc
15. Omega 欧米茄
16. Polo Ralph Lauren 拉夫·劳伦马球男装
17. Rolex 劳力士
18. Salvatore Ferragamo 菲拉格慕
19. Swarovski 施华洛世奇
20. Swiss Watch Gallery 瑞士名表廊
21. That's a Wrap 全景店
22. The Rock Shop
23. Tory Burch
24. Tumi
25. Versace 范思哲
26. Victoria's Secret 维多利亚的秘密

**FACILITIES & SERVICES 设施与服务**
1. Adventure Cove Waterpark and Dolphin Island Guest Services 水上探险乐园和海豚岛礼宾服务
2. Member's Lounge 会员厅
3. MindChamps
4. S.E.A. Aquarium Guest Services S.E.A. 海洋馆海洋馆宾客服务
5. 7-Eleven 便利店
6. Parkway Shenton Medical Clinic 百汇康顿医疗诊所 (located in B2 & T3 3rd)
7. Universal Studios Singapore Guest Services 新加坡环球影城宾客中心
8. Universal Studios Singapore Ticketing Kiosk 新加坡环球影城票务中心
9. Universal Studios Singapore VIP Reception 新加坡环球影城贵宾接待处

**HOTELS 酒店**
1. Crockfords Tower 唐乐德豪华酒店
2. Equarius Hotel 德豪酒店
3. Equarius Ocean Suites 德豪海景套房
4. Equarius TreeTop Lofts 德豪树顶豪房
5. Equarius Villas 德豪别墅
6. Hard Rock Hotel Singapore 新加坡Hard Rock酒店
7. Hotel Michael 迈克尔酒店
8. Hotel Ora 欧秀酒店

**DINING 美食佳肴**

*Signature Restaurants 名厨餐厅*
1. Chita 秘鲁中式融合料理
2. Feng Shui Inn 风水环境品味宴饮
3. Ocean Restaurant 海之味水族餐厅
4. Osia Steak and Seafood Grill 澳西亚牛扒海鲜烧烤
5. Syun Syun春日本料理
6. tablea65
7. TungLok Heen 同乐轩

*ASIAN CUISINE 亚洲美食*
8. Malaysian Food Street 马来美食街
9. PUTIEN 莆田
10. Soi Social

*FUSION RESTAURANT 中西餐*
11. Sessions Sessions 辣椒螃蟹斯味餐厅

*BARS & LOUNGES 酒吧及酒廊*
12. Michael's Lounge 迈克尔休息厅
13. tablea65 Cocktail Bar tablea65 鸡尾酒酒吧
14. The Rock Bar 摇滚吧

*CAFÉ & QUICK BITES 咖啡厅及快餐*
15. Krispy Kreme
16. Ilaoilao
17. Starbucks 星巴克
18. Twelve Cupcakes

### LEGEND图例

- ATM 提款机
- Bus 巴士
- Shuttle Bus 接驳巴士
- Taxi 德士
- Drop-Off 下车处
- Escalator 电动扶梯
- Locker 储物柜
- Interactive Kiosk 互动查询机
- Lift 乘客电梯
- Smoking Zone 吸烟区
- Toilet 洗手间
- Family Room 育婴室
- Prayer Room 祈祷室
- Valet Parking 代客泊车
- Parking 停车场
- Carpark Kiosk 停车场查询机
- BlueSG 电动车共享服务
- EV Charging Stations by SP Group 电动汽车充电站

*Temporarily suspended until further notice
暂时关闭，直至另行通知

# Author Index


Çağatan, 72
Çano, 72
Çelikkol, 73
Štefánik, 272
Šuppa, 262
Švec, 262
Öhman, 223
Östling, 244
Üstün, 17
İnce, 410

Dankers, 80, 81
Züfle, 80, 81

AAlAbdulsalam, 112
Aalst, 361
Ababu, 89
Abadi, 89
Abboud, 318
Abdaljalil, 111
Abdalla, 170
Abdel-Salam, 114
Abdelali, 113
Abdelaty, 88
Abdelaziz, 291
Abdelhalim, 114
Abdelkadir, 89
Abdelzaher, 356
Abdo, 114
Abdul-Mageed, 110, 111, 113, 127, 302, 334, 367
Abdullah, 415
Abdulmumin, 218
Abebe, 153
Abend, 93, 142, 258, 286
Abercrombie, 308
Aberer, 270, 286
Abiderexiti, 94
Abiola, 113
Ablez, 94
Abrami, 430
Abu-Hanna, 411
Abudouwaili, 94
AbuOdeh, 111
Abzaliev, 399
Acharya, 197
Adams, 83, 413
Adauto, 421
Adebanji, 113
Adebara, 19
Adebayo, 19
Adel, 18, 114, 362
Adelani, 19, 89, 152, 218, 284
Adesam, 223
Adeyemi, 152, 284
Adhiambo, 284

Adi, 207, 219
Adithan, 208
Aditya, 17
Adly, 115
Aepli, 100
Afanasev, 83
Afonja, 281
Afzal, 400
Agarwal, 104, 147, 177, 287, 326, 368, 400, 411, 417
Agerri, 172, 337
Aggarwal, 261, 340, 342, 366, 400, 401, 405
Aggazzotti, 346
Agirre, 213, 292
Agrawal, 19, 99, 127, 150, 182, 190, 194, 231, 244, 333, 402
Aharoni, 134, 151, 203, 271, 313
Ahia, 152, 265, 284, 426
Ahmad, 113, 114, 218, 280, 284, 352, 358
Ahmadi, 217
Ahmed, 99, 113, 119, 280, 326, 380, 390, 395, 421
Ahn, 231, 251, 409
Ahrabian, 364
Ahrens, 382
Ahuja, 326
Ai, 353
Aida, 159
Ainslie, 219, 308, 309
Aisyah, 394
Aizawa, 240
Ajayi, 284
Aji, 19, 104, 138, 153, 302
Ajisafe, 284
Ajith, 183
Akavarapu, 171
Akbari, 71
Akbik, 20, 224, 351
Akbiyik, 333
Akhtar, 203, 216, 378, 388
Akiki, 173
Akimoto, 330
Akinsanya, 113
Aksu, 229
Akter, 269
Akyürek, 273
Al Moubayed, 239
Al-Fuqaha, 115
Al-Kharusi, 112
Al-Khatib, 340
Al-Matham, 111
Al-Thubaity, 110
Al-Twairesh, 19
Alabi, 89, 284
Alacam, 83, 334
Alaifi, 111
Alajrami, 150
Alam, 19, 111, 112, 118, 119, 147, 227
Alami, 112
Alaqil, 113
Alastruey, 78, 104
Alateeq, 110
Alawwad, 115
Albalak, 252, 407
Albanyan, 233
Albared, 111
Alberti, 244
Alcaide, 252
Aldarmaki, 19, 110, 111
Alessa, 113
Aletras, 20, 150, 320, 427
Alfter, 274
Algayres, 219, 243
Alhafni, 258
Alhamadani, 113
Alhammadi, 110
Alharbi, 111, 113–115
Alhassoun, 110
Alhumoud, 111
Ali, 110, 111, 115, 118, 119, 218, 401, 412
Aliannejadi, 358
Alikhani, 75, 129
Alishahi, 392
Alkanhal, 113
AlKhamissi, 114
Alkhanen, 110
Alkhereyf, 110
Allan, 380
Allaway, 411
Almahairi, 413
Almahmoud, 113
Almarwani, 112
Almasian, 174
Almazrouei, 110
Almutairi, 110
Alnefaie, 112
Alobeidli, 110
Aloisi, 160
Alon, 368, 383
AlOsaimy, 111
Aloufi, 112
AlQuabeh, 115
Alrowili, 113
Alsaleh, 112
Alsalka, 112
Alsayyahi, 173

Alshahrani, 113
Alshalawi, 110
Alshammari, 111
Alshmrani, 110
Alshomary, 410
Alsuwailem, 110
Altahhan, 112
Altamimi, 111
Altuna, 306
Alukaev, 321
Alva-Manchego, 19, 231
Alvarez-Valle, 200, 225
Alves, 200, 427
Alvez, 306
Alwajih, 110
Aly, 143, 207
Alyafeai, 113
Amalvy, 154
Amar, 307, 319
Amba Hombaiah, 399
Amelot, 248
Amin, 118, 119, 399
Amini, 350
Amiri, 366
Ammanabrolu, 19, 363
Ammar, 236
Amplayo, 20
Amrhein, 97, 100
Amroush, 115
An, 77, 190, 235, 264, 327, 363, 406
Anand, 86, 271
Anandkumar, 268
Ananiadou, 127, 287
Anastasopoulos, 56, 119, 142, 147, 153, 166
Anati, 325
Anderson, 18, 400
Andreas, 93, 277, 321
Andrews, 78, 220, 346
Androutsopoulos, 410
Ang, 246
Anikina, 285
Anish, 106
Anjum, 118
Ansell, 147
Antao, 301
Antetomaso, 106
Anthony, 184, 252
Antoniak, 18
Antypas, 382
Anuva, 390
Anwer, 113, 204
Ao, 285
Appalaraju, 206
Araabi, 76
Arakelyan, 373
Araki, 17, 343
Aralikatte, 72
Arase, 175
Arbelle, 131
Arcadinho, 252
Aremu, 89, 284
Arias, 256
Arik, 164, 317, 420
Arkhipkin, 352
Arnett, 194
Arora, 236, 308, 328, 336, 346, 359, 374
Aroyo, 258, 396
Arras, 18
Artetxe, 19, 213, 256
Arthur, 218, 284
Artzi, 219
Arviv, 286
Asai, 19, 284, 358
Asami, 71
Asgari, 20
Ash, 108, 265, 426
Asher, 245
Ashpole, 305
Ashraf Vaghefi, 129
Askari, 358
Asl, 385
Assabie, 88
Assenmacher, 361
Assent, 313
Assogba, 100
Astudillo, 299
Atanasova, 18, 300
Atri, 348
Attanasio, 327
Atwell, 112
Atzeni, 140
Augenstein, 20, 300, 336, 346, 357, 361
Auli, 158
Aumiller, 231
Aunti, 136
Authur, 262
Auvray, 162
Avramidis, 75, 76, 97
Aw, 180, 207, 225, 231
Awadalla, 341, 388
Awadallah, 189
Awokoya, 284
Axmed, 20
Ayed, 355
Ayele, 88, 89, 218
Ayyubi, 253

Azadi, 97
Azeemi, 429
Aziz, 300
Azizah, 240
Azizov, 113–115
ASSenmacher, 256

B a t i s t a - N a v a r r o, 98
Bénard, 77
Baan, 300
Babu, 158
Badel, 270
Bader, 384
Bae, 185, 232, 273, 385, 403
Baek, 224, 278, 390, 412, 429
Baes, 82
Baesens, 235, 302
Baevski, 158
Bagdasarov, 76
Bahaaulddin, 113, 114
Bahdanau, 81, 213, 362
Baheti, 250
Bai, 155, 197, 220, 230, 282, 310, 312, 325, 380, 416
Bajracharya, 243
Bak, 194, 304, 357, 390
Baksi, 107
Balachandran, 56, 325
Balalau, 158
Balashankar, 108
Balasubramanian, 17, 217, 280
Balch, 150
Balcha, 89, 218
Baldo, 110
Baldwin, 107, 262, 284, 294, 378, 394
Balepur, 185
Bali, 262, 326
Ball, 280
Ballesteros, 196, 314
Ballier, 77
Bambauer, 107, 108
Bamman, 354
Banaei, 286
Bandyopadhyay, 99
Banerjee, 428
Bang, 232, 339
Bangalore, 400
Bangoura, 77
Banitalebi-Dehkordi, 71
Bannur, 200, 225
Bansal, 119, 196, 208, 273, 300, 311, 326, 335, 364, 425
Bao, 194, 233, 247, 249, 283, 293, 304, 309, 381
Bar-Haim, 306, 406
Bar-Yossef, 381
Barale, 107, 108
Baraniuk, 383
Baranov, 306
Barba, 330
Barbieri, 382
Barbour, 395
Barka, 113
Barlacchi, 297
Barman, 226
Baroni, 251, 300
Barrault, 18, 219
Barriere, 324
Barry, 112
Bartolo, 182
Bartsch, 86
Barzilay, 389
Barzon, 82
Basak, 331
Basile, 20, 148
Basiou, 162
Basit, 227
Bast, 233
Bastings, 18, 210
Basu, 342
Basu Roy Chowdhury, 421
Basulto, 357
Bathala, 185
Batista-Navarro, 114, 173, 221, 287
Bau, 94
Baucells, 295
Bauer, 375
Baumann, 237
Baumartz, 430
Baumgartner, 108
Baumler, 150
Bawden, 18, 75, 76
Baylor, 297
Bayo, 77
Bazzoli, 108
Bean, 324
Bearman, 381
Beauchamp, 361, 387
Beauchemin, 85
Beck, 83, 98, 336
Bedekar, 422
Begus, 20
Behjati, 295
BehnamGhader, 287
Beinborn, 95, 145, 320

Beirami, 321
Bejan, 221, 353
Bekal, 397
Bel, 411
Beladev, 405
Belapurkar, 342
Belay, 88, 89, 218
Belhaj, 113, 114
Belinkov, 313, 322, 329
Bellet, 315
Bellew, 85
Beloucif, 19, 218, 237
Beltagy, 19
Belyi, 403
Belyy, 130
Ben Rim, 80
Ben-Michael, 316
Benajiba, 196
Bendersky, 399
Benelallam, 88
Bengtson, 382
Bennett, 189, 358
Benotti, 90
Bensalem, 113
Bentivogli, 76, 127, 274
Bentz, 71, 244
Berant, 224
Berchansky, 271
Berdicevskis, 223
Berg-Kirkpatrick, 93, 273
Bergen, 194, 340
Berlowitz, 243
Bernstein, 354
Berrada, 112
Bertsch, 157, 393, 396
Besacier, 18
Bethard, 208
Betka, 113
Beutel, 321
Bhagavatula, 250, 384
Bhambhoria, 107, 162
Bhar, 245
Bhaskar, 234
Bhat, 227, 265, 319, 398, 425
Bhatia, 110, 168, 340
Bhattacharjee, 174, 396
Bhattacharya, 80, 107, 108, 171
Bhattacharyya, 97, 99, 156, 199, 210, 225, 327, 347, 388, 426, 428
Bhattamishra, 213
Bhaumik, 242
Bhosale, 256
Bhushan TN, 398
Bhuta, 107, 108
Bi, 140, 224, 264, 414
Bian, 178, 233, 306, 429
Biderman, 184, 252
Bielaniewicz, 142
Bielikova, 303, 391
Biemann, 89
Biester, 20
Bigham, 417
Bijoy, 119
Bikel, 19
Bilen, 137
Billah Nagoudi, 367
Bin, 126, 235, 375
Bing, 20, 249, 279, 304, 307, 312, 409
Bingler, 129, 258
Birch, 18, 78, 200, 330
Birim, 112
Bisazza, 79, 80, 147, 200
Bishop, 346
Bisk, 169
Bissyandé, 210
Biswas, 99, 162
Bitton, 134, 288
Bjerva, 297
Blain, 97, 388
Blakeney, 17
Blanchard, 20
Blanco, 20, 207, 233, 360, 383, 385
Blevins, 287
Blinder, 414
Blodgett, 17, 295
Blukis, 18
Blum, 83
Blume, 401
Blumm, 321
Bodapati, 397, 403, 404
Bodenreider, 332
Bogin, 224
Bogoychev, 99
Bohnet, 203
Boholm, 82
Bohra, 341
Boisson, 275
Bojar, 75, 76, 99
Bolding, 159
Bollegala, 159, 323
Bolliger, 140
Bollmann, 85
Bolukbasi, 191
Bommasani, 20
Bonial, 150
Bonn, 395

Bono, 418
Bontcheva, 339, 363
Boo, 129
Borchert, 172, 235, 302
Borenstein, 357
Boreshban, 217
Borin, 223
Borse, 174
Borsos, 243
Boruah, 99
Boschetti, 19
Bosco, 148
Bosselut, 168, 270, 286, 289, 321
Bothwell, 276
Botzer, 161
Bouamor, 111
Bouchard, 325
Boudiaf, 355
Boudin, 20
Bougares, 115
Boughorbel, 110
Boujelbane, 114
Bouklouha, 113
Bouma, 223
Boumber, 111
Bouraoui, 345, 378
Bout, 361
Bouthors, 190
Boutiba, 113
Bouyamourn, 315
Bouyarmane, 399
Bouzid, 225
Bowden, 75
Bowen, 107
Boyd-Graber, 20, 196, 325, 373, 407
Boyden, 83
Boyle, 322
Brachat, 182
Braffort, 75
Bragilovski, 108
Brahma, 308, 408
Brahman, 287, 363, 424, 425
Brandl, 146, 318
Bransom, 148, 262
Brantley, 17, 303
Bras, 17, 129, 250, 259
Braslavski, 306
Brazdil, 218
Breazeal, 261
Brentari, 77
Briakou, 127, 150
Bridgers, 412
Brimacombe, 385
Briot, 169
Broscoteanu, 416
Brower-Sinning, 205
Brown, 218, 219, 325, 400
Browne, 271
Bruckner, 80
Bruni, 72, 80
Brunila, 355
Brunner, 108
Brusilovsky, 181
Bruton, 237
Bryan, 308
Bu, 134, 229
Buaphet, 80
Bucur, 89
Budak, 90
Bugliarello, 78, 145, 146
Bugueño, 412
Bui, 86, 161, 213, 339
Bulatov, 427
Bunzeck, 94
Burger, 315
Burja, 216
Burke, 395
Burns, 219
Burtsev, 18, 364, 427
Butoi, 275
Buttery, 95
Buys, 18
Buzaaba, 284
Bylinina, 345
Bölücü, 291
Börjeson, 223

C o s t a - J u s s à, 78
C. De Souza, 97, 98
Cabello, 146, 318
Cabrio, 20, 346
Caciularu, 198, 203, 259, 271, 333
Cacoullos, 149
Cahyawijaya, 104, 153, 302
Cai, 82, 93, 146, 163, 176, 188, 190, 209, 240, 243, 266, 303, 360, 395, 409
Caillout, 99
Caines, 95
Calderon, 271, 324
Calixto, 411
Callan, 237
Callison-Burch, 85, 422
Calvo, 113
Cam-Tu, 234, 335
Camacho-Collados, 275, 288, 323, 382
Camassa, 106

Cambria, 295
Cambronero, 257
Campagna, 341
Campbell, 323, 414
Campesan, 110
Campese, 286
Campos, 292
Canby, 162
Cancedda, 140
Candel, 130
Candra, 262
Canny, 134, 272
Canute, 413
Cao, 17, 107, 126, 136, 138, 165, 187, 189, 198, 214, 219, 230, 241, 243, 252, 266, 268, 271, 290, 291, 295, 298, 304, 305, 307, 319, 325, 372, 378, 383, 398, 425
Caragea, 86, 281, 361, 365
Carbune, 72
Card, 18
Cardenas, 315
Cardie, 242, 426
Carenini, 107, 158
Carlson, 308, 375
Carnahan, 325
Carpuat, 17, 127, 134, 150, 182, 196, 314
Carvalhais, 130
Casacuberta, 97, 98
Caselli, 295
Casola, 148
Casper, 321
Castelli, 134, 196, 314
Castellucci, 18
Castricato, 85, 184
Castro, 225
Caswell, 179
Catanzaro, 268
Cattan, 306
Caverlee, 195, 340, 413
Cegin, 181
Celikyilmaz, 168, 257, 325, 386
Ceron, 364
Cettolo, 127
CH-Wang, 272, 355, 357
Chadha, 104, 226, 227, 249, 267
Chae, 240
Chai, 152, 272, 285, 334, 375, 381
Chakrabarti, 327
Chakrabarty, 56, 360, 412
Chakraborty, 17, 216, 226, 227, 267, 327, 330, 332, 348, 388
Chakravarthy, 106
Chala, 218
Chalamalasetti, 269
Chalkapurkar, 342
Chalkidis, 107, 157, 318, 320
Chambers, 217, 280
Chamoun, 373
Chan, 134, 231, 272, 309, 328, 356, 359, 393, 397
Chandar, 191, 393
Chandra, 152
Chandrasekar, 373
Chandrasekhar, 262
Chandu, 236, 250, 363
Chang, 18, 82, 99, 130, 137, 143, 144, 151, 166, 185, 194, 196, 236, 247, 249, 253, 262, 269, 272, 281, 307, 309, 310, 314, 322, 352, 354, 375, 402–404, 409
Changpinyo, 143, 248
Chao, 175, 200, 306, 328
Chaoji, 297
Chaovavanich, 85
Charnois, 110, 111
Charpentier, 94
Chatterjee, 97, 104, 223, 227, 345
Chaturvedi, 245, 343, 421, 424
Chaudhary, 20, 127, 166, 287, 400
Chaudhuri, 18
Chaudhury, 291
Chauhan, 400
Chawla, 187
Chazan, 225
Che, 138, 267, 302, 416
Cheang, 328
Chelliah, 428
CHen, 187
Chen, 17–19, 76–78, 82, 83, 89, 90, 93, 98, 99, 107, 108, 125, 126, 128–132, 135, 137, 138, 140, 142–146, 149, 154, 155, 159, 161, 164, 165, 167, 172–178, 180–183, 185–194, 199–204, 206, 209, 210, 214, 219, 222, 224–226, 230, 233, 235, 237, 238, 240, 241, 243, 246–250, 253, 257, 258, 264–266, 269, 274, 275, 278–281, 283, 284, 291–294, 296, 299, 302, 307, 311, 313, 315, 318, 320, 321, 323, 328, 329, 331, 332, 339, 345, 347, 348, 352, 353, 359, 360, 362, 363, 365–367, 369, 370, 374, 376–379, 387, 390, 392–394, 397–399, 401, 402, 409, 413, 416, 417, 419–422, 424, 427, 428

Cheng, 17, 18, 72, 129, 139, 150, 160, 175, 177, 209, 212, 216, 222, 224, 227, 233, 234, 246, 249, 252, 264, 265, 275, 297, 300, 305, 307, 309, 310, 315, 316, 326, 332, 336, 340, 353, 359, 387, 397, 402, 406, 409, 417, 427
Cheon, 292
Chepurova, 427
Cherry, 18
Chersoni, 20, 83
Chetlur, 404
Cheung, 72, 295, 337
Chevalier, 183
Chew, 336
Chheang, 246
Chhel, 169
Chi, 231, 281
Chiang, 275, 276, 356, 367, 383, 385
Chien, 158
Chieu, 292
Chilimbi, 230
Chiniya, 355
Chinmay, 249
Chinnappa, 207
Chitale, 99
Chitta, 108
Chiu, 142, 173, 242, 404
Chng, 288
Cho, 150, 216, 236, 290, 295, 308, 389, 401, 409, 412, 422, 425
Chobey, 72
Choe, 236, 251
Choenni, 141, 411
Choi, 19, 107, 129, 160, 168, 188, 191, 212, 219, 236, 239, 250, 259, 270, 278, 287, 361, 363, 384, 387, 390, 415, 422, 424, 425, 430
Cholakkal, 113, 204
Chollampatt, 190, 317
Chong, 323
Choo, 160, 429
Chopra, 285
Chormai, 85
Choshen, 93, 142, 289, 349
Chou, 158
Choudhury, 138, 300, 400, 411
Chowdhary, 170
Chowdhery, 196, 330
Chowdhury, 20, 111, 119, 142, 256, 329, 382, 405, 412
Christen, 108
Christodoulopoulos, 17, 160
Christofidellis, 157
Christoph, 328
Chronopoulou, 338
Chrupała, 392
Chu, 128, 146, 177, 245, 376
Chua, 107, 214, 233, 319, 417, 425, 427
Chuang, 298
Chuangsuwanich, 281
Chukwuneke, 284
Chulvi, 89
Chun, 337, 413, 430
Chung, 134, 137, 196, 212, 252, 290, 398
Cignarella, 148
Cihan Camgöz, 75
Cimiano, 194
Ciosici, 19
Cissé, 77
Clapper, 322
Clark, 18, 19, 100, 151, 156, 184, 220, 226, 271, 284, 321, 324
Clavel, 289
Clement, 19, 188, 268
Coavoux, 108
Cocarascu, 203, 378
Cohan, 18, 246, 277, 311, 324, 384, 402
Cohen, 161, 227, 272, 319, 340, 374, 376, 379, 396
Cohen-Ganor, 134
Coheur, 98, 296
Cohn, 107, 163, 282, 358, 378, 400
Cojocaru, 110
Cole, 237, 361
Colen, 399
Colesanti-Senni, 129
Coley, 389
Collier, 176, 393
Collins, 258, 331
Colombo, 184, 200, 301, 355, 427
Colon-Hernandez, 261
Colunga, 369
Coman, 297
Comi, 157
Compher, 430
Comsa, 280
Conde, 77, 130
Cong, 135
Conger, 185
Conia, 276
Conneau, 158
Constantin, 351
Constantinides, 316
Conti, 275
Contractor, 265

Copet, 219
Corro, 243, 350
Cosma, 89
Costa, 186
Costa-jussà, 219, 220, 417
Cotterell, 149, 154, 156, 196, 198, 215, 275, 278, 350
Courville, 275
Coussement, 172
Coustaty, 182
Cowap, 161
Cramer, 354
Crego, 190
Crestani, 18
Creutz, 80
Cripwell, 361
Cristea, 231
Crnkovich, 276
Croce, 18
Crook, 17
Crouse, 291
Cruz, 75, 104, 302
Csordás, 337, 364
Cui, 136, 140, 178, 200, 225, 230, 253, 291, 295, 304, 365, 371, 416
Culhane, 238
Curry, 308
Curtis, 85
Côté, 195

D'Arcy, 277
D'Haro, 417
D'Oosterlinck, 331
Dönmez, 93
Da San Martino, 111
Dabre, 99, 180, 285
Dada, 428
dadadadada, 55, 56
Dadure, 97
Dagan, 198, 271, 306, 307, 319, 333, 335
Dahan, 107, 162
Daheim, 329
Dahmani, 113
Dahou, 80, 81
Dai, 140, 144, 161, 215, 263, 267, 278, 309, 313, 317, 336, 383, 399, 420
Dainese, 156
Dalal, 227
Dale, 219
Dalton, 242, 286
Dalvi, 18, 211
Damavandi, 381
Dandapat, 197
Dandeniya, 89
Danial, 111
Dankers, 288, 313
Dannélls, 223
Dao, 140
Darrell, 131, 272
Darrin, 184
Das, 99, 104, 106, 118, 152, 226, 227, 244, 248, 258, 267, 271, 287, 305, 373
Dasgupta, 330
Dash, 97
Dasigi, 19
Datta, 226
Dau, 86, 339
Daumé III, 17, 134, 150, 314, 342
Dave, 227
David, 218
Davila, 17
Davis, 71, 95
De, 108, 331
de Araujo, 280
De Bruyne, 299
De Caigny, 172
de Castro, 284
De Clercq, 299
de Gispert, 297
de Jong, 308, 309
de la Iglesia, 381
De la Peña Sarracén, 311
de Lacalle, 292
de Langis, 71
De Melo, 78
de Melo, 302, 412
de Rijke, 206, 209
De Silva, 367
de Souza, 191, 427
de Valk, 93
de Vries, 226
De Weerdt, 172, 302
de-Dios-Flores, 19
Deas, 324
DeBenedetto, 94, 276
Degan, 404
Deguchi, 75, 129
Dehan, 119
Dehghani, 196, 349
Dehouck, 82, 94
Deka, 114
del Grosso, 19
Deleu, 331
Deligianni, 286
Dell, 308, 359
Demberg, 19

Demeester, 331
Demszky, 322
Deng, 18, 144, 164, 167, 171, 178, 183, 185, 200, 233, 238, 243, 253, 261, 264–266, 270, 274, 279, 282, 290, 291, 336, 345, 358, 365, 371, 387, 390, 414, 418
Denis, 159, 202, 233, 315
Deoghare, 99, 388
Derczynski, 252, 313
Dernoncourt, 217
Derouich, 114
Desarkar, 408
Desbordes, 216
Deshpande, 106, 113, 114, 237, 286, 383
Dessi, 413
Dettmers, 17, 176
Deutch, 224
Deutsch, 78, 97–100, 196
Dev, 20
Devare, 217
Develder, 331
Devereux, 301
Dey, 113
Dhawan, 105
Dhingra, 149, 237, 327
Dhuliawala, 261
Di Gangi, 78
Di Marco, 317
Di Nunzio, 75
Diané, 77
Diao, 221, 328, 329
Dibia, 189
Diddee, 262, 326
Dieke, 113
Diep, 342
Diera, 80, 81
Diesner, 86, 322
Dietze, 407
Dijck, 106
Dikkala, 264
Dimitriadis, 422
Dimitrov, 352
Dinarelli, 77
Ding, 126, 165, 198, 204, 218, 222, 240, 252, 260, 268, 278, 289, 299, 309, 363, 375, 396, 417, 418
Dingliwal, 404
Dinh, 351
Dinkar, 308
Dinu, 231, 396
DIOP, 284
Djanibekov, 110
Djeradi, 114
Do, 312
Doan, 127, 312
Dobler, 302
Dodeja, 206
Dolan, 303
Domhan, 75, 99
Domingo, 98
Domingo-Fernández, 185
Don-Yehiya, 142, 349
Donatelli, 222, 272
Dong, 20, 135, 137, 156, 159, 162, 178, 191, 207, 221, 240, 252, 266, 268, 310, 311, 356, 366, 387, 394, 413, 418, 424
Donnelly, 108
Doshi, 99
Dossou, 89, 281, 284
Dou, 152, 214, 218, 267, 274, 325, 357, 416
Doumbouya, 77
Downey, 148, 262, 277, 306
Doğruöz, 285
Dragut, 281
Dreano, 97, 98
Dreyer, 19, 140, 311, 325
Dror, 19
Drozdov, 195, 383
Drummond, 163
Du, 126, 132, 139, 168, 182, 190, 198, 208, 252, 269, 279, 307, 347, 363, 370, 426
Duan, 86, 93, 132, 201, 229, 356, 359, 382, 421
Dube, 287
Dubey, 18, 421
Duderstadt, 86
Duesterwald, 133
Dufour, 154
Dugan, 85
Duh, 78, 383
Dulka, 284
Dunstan, 89
Dupoux, 219, 243
Dupuy, 398
Durmus, 281
Durrani, 211
Durrett, 18, 93, 167, 177, 217, 345
Durumeric, 273
Dusek, 119, 362
Dutt, 240
Dutta, 327, 345, 353
Dvorkovich, 75
Dwivedi, 277, 403

Dwivedi-Yu, 170, 237
Dzialo, 403
Dziri, 17, 287, 363
Düsterhus, 331

E, 145
E s p a ña - B o n e t, 75
E s p e r a n ç a - R o d i e r, 77
E Sobhani, 119
Eaton, 89
Eberle, 318
Ebling, 75, 250
Ebrahem, 114
Ebrahimi, 164
Eckman, 336
Eckstein, 175
Eden, 306, 406
Eder, 279
Edwards, 283
Eetemadi, 83, 89, 90, 115
Eger, 98, 420
Egger, 429
Ehrenworth, 83
Ehsan, 115
Eickhoff, 318, 370
Ein-Dor, 303
Eisenschlos, 18, 237, 317
Eisenstein, 18, 237
Eisner, 391
Ekbal, 17, 155, 216, 279, 280
El Khbir, 111
El Mesbahi, 72
El-Assady, 261
El-Kishky, 273
El-Kurdi, 299
El-Makky, 115
El-Sayed, 113
El-Shangiti, 110, 113, 334
Elbakry, 114
Elbayad, 246
Elenberg, 298
Elgaar, 366
Elhadad, 205
Elkaref, 112
Elkhbir, 110, 111
Elkind, 313
Elkomy, 115
Elkordi, 115
Ellershaw, 331
Elleuch, 115
Elliott, 78, 146, 192, 357
Elluru, 397
Elmadany, 110, 111, 113, 114, 334, 367
ElNokrashy, 114
Elnokrashy, 98
Elsayed, 111, 112
Elshabrawy, 111
Emerson, 107, 222, 393
Emerton, 108, 293
Emezue, 281, 284
Emonet, 349
Engländer, 173
Eo, 221, 222
Erdem, 410
Erden, 112
Erk, 217
Ermis, 17, 227, 333
Ernandes, 344, 403
Ernst, 319, 335
Ersoy, 382
Erwin, 291
Escolano, 417
Eshghi, 148
Esiobu, 170
Eskelinen, 301
España-Bonet, 252, 256
Espinosa-Anke, 275, 345, 382
Espinoza, 346
Estève, 19
Etori, 281
Ettinger, 20, 211, 384
Etxaniz, 292
Eugenio, 86
Evci, 17
Evuru, 136
Ezeani, 284
Ezra, 108

Fabbri, 177, 329
Fadeeva, 262
Fahim, 119
Fainman, 405
Faisal, 153
Falenska, 20
Falissard, 379
Faltings, 303
Fan, 18, 19, 133, 155, 165, 167, 192, 199, 207, 209, 216, 225, 229, 251, 256, 260, 296, 310, 355, 360, 374, 398, 401, 420
Fanconi, 313
Fang, 18, 72, 107, 135, 137, 181, 191, 204, 231, 239, 264, 266, 289, 328, 373, 378, 381
Farah, 295
Faria, 119

Farinhas, 157, 191
Farooq, 110
Farooqui, 107
Farrús, 411
Fatahi Bayat, 130
Fateen, 110
Fathi, 83
Fatima, 227
Faust, 330
Faysse, 301
Federico, 19, 238
Federmann, 75, 98
Fedorenko, 215
Fedyanin, 262
Fehr, 295
Fei, 251, 319, 321
Feldhus, 285
Feldman, 113, 148, 277
Feng, 17, 18, 86, 126, 130, 136, 140, 144, 155, 163, 169, 181, 199, 214, 232, 239, 242, 252–254, 260, 269, 284, 294, 304, 325, 338, 343, 370, 386, 387, 410, 414, 416, 426, 428
Fenogenova, 157
Ferdoush, 119
Fergadiotis, 85
Ferhat, 113
Feris, 131
Fernández, 79, 80
Fernandes, 100, 157, 170, 283
Fernandez, 169
Fernandez Astudillo, 93
Fernando, 137
Fernández, 147–149, 215, 300
Ferrando, 100, 417
Ferrara, 368
Ferraro, 18, 217, 280
Fetahu, 250, 402
Fiedel, 330
Field, 17
Fields, 94
Figueras, 295
Filice, 19
Filippova, 210, 353
Finkelstein, 78, 98–100
Finlayson, 324
Finn, 197, 368
Firat, 18, 100, 179
Firdaus, 341
Firdous, 76
Firooz, 138
Fishel, 19, 75
Fisher, 363
Fisichella, 297
Flanigan, 90
Fleisig, 153, 326
Flores, 246
Florez, 273
Florian, 299, 354
Fok, 324
Fokkens, 18, 320, 334
Fokoue, 291
Folny, 274
Forde, 104
Forkel, 83
Foroosh, 308
Foroutan, 286
Forristal, 93
Forsberg, 223
Forte, 107, 108
Fosler-Lussier, 166
Foster, 19, 97, 107, 161, 196, 357
Fowlie, 272
Fox, 382
Fradet, 169
Fraiberger, 284
François, 82
Frank, 80, 301
Fransen, 430
Franz, 354
Franzon, 251
François, 274
Fraser, 17, 317, 338
Frassinelli, 71
Frayerman, 405
Freedman, 152
Freihat, 111
Freitag, 19, 75, 78, 97–100, 196, 283, 303
Freitas, 220, 374
Frenda, 148
Frermann, 17, 107, 378
Fried, 18, 142, 204, 230, 377, 407
Friedl, 398
Friedler, 384
Friedman, 90
Friedrich, 429
Frost, 80, 220
Fu, 86, 128, 140, 175, 176, 178, 179, 199, 200, 221, 248, 252, 264, 266, 304, 310, 317, 369, 371, 380, 398, 409
Fucci, 127, 274
Fujimoto, 339
Fujinuma, 19, 206, 314
Fukatsu, 72
Fukumoto, 295
Funakoshi, 346

Funayama, 211
Fung, 134, 158, 252, 283, 332, 339, 356, 360
Funkquist, 223
Furman, 319
Futrell, 156
Fyshe, 150
Féré, 355

G, 227
Göhring, 75
Günther, 85
Gaanoun, 88
Gabr, 114
Gaido, 76, 127
Gaikwad, 99
Gajbhiye, 223, 345
Gala, 99
Galappaththi, 89
Galassi, 107, 157
Gales, 263
Galke, 80, 81
Galley, 303
Gallé, 18
Galstyan, 401
Gan, 18, 108, 249, 275, 334, 335, 407
Ganapathy, 191, 228
Ganchev, 244
Gandhi, 155
Ganguly, 17, 408
Gantt, 178, 284
Ganu, 326
Gao, 20, 71, 108, 126, 128, 130, 134, 135, 139, 150, 163, 168, 179, 182, 202, 214, 219–221, 224, 231, 232, 237, 245, 246, 251, 277, 283, 296, 303, 304, 323, 329, 358, 362, 377, 402
Garain, 245
Garcia, 104, 174, 196
Garcia Contreras, 297
Garcia-Olano, 86
García-Ferrero, 292, 306, 337
Gardent, 361, 379
Gardner, 381
Garera, 428
Garg, 19, 100, 133, 400
Garimella, 20, 195, 365, 375
Garland, 383
Garmash, 345
Garner, 209
Garrette, 19, 141
Gashteovski, 181
Gasic, 130
Gaur, 147
Gautam, 160, 175, 226
Gauthier, 215
Ge, 131, 247, 322, 328, 358, 416
Gebremichael, 218
Gee, 310, 403
Gehrmann, 271
Geishauser, 130
Gekhman, 305, 313
Gella, 148, 333
Gemmell, 242
Gemulla, 412
Geng, 98, 182, 282, 365, 406
Gentili, 272
Georgescu, 231
Gera, 303
Gertz, 174
Gessler, 72
Getane, 89
Geuter, 85
Geva, 18, 168, 203, 210, 227, 428
Ghaddar, 72
Ghaffari, 86
Ghaisas, 106
Ghalandari, 86
Ghanekar, 343
Ghanem, 112
Ghareeb, 318
Ghasemi Madani, 94
Ghassem-Sani, 217
Ghazvininejad, 142
Ghodsi, 331
Gholamian, 107
Ghorbani, 283
Ghosal, 19, 155, 181
Ghosh, 90, 93, 108, 134, 136, 147, 217, 226, 264, 279, 280, 298, 342, 355, 391
Giannotti, 407
Gibson, 71
Giles, 384
Gilleron, 202
Gillick, 237, 361
Gillis, 355
Gimpel, 134
Ginger, 318
Gini, 72
Ginn, 79
Ginsburg, 105
Ginter, 301
Giorgi, 85, 384
Gipp, 170, 235
Girgin, 243
Giulianelli, 72, 215, 300
Gkoumas, 128, 372

Glass, 298
Glavaš, 20, 352
Glavaš, 181, 311, 345
Gligoric, 17
Globerson, 131, 210, 227, 428
Glockner, 282
Goanta, 320
Godbole, 221
Godbout, 328
Godfrey, 325
Goel, 137, 236, 367
Goffredo, 346
Goharian, 19
Goldberg, 250, 259, 284, 340, 379
Golde, 351
Goldfarb-Tarrant, 311
Goldman, 198, 249, 259
Goldsack, 180, 328
Goldwater, 154
Gollan, 104
Gollapalli, 246
Gombolay, 206
Gomez-Cabrero, 392
Goncharova, 262
Gonen, 142, 265, 287, 426
Gong, 201, 229, 298, 314, 354, 356, 420, 428
González, 77
González Hernández, 106
González Juclà, 106
Gonzalez, 247, 272
Gonzalez-Dios, 306
Gonçalves, 169
Good, 398
Goodarzi, 413
Gooding, 72
Gopalan, 404
Gorelik, 108
Gori, 110
Goriely, 95
Gorinski, 18
Gostlow, 129
Goswami, 133, 246, 256, 408, 430
Gotmare, 213
Gou, 234, 335
Gourru, 349
Gowda, 75, 86, 98
Goyal, 20, 142, 150, 222, 331, 345, 348, 408, 428
Grabmair, 107, 108, 154, 278
Graf, 90, 237
Graham, 93, 161
Grant, 108, 293
Grari, 338
Gratch, 187
Gravier, 349
Gray, 291
Greco, 106
Greene, 236
Grella, 252
Grenon-Godbout, 338, 406
Gribomont, 83
Grieser, 324
Grinberg, 108
Gritta, 143, 282
Grobol, 85
Groschwitz, 272
Grubisic, 335
Grundkiewicz, 75, 86
Gruza, 142
Grünwald, 420
Gu, 75, 128, 133, 169, 178, 179, 188, 200, 202, 209, 259, 268, 277, 287, 307, 353, 410, 426, 429
Guan, 234, 341
Guerin, 311
Guerini, 271, 281
Guerreiro, 97, 98, 200, 427
Gueta, 289
Guha, 108
Gui, 160, 247, 251, 296, 343, 347, 366
Guigue, 379
Guillou, 97
Gulati, 287
Gulla, 374
Gulwani, 257, 429
Gunapati, 217
Gundroo, 343
Gung, 134, 342
Gunter, 219
Guntuku, 85, 216
Guo, 19, 76, 77, 86, 119, 126, 132, 163, 176, 181, 193, 203, 209, 210, 219, 231, 234, 237, 238, 266, 283, 284, 289, 305, 307–309, 314, 319, 320, 336, 339, 347, 349, 351, 360, 398, 401, 403, 423, 427, 429
Gupta, 18, 104, 133, 175, 189, 197, 222, 226, 236, 241, 251, 274, 328, 331, 333, 349, 358, 377, 378, 381, 390, 398, 400, 417, 423, 428, 429
Gur, 330
Gurevych, 127, 173, 179, 223, 241, 282, 329, 345, 373
Gurumurthy, 227
Gururaja, 393
Gutierrez, 332

Gutierrez-Vasques, 244
Gutowski, 169
GV, 252
Gwadabe, 218, 284
Gwak, 160
Gygli, 108
Göldner, 174
Götze, 269
Gómez-Rodríguez, 350

Ha, 19, 193
Habash, 111, 258, 422
Haber, 262
Hacheme, 89, 284
Hackinen, 107
Hada, 262, 326
Haddad, 107
Haddadan, 210
Haddow, 19, 200, 376
Hadfield-Menell, 321
Hadgu, 89
Hadjer, 113
Haf, 238, 424
Haffari, 19, 72, 130, 350
Haga, 72
Hagag, 339
Hagström, 193
Hahn, 20, 156, 174, 292, 372
Hahnloser, 179
Hai, 86
Hai Long, 107
Hajialigol, 395
Hajishirzi, 168, 212, 287, 358
Hajizadegan, 83
Hakami, 112
Hakim, 208
Hakimi, 134, 340
Hakimi Parizi, 107
Hakimov, 269
Hakkani-Tur, 148, 229, 333
Hale, 324
Halim, 429
Haller, 140, 351
Hallinan, 363, 425
Halterman, 17
Hamad, 111
Hamborg, 351
Hamed, 111, 422
Hammouda, 110
Hammoudeh, 414
Hamoud, 111
Hamri, 227
Han, 17, 19, 126, 130, 144, 158, 168, 169, 178, 185, 187, 189, 196, 212, 218, 221, 235, 248, 258, 266, 268, 283, 285, 291, 292, 295, 311, 324, 341, 351, 358, 362, 369, 372, 379, 385, 395, 403, 408, 414, 430
Handa, 151, 322
Handler, 17
Hanif, 183
Hanley, 273
Hanna, 94, 322
Hansanti, 78, 219, 220
Hansen, 77
Hao, 108, 138, 187, 207, 245, 259, 283, 390, 401
Haq, 112, 210
Haque, 99
Harabagiu, 346
Haraguchi, 203
Harris, 415
Harshvardhan, 205
Harutyunyan, 341
Harwath, 20
Hasan, 100, 118, 401
Hasanain, 111
Hasegawa, 139
Hashimoto, 18, 213, 281
Hasib, 119
Haslam, 82
Hasler, 99
Haslum, 225
Hassan, 17, 18, 233, 373
Hassanpour, 164
Hatekar, 114
Hauer, 256
Hauptmann, 399
Havaldar, 180
Havens, 375
Havrilla, 184
Hawasly, 110
Hayashi, 20
Hazarika, 17, 229, 333
He, 18, 55, 56, 71, 77, 93, 113, 130, 136, 153, 160, 163, 165, 178, 188, 193, 194, 209, 229, 230, 232, 236, 241, 245, 249, 252, 253, 260, 266, 269, 272, 273, 293, 294, 298, 304, 306, 317, 327, 343, 346, 356, 358–360, 364, 366, 389, 390, 402, 406, 422
Hearst, 262
Heck, 130
Heiliger, 428
Heineman, 218, 274

Heinisch, 194
Heinonen, 301
Heinzerling, 382
Held, 143, 151
Hemanthage, 148
Hemmer, 182
Henao, 18
Hendel, 428
Henderson, 295
Hendler, 298
Hendley, 112
Hendricks, 145
Hendrycks, 280
Hengchen, 223
Hengle, 240
Henry, 94
Heppell, 363
Heras, 174
Hergul, 282
Heri, 278
Hermann, 331
Herold, 77
Herrera-Berg, 271
Hershcovich, 210
Hertel, 233
Herzig, 19, 131, 151, 203, 313
Hessel, 18, 129, 212, 250, 314, 322
Heumann, 256
Hewitt, 18
Hidey, 398
Higashinaka, 241
Hill, 83
Himakunthala, 153
Hinrichs, 200
Hirao, 248
Hirasawa, 78
Hirsch, 333
Ho, 107, 138, 209, 288
Hockenmaier, 162, 290
Hoeken, 83, 334
Hoffman, 421
Hofmann, 240
Hofmann-Coyle, 323
Hogan, 177
Hoi, 137, 183, 213, 312
Hoiem, 172
Hokamp, 86
Holat, 111
holtzclaw, 413
Holzenberger, 108
Homan, 353
Hong, 94, 165, 166, 194, 204, 207, 225, 259, 290, 295, 307, 360, 379, 417
Hooi, 93, 176
Hooker, 227, 333, 419
Hope, 306, 421
Hoque, 55, 119, 152, 161
Horbach, 19
Horiat, 20
Hoshi, 129
Hosking, 243
Hossain, 383
Hosseini, 170, 340
Hou, 126, 127, 142, 152, 169, 175, 179, 204, 223, 252, 262, 297, 315, 321, 413, 416
Houlsby, 196
Houndayi, 89
Hourrane, 218
Hovsepian, 367
Hovy, 17, 283, 331
Howard, 208
Hoyle, 265, 367
Hoyos-Idrobo, 375
Hraška, 262
Hromadka, 391
Hruschka, 18
Hsieh, 359, 402
Hsiung, 402
Hsu, 20, 158, 159, 272, 336, 384
Hu, 76, 110, 125, 135, 136, 143–145, 152, 156, 163, 170, 172, 173, 190, 191, 193, 197, 202, 222, 244, 259, 269, 270, 272, 282, 291, 296, 299, 304, 308, 309, 312, 319, 353, 357, 368, 370, 395, 397, 405–407, 418
Hua, 107, 282
Huai, 182
Huang, 17–19, 77, 83, 97, 98, 126, 128–130, 132, 138, 145, 153, 156, 158, 161, 164, 167, 169, 173, 175, 182, 183, 185, 186, 188–192, 197–199, 201, 207–209, 212, 215, 224, 229, 230, 232–235, 237–241, 245–247, 251–253, 260, 264, 265, 267–270, 276, 286, 289, 291, 292, 294, 296, 302, 304, 305, 307, 309, 310, 315, 317, 318, 326, 330, 335, 336, 338, 345, 353, 361, 367, 369, 376, 384, 393, 398, 401, 406, 410, 413, 415, 424, 427
Huber, 351
Hudelot, 301
Hudson, 94, 239
Huff, 262
Hughes, 323

Hugues-Nuger, 405
Hui, 170, 178, 358, 383, 387
Hung, 80, 181, 191, 202
Hunter, 338
Huot, 244
Hupkes, 18, 72, 80, 313
Hur, 135, 203
Husmann, 313
Hussein, 115
Huth, 71
Hwang, 85, 131, 161, 180, 224, 232, 278, 371, 378, 384, 403, 404, 412, 430
Hyland, 200, 225
Hyslop, 258
Hämmerl, 317

Iacobacci, 143, 282
Iana, 352
Ibragimov, 321
Ichim, 154, 278
Idrissi-Yaghir, 428
Ignat, 20, 212
Ikbal, 234, 291
Ilagan, 425
Ilaslan, 135
Ilharco, 358
Ilievski, 277
Ilin, 156
Ilinykh, 212
Ilyas, 130, 276
Imai, 318
Imamura, 75
Imani, 331
Imhof, 173
Imouza, 328
Imperial, 202, 256
Imran, 112
Inan, 148
Inciarte, 110, 367
Indurthi, 20, 190, 317
Inoue, 111, 203, 258, 406
Inui, 382, 419, 420
Ionescu, 416
Iordache, 89, 231
Irie, 337, 364
Iro, 284
Isahagian, 133, 242, 287
Isbister, 223
Ishihata, 388
Ishii, 20, 94, 134
Islam, 119, 401
Islam Khondaker, 367
Ismail, 112
Ismayilzada, 168
Iter, 175, 195, 296, 324, 334
Ito, 76, 419
Ivanov, 321
Iwamoto, 416
Iwasawa, 198
Iyer, 78, 287, 330, 340, 348
Iyyer, 19, 77, 195, 212, 383
Izsak, 271

J u n c z y s - D o w m u n t, 98
Jaber, 112
Jacob, 372
Jacovi, 203, 259
Jafari, 83
Jahan, 161
Jaidka, 408
Jaimes, 252
Jain, 86, 104, 156, 197, 205, 226, 232, 326, 405, 423
Jakobi, 140
Jalota, 256
Jang, 129, 135, 251, 301, 320, 338, 392
Jansen, 195
Jarrar, 110–112
Jatowt, 197, 388
Jauhar, 169
Javier Vazquez Martinez, 80
Jawahar, 19
Jayakumar, 107
Jayanthi, 404
Jean, 342
Jeblick, 130
Jenkins, 242
Jeon, 233, 258, 268, 313, 349
Jeong, 108, 224, 385, 412, 425
Jeoung, 322
Jha, 156, 164
Jhamtani, 18
Jheng, 359
Ji, 18, 126, 127, 134, 158, 172, 185, 218, 230, 232, 257, 268, 283, 324, 343, 356, 360, 363, 371, 375, 380, 386, 388, 392, 395, 401, 430
Jia, 19, 152, 183, 221, 239, 248, 253, 276, 312, 348, 350, 380
Jian, 133, 167, 347, 350, 379
Jiang, 17, 81, 97, 129, 130, 137, 138, 141, 163, 164, 167, 170, 176, 186, 201, 202, 212, 220, 223, 237, 240, 243, 250, 260, 263, 277, 285, 287, 292, 298, 300, 303, 308, 334, 354, 363,

369–371, 386, 393, 395, 401, 409, 421
Jianhe, 201
Jiao, 126, 245, 288, 324, 356, 395
Jiayang, 231
Jimenez, 237, 383
Jimeno Yepes, 75
Jin, 20, 77, 165, 166, 176, 188, 231, 235, 269, 282, 296, 304, 316, 333, 343, 356, 360, 364, 389, 400, 405, 406, 413, 421
Jindal, 151, 298
Jing, 155, 192, 199
Jingxuan, 97
Jionghao, 230
Jo, 162, 166, 290, 338, 357, 424
Johansson, 193
John, 196, 314
Johnson, 17, 18, 340, 368
Jon, 76
Jones, 179
Jonker, 280
Joo, 220, 320
Jorge, 218
Jose, 219
Joseph, 184
Joshi, 18, 241, 399, 405
Josifoski, 149, 365
Joty, 55, 137, 152, 177, 183, 188, 254, 329
Jouravel, 83
Jovanovic, 86
Ju, 243
Jukić, 260
Jullien, 220
Jumelet, 72, 94
Junaed, 118, 119
Junczys-Dowmunt, 86
Juneja, 327
Jung, 147, 161, 180, 181, 220, 232, 236, 238, 251, 290, 295, 363, 376, 425
Jurafsky, 104, 201, 213
Juraska, 78, 98, 99
Jurgens, 90, 390
Jyothi, 20, 226, 228
Jäger, 140, 261

K, 196
Köhn, 85
Köllner, 83
Kabir, 119
Kabra, 240, 298
Kadaoui, 110
Kadlčík, 272
Kaestner, 205
Kaffee, 336, 346
Kagita, 413
Kahhoul, 113
Kahira, 284
Kahn, 169, 374
Kaiser-Schatzlein, 405
Kaji, 394
Kajiwara, 175
Kaknes, 395
Kalbassi, 78, 219, 220
Kalita, 381
Kalkar, 76
Kalo, 284
Kalra, 94
Kalyan, 82, 106, 226, 237, 286
Kamal, 119
Kamali, 80
Kamani, 217
Kamath, 314, 322
Kamel, 112
Kamigaito, 19
Kamoda, 382
Kamoi, 345
Kamp, 320
Kamper, 20
Kampfmeier, 279
Kamranian, 71
Kan, 144, 231, 247, 289, 328, 351, 365, 396
Kanayama, 416
Kanclerz, 142
Kandoi, 390
Kandukuri, 104
Kane, 184
Kaneda, 93
Kaneko, 19, 370
Kanerva, 301
Kanezaki, 297
Kang, 18, 71, 72, 108, 135, 145, 165, 185, 201, 203, 224, 233, 240, 263, 265, 268, 274, 293, 332, 355, 357, 385, 406, 415
Kangas, 375
Kannen, 234, 428
Kanojia, 97, 98, 199, 388
Kanoulas, 358
Kantharuban, 240, 407
Kantor, 306, 406
Kao, 171
Kapadnis, 408
Kapanipathi, 291
Kapoor, 243
Kapur, 329

Kar, 250
Karadayi, 243
Karamolegkou, 171, 210
Karanam, 234, 291
Karande, 106
Karanowski, 142
Kargaran, 331
Kargupta, 395, 408
Karidi, 93, 286
Karim, 119
Karl, 80, 81
Karlinsky, 131
Karmakar, 407
Karmaker Santu, 242, 254, 269
Karpinska, 77
Kartchner, 185
Karypis, 352
Kasai, 265, 397
Kashyap, 85, 99
Kassem, 197
Kassner, 140, 321
Kaszefski-Yaschuk, 271
Kate, 287
Katinskaia, 179
Kato, 245
Katsis, 298
Katsurai, 297
Katz, 224, 289, 329, 340, 349, 406
Kaur, 147
Kaushik, 18
Kavehzadeh, 152
Kawabata, 219
Kawaguchi, 319
Kawano, 297
Kawasaki, 131
Kazakova, 174
Kazanas, 184
Kazemeini, 83
Kazienko, 142, 252
Kchaou, 114
Ke, 18, 197, 386
Keh, 328
Keith, 17, 83
Keizer, 17
Keleg, 110, 154
Kelleher, 99
Keller, 137, 293, 305, 315
Kembhavi, 18
Kennington, 94
Keren, 380
Kern, 336
Kertkeidkachorn, 203
Kervadec, 251, 300
Keung, 352
Keutzer, 267
Kew, 231
Keydar, 258
Khabsa, 142, 252
Khadilkar, 327
Khadivi, 77
Khairallah, 258
Khaled, 115
Khalifa, 171, 304, 341
Khalilia, 110–112
Khan, 113, 119, 204, 291, 325, 390, 392, 429
Khandelwal, 18, 133, 151, 234, 400, 411
Khanna, 225
Khanuja, 153
Khapra, 20, 422
Kharitonov, 243
Khashabi, 19
Khatib, 348
Khatravath, 398
Khatri, 113
Khatry, 429
Khattab, 354
Khayrallah, 75, 86
Kheir, 111, 412
Khenglawt, 97
Khered, 114
Khiem, 312
Kho, 332
Khondaker, 110, 302
Khoong, 150
Khorshidi, 130
KhudaBukhsh, 353
Khurfan, 113, 114
Ki, 166
Kiani, 392
Kieu, 393
Kil, 147
Kilavuz, 333
Killamsetty, 340
Kim, 18–20, 89, 93, 125, 129, 131, 136, 143, 147, 160, 161, 181, 185, 191, 194, 201, 203, 212, 222, 228–233, 236, 240, 250–252, 258, 259, 263, 270, 278, 292, 298, 301, 313, 317, 320, 332, 338, 343, 349, 358, 372, 376, 380, 384, 385, 389, 390, 400, 403, 404, 409, 417, 422, 424, 428, 429
Kimura, 19
King, 291
Kirk, 324
Kiselev, 321
Klaisoongnoen, 107

Klakow, 94, 160
Kleesiek, 429
Klein, 153, 210, 247, 326, 334, 391, 399, 425
Kleiner, 324
Kleinfeld, 405
Klejch, 20
Kletz, 82
Kloots, 94
Kniazhevsky, 306
Knowles, 98
Knuples, 71
Ko, 128, 143, 251, 263, 273, 292, 409
Kobayashi, 211, 420
Kobi, 108
Kobyzev, 272
Kocaman, 112
Kochedykov, 398
Kochkina, 20
Kochmar, 19, 256
Kochsiek, 412
Kocmi, 75, 97, 98
Kocon, 142, 252
Kodner, 80, 171
Koehler, 415
Koehn, 75, 78, 97, 126, 187
Koh, 181, 212, 220
Koivula, 106
Kojima, 198
Kokhlikyan, 86
Kokush, 98
Kola, 401
Koller, 222, 239
Kolter, 368
Komachi, 78, 408
Komarlu, 408
Komulainen, 301
Kondrak, 256
Koneru, 351
Kong, 18, 133, 135, 167, 186, 221, 252, 289, 295, 390, 428
Kongyoung, 292
Konstas, 148
Konz, 325
Koo, 222, 399
Koopman, 202, 261
Kopparthi, 185
Koptyra, 252
Koraş, 250
Kordjamshidi, 80, 164, 381
Korenčić, 335
Korhonen, 143, 147, 282, 319, 333, 391, 407, 410
Kornaev, 321
Kost, 325
Kotamraju, 207
Kotek, 399
Kothyari, 327
Kotikalapudi, 321
Koto, 394
Kottur, 236
Kotula, 346
Koulogeorge, 170
Koupaee, 217, 280
Kouwenhoven, 93, 393
Kovacs, 97
Kovalenko, 76
Kozlova, 157
Kraus, 129, 258
Kreuter, 336
Krishna, 212, 251, 377, 404, 417
Krishnamurthy, 340, 347
Krishnaswamy, 402
Kritharoula, 348
Kriz, 284
Krubiński, 111
Kruk, 146
Kruse, 86
Kryscinski, 177
Ku, 20, 161, 191, 266, 336
Kuang, 108, 127, 334, 370, 389
Kubis, 232
Kuchaiev, 162, 268
Kudo, 76, 228
Kuehl, 148, 262
Kuleshov, 141
Kulkarni, 110, 111, 240, 323, 405
Kulshreshtha, 404
Kumar, 17, 56, 86, 107, 136, 155, 216, 223, 226, 241, 248, 265, 293, 355, 390, 392, 395, 405, 422, 426
Kumarage, 383
Kumari, 80
Kummerfeld, 242
Kunchukuttan, 180, 285, 422
Kung, 269
Kuo, 208
Kupari, 301
Kuparinen, 112, 290
Kurata, 20, 226
Kuratov, 427
Kurdy, 110
Kurfali, 244
Kurfalı, 17
Kuribayashi, 211
Kurimo, 80
Kurita, 245, 420, 423

Kurohashi, 177, 240, 245, 326
Kurtz, 223
Kuts, 352
Kutuzov, 20
Kuwanto, 273
Kuznetsov, 223, 352
Kvapilíková, 99
Kvinge, 325
Kwak, 107, 108, 185, 208
Kwan, 209, 291
Kwiatkowski, 19, 151, 205, 361
Kwok, 376
Kwon, 110, 240, 424, 430
Käser, 209
Köksal, 333, 421

Lapshinova-Koltunski, 76
Le-Minh, 77
Laban, 131, 177, 206
Labatut, 154
LaCasse, 149
Laclau, 349
Ladhak, 281
Lahiri, 107
Lahoti, 321, 396
Lai, 98, 182, 189, 217, 229, 234, 266, 268, 338
Laili, 178
Laippala, 301
Laitonjam, 97
Lakhdar, 113
Lakshmanan, 19
Lakshminarayanan, 377
Lal, 217
Lalwani, 423
Lam, 179, 206, 219, 238, 304, 386
Lamm, 258
Lampinen, 175
Lamsiyah, 112
Lan, 178, 312, 364, 369, 372, 414
Lanchantin, 413
Landers, 220
Landes, 86
Lange, 148, 343, 362
Langedijk, 94
Langlais, 72
Lango, 362
Lao, 18
Laouar, 355
Laouirine, 115
Lapata, 232, 243, 244, 293, 330
Lapshinova-Koltunski, 17
Laradji, 161, 362, 411
Larionov, 98, 420
Larkin, 98
Larson, 83
Lash, 108
Laskar, 97, 119, 161, 390, 398
Lasri, 284
Lastmann Assaraf, 405
Lastras, 241
Lateckit, 281
Latrache, 114
Lattimer, 187
Lau, 20, 252, 282, 284, 331
Laugier, 270
Launay, 110
Laurençon, 243
Lauriola, 19, 286
Lauscher, 17, 179, 327
Lauw, 311
Lavania, 277
LaViolette, 355
Lawan, 218, 284
Lawonn, 130
Lawrence, 80, 181, 400
Lawrie, 226
Le, 140, 175, 196, 213, 257, 301, 303, 315, 339, 388, 393, 406, 407, 429
Lea Heuser, 80
Lebron, 309
Lee, 17, 18, 72, 99, 107, 130, 131, 135, 143, 147, 155, 169, 175, 180, 181, 188, 202, 217, 221, 228, 230, 232, 235, 236, 240, 241, 247, 258, 259, 263, 268, 270, 272, 276, 278, 281, 287, 290, 299, 303, 304, 312, 313, 317, 327, 337, 339, 341, 349, 361, 364, 367, 371, 376, 382, 385, 389, 390, 403, 407, 409, 426, 430
Lee-Thorp, 308, 309
Lefever, 299
Legovini, 284
Legrand, 361
Lehman, 199
Lehmann, 160, 423
Lei, 18, 76, 77, 135, 208, 235, 269, 275, 308, 315, 328, 369, 406, 414, 418
Leidinger, 376, 411
Leippold, 129, 258
Leiter, 420
Lelkes, 299, 358
Lemon, 148
Lenci, 20
Lendvai, 83
Lensch, 147
Leonardelli, 82

Leong, 139, 177
Leonhardt, 430
Lerman, 390
Lernould, 166
Lertvittayakumjorn, 18
Leslie, 411
Lestari, 240
Leteno, 349
Levi, 399
Levkin, 280
Levy, 156, 215, 249, 259, 277, 284, 314, 330, 335
Lewenberg, 134
Lewis, 205, 212, 227, 323, 333, 374, 414
León-Villagrá, 271
Li, 17, 18, 20, 76, 77, 90, 98, 99, 104, 107, 108, 113–115, 119, 125, 126, 128, 130, 131, 133, 135, 136, 138–140, 144, 145, 148, 152, 156, 158–160, 163–165, 167, 168, 170–173, 175–179, 181, 184–187, 189, 191–194, 197, 198, 202, 204, 207, 208, 212–215, 217, 218, 221–223, 225, 229, 230, 234, 237–239, 242, 243, 246, 248–250, 253, 254, 256, 257, 262–265, 267, 268, 270, 272, 274–278, 280, 282, 283, 285, 289, 293–295, 297–299, 303–305, 307–314, 319, 321, 323, 325, 326, 328, 331, 334–336, 338–341, 345–347, 350, 351, 354–357, 360, 365–367, 369, 371, 372, 377, 378, 380, 381, 386, 389–392, 394, 395, 397, 400, 401, 403, 405, 406, 409, 414–418, 420, 422, 424, 428, 430
li, 165
Liakata, 128, 372
Lialin, 273
Lian, 168, 200
Liang, 18, 93, 113–115, 119, 142, 148, 162, 176–178, 189, 192, 226, 239, 245, 252, 266, 281, 283, 290, 304, 305, 321, 341, 362, 389, 397, 406
Liao, 127, 138, 140, 159, 176, 191, 199, 229, 269
Libovický, 19, 298
Lichouri, 114
Licht, 220, 417
Liem, 395
Lignos, 19, 86
Lillemark, 267
Lim, 135, 201, 203, 221, 222, 251, 311, 312, 337, 395
Limkonchotiwat, 80, 85, 86, 160, 281
Lin, 17, 88, 130, 133, 138, 139, 152, 159, 165, 171, 173, 176, 180, 187, 201, 205, 207, 214, 232, 236, 239, 241, 244, 247–250, 252, 260, 264, 268, 291, 299, 304, 309, 311, 316, 328, 335, 353, 355, 358, 359, 363, 371, 372, 377, 381, 394, 400, 402, 403, 406, 419, 422, 423, 429
Lindahl, 223
Ling, 128, 202, 253
Linzen, 207, 222, 288
Lipkin, 277
Lipton, 368, 417
Liqreina, 110
List, 83
Litschko, 270, 311
Litvak, 20
Liu, 17–20, 55, 56, 71, 75, 97–99, 107, 125–128, 133–136, 138, 140, 143–145, 149–151, 153–160, 162, 164, 165, 167, 168, 170, 174, 176, 178, 182, 183, 186, 188–190, 192, 194–200, 202, 204, 208, 211, 214, 219, 222, 224, 225, 228–230, 232, 235, 237–241, 243, 246–249, 253, 254, 257, 260, 264, 265, 267–270, 272, 278, 282, 283, 285–287, 289, 291–298, 304, 305, 307–312, 314, 318, 319, 321, 324, 326, 328, 329, 333, 334, 336, 337, 341, 342, 345, 348, 350–355, 357, 359, 361–363, 365, 368, 369, 372, 374, 377, 380, 383, 386, 389, 391, 393, 395–398, 400, 402, 404, 406, 409, 415, 416, 418, 419, 424–428
Liusie, 263
Livescu, 77, 158
Liyanage, 89
Lo, 97, 98, 148, 247, 262, 324, 384, 387
Loáiciga, 94
Loakman, 205
Lodha, 80, 342
Loftsson, 351
Logeswaran, 155, 228, 304, 341
Lohiya, 185
Lohr, 174
Long, 18, 152, 198, 279, 360, 382
Longpre, 20
Longtin, 108
Loo, 412
Lopez, 77, 311
Lopez-Avila, 275

Loreaux, 299
Lothritz, 210
Lotz, 192
Lou, 148, 225, 239, 264, 277, 352, 405
Loukas, 85
Lounnas, 114
Lovenia, 104, 134
Lowd, 414
Loweimi, 20
Lowphansirikul, 85, 281
Loyola, 375
Loáiciga, 17
Lu, 18, 113, 125, 129, 132, 136, 141, 144, 153, 164, 175, 203, 204, 248, 250, 263, 266, 272, 282, 298, 306, 312, 351, 363, 389, 390, 398, 401, 405, 421, 425
Luan, 143, 181
Lubis, 130
Lucas, 187, 259, 288, 303, 375
Lucchese, 106
Luccioni, 384
Lukasiewicz, 301, 392
Lungren, 225
Luo, 107, 126, 155, 159, 162, 174, 175, 192, 228, 234, 248, 293, 296, 298, 306, 319, 337, 341, 365, 377, 387, 395, 402, 406, 424
Luoma, 301
Lusoli, 413
Luss, 291
Lutati, 260
Luu, 108, 310, 334, 356, 388, 424
Luukkonen, 301
Lv, 158, 216, 230, 253, 269, 278, 307, 309, 369
Lymperaiou, 348
Lyu, 75, 98, 108, 173, 212, 228, 268, 283, 289, 357

M, 285
M a r r e s e - T a y l o r, 78
M i n h - C o n g, 77
M S, 402
Möller, 76, 97
Müller, 75
Ma, 17, 71, 75, 77, 94, 125, 126, 132, 141, 153, 154, 160, 170, 175, 180, 188, 194, 196, 197, 214, 218, 229, 244, 257, 259, 262, 266, 277, 290, 314, 321, 334, 336, 345, 353, 359, 363, 365, 366, 377, 380, 383, 397, 398, 401, 403, 415, 418
Mabuya, 284
MacDonald, 292
Macina, 329
Macketanz, 76, 97
Macko, 303
Madaan, 157, 261, 331
Madabushi, 202
Madaio, 17
Madanagopal, 195
Maddela, 274
Madhani, 422
Madnani, 85
Madotto, 236, 381
Madureira, 73, 269
Madusanka, 221
Maeda, 139, 423
Maekawa, 318
Magdy, 110, 154
Mager, 19
Maggini, 110
Mahamud, 401
Maharaj, 426
Maharana, 208
Mahari, 108, 426
Mahbub, 390
Mahdaouy, 112
Maheshwari, 315
Mahmoud, 112, 197
Mahmoudi, 115
Mahmud, 72
Mahor, 227
Mahowald, 149, 184
Mai, 393, 419
Maillard, 246, 417
Maimon, 167, 207
Maina, 90
Maini, 368
Maiti, 20
Maity, 156
Majmudar, 398
Majumdar, 119
Majumder, 17, 297, 327, 378
Makhervaks, 355
Makondo, 291
Maladry, 205
Malakasiotis, 85, 410
Malandrakis, 162
Malartic, 110
Malaysha, 110
Malhas, 111
Malin, 248
Mallick, 383
Malmasi, 250, 402
Malmsten, 223

Maltseva, 352
Mamié, 108
Mamidi, 72
Mamta, 280
Manakhimova, 76, 97
Manakul, 263
Manchanda, 327
Mandal, 99
Mangalvedhekar, 113, 114
Manggala, 411
Mangla, 167
Manh, 86
Manica, 157
Manjunatha, 195
Manning, 104, 185, 197, 244, 277, 368, 375
Manocha, 136, 142, 355, 380
Manolache, 221
Manolescu, 158
Manrique, 113
Mansimov, 134
Mansour, 111, 377
Mantri, 252
Mao, 125, 142, 178, 179, 186, 203, 245, 252, 257, 260, 265, 312, 319, 326, 332, 338, 352, 354, 362, 371, 373, 391, 400, 404, 416
Maratha, 119
Margatina, 150
Marinier, 243
Marivate, 20
Markosyan, 86
Markovitch, 211
Markowska, 253
Marrese-Taylor, 375
Marshall, 146
Martin, 140, 395
Martindale, 127
Martinek, 272
Martinez, 95, 112
Martins, 98, 100, 157, 191, 200, 427
Marttinen, 156
Martínez del Rincón, 301
Masip Gomez, 400
Masry, 152
Mastrapas, 85
Mathur, 86, 97, 207, 238, 380, 383
Matoshi, 107, 157
Matsubara, 86
Matsumoto, 186
Matsuo, 78, 198
Matsuzaki, 76, 300
Matthews, 113
Maurya, 408
Mavi, 107
May, 188, 236, 368, 389
Mayeesha, 382
Maynard, 413
Maynez, 244, 271, 283
Mazzotta, 110
Mbataku, 281
Mbonu, 284
McAfee, 268
McAuley, 132, 133, 272, 327
Mcbride, 108
McCallum, 287, 383
McCarthy, 422
McCrae, 430
McDanel, 17
McGovern, 95
McInerney, 417
McKenna, 340
McKenzie, 355
McKeown, 281, 324, 422
McKinney, 130
McMains, 94
Md Salleh, 402
Meade, 333
Meadows, 374
Medvedeva, 108
Mehandru, 150
Mehdad, 377
Mehler, 430
Mehra, 323
Mehri, 229
Mehta, 349
Mei, 153, 330, 349
Meister, 149, 154, 156
Meka, 264
Mekala, 207, 222
Mel, 112
Mešek, 391
MELLAH, 112
Melo, 383
Mendelsohn, 90
Mendelson, 131
Menezes, 341, 388
Meng, 18, 125, 154, 176, 254, 298, 313, 332, 358, 369, 378
Menini, 82
Menon, 119, 316, 343, 391
Merello, 140
Merhav-Fine, 379
Merioksa, 301
Merler, 396
Merlo, 81, 298, 335
Merullo, 323

Messelle, 218
Metallinou, 162
Metze, 20
Metzler, 196, 349, 358, 383
Meunier, 295
Meuschke, 130
Meza Ruiz, 20
Mi, 20, 217, 279, 282, 290–292, 311
Miandoab, 415
Miao, 179, 181, 230, 330, 356, 399
Miaomiao, 98, 99
Michael, 239
Michaelov, 194, 340
Micher, 150
Mickus, 203
Mieskes, 17, 302
Miglani, 86
Mihalcea, 164, 183, 212, 253, 371, 419, 421
Mikhailov, 157
Milbauer, 396
Miletic, 242
Miletić, 112, 290, 345
Milios, 81
Miller, 86
Milliken, 85
Milunovic, 401
Mimno, 141, 195, 323
Min, 19, 76, 94, 107, 206, 212, 347, 419
Mina, 110
Minakova, 271
Minervini, 18, 94, 107
Mingsheng, 209
Minhas, 276
Minixhofer, 136
Minkova, 106
Minn, 413
Mirbostani, 217
Mircea, 337
Mireshghallah, 93, 273
Miret, 287, 342
Mirroshandel, 217, 253
Mirzaee, 164
Mirzaei, 90
Mirzazadeh, 98
Mishra, 86, 216, 265, 411, 426
Mita, 408
Mitchell, 185, 197, 368
Mitra, 293, 320
Mittal, 192, 226
Miwa, 76
Miyanishi, 423
Miyano, 175
Miyao, 94, 285
Miyashita, 129
Mizokuchi, 406
Mizrachi, 405
Miłkowski, 142
Mo, 333, 416
Mobasher, 90
Modi, 18, 133, 277
Moeller, 320
Moens, 172
Moghe, 97
Mohamed, 110, 243
Mohammad, 127, 150, 170, 218, 384
Mohammed, 104, 118, 119, 400
Mohan, 401
Mohebbi, 392
Mohseni, 77
Moisio, 80
Mok, 270, 349
Molaei, 99
Molchanov, 76
Molloy, 97, 98
Molnar, 72
Mom, 252
Monath, 421
Mondal, 80, 133, 274
Moniz, 405
Monsur, 119
Montariol, 168, 270
Monz, 76, 159, 182, 183, 411
Moon, 135, 221, 222, 381, 389
Mooney, 217
Moor, 208
Moortgat, 72
Moosavi, 20, 241
Moraffah, 383
Morante, 20
Morency, 316, 422
Morger, 223
Moriceau, 295
Morioka, 129
Morishita, 76
Moro, 303, 391
Morris, 141, 358
Morrison, 72, 108
Morshedzadeh, 115
Morstatter, 364
Mortensen, 77, 82, 240, 265, 386
Moryossef, 250
Mosbach, 94
Moschitti, 286
Moses, 112
Moshtaghi, 19
Moskvichev, 419

Moslem, 99
Mou, 19, 266, 341
Moubayed, 94
Mousi, 211
Mout, 113
Mu, 194, 339
Mubarak, 111
Mueller, 72
Muennighoff, 301
Muhammad, 218, 284
Muhtaseem, 119
Mukherjee, 76, 98, 119, 142, 205, 226, 227, 305, 318, 428
Mukku, 401
Mullappilly, 113, 204
Muller, 19, 78, 151, 276, 382
Mullov, 351
Mulyar, 86
Mun, 411
Munawar, 299
Mundnich, 397
Muntasir, 119
Muppidi, 403
Murahari, 17, 237, 286, 383
Murakhovs'ka, 131, 206
Muraoka, 416
Murawaki, 19
Murayshid, 110
Muresan, 272, 357, 360, 412
Murphy, 97, 98
Murray, 158, 187, 246
Murthy, 265, 412
Murty, 277, 423
Murukannaiah, 280
Murumkar, 113, 114
Muthusamy, 287
Mwase, 284
Myers, 134
Möller, 285
Müller, 250
Müller-Eberstein, 270, 334

N, 72, 227
Névéol, 75
Na, 169, 293, 384, 393
Nachum, 307, 330
Nag, 142
Nagasawa, 228
Nagata, 19, 131
Nagoudi, 110, 111, 113, 302, 334
Naik, 148, 199, 377
Nair, 408, 413, 424
Najork, 349, 399
Nakashole, 343
Nakhlé, 99
Nakhle, 77
Nakhost, 317, 420
Nakov, 111, 144, 183, 192, 247, 294, 332, 351
Nam, 339
Namazi-Rad, 137, 264
Namazifar, 229
Nambi, 326
Namburi, 252
Namdar, 77
Namdarzadeh, 77
Nan, 311, 402
Nanayakkara, 428
Nandi, 119, 293
Nanduri, 141
Nandwani, 241
Nandy, 408
Nanniyur, 227
Naradowsky, 285
Narang, 330
Narasimhan, 106, 237, 286, 383
Narayan, 20, 283
Narayanan, 280
Naseem, 17, 93, 113–115, 119, 299
Nasim, 119
Naskar, 98, 99
Nasr, 113
Nastase, 81, 298
Naszadi, 411
Natarajan, 343
Nath, 327
Nathan, 395
Nathani, 187
Natouf, 94
Naumann, 200
Naushan, 284
Navarro, 98
Navigli, 330
Nawrot, 85
Nayak, 80, 293
Nayouf, 110
NC, 422
Neel, 267
Neelam, 234, 291
Negia, 89
Negreanu, 257
Negri, 76, 97, 127, 274
Nejadgholi, 18
Nejdl, 297
Nema, 181
Nematzadeh, 145, 407
Nenkova, 259

Neshaei, 209
Neubig, 77, 100, 127, 153, 157, 166, 204, 237, 368, 396
Neves, 75, 382
Newman, 262, 324
Ney, 77
Ng, 17, 104, 129, 235, 246, 370, 372, 378, 424
Nghiem, 86, 339
Ngo, 217, 271
Nguyen, 17, 18, 86, 107, 134, 144, 213, 217–219, 228, 249, 271, 312, 332, 339, 342, 351, 393, 406, 409, 424, 428
Ni, 129, 140, 219
Nie, 18, 94, 119, 138, 140, 260, 305, 333, 380
Niehues, 351
Nieminen, 99
Nigam, 108
Nigatu, 89, 381
Nikandrou, 148
Niklaus, 107, 108, 157
Nikolaev, 258, 271, 286, 320, 364
Nikolenko, 361
Nikolentzos, 18
Ning, 18, 242, 243
Ningthoujam, 99
Nishida, 75, 139
Nissim, 226
Nitisaroj, 236
Niu, 190, 238, 268
Njoo, 56, 376
Noble, 212
Noh, 251
Nokku, 107
Nong, 141
Nori, 200, 225
Norlund, 193
North, 119, 395
Noune, 110
Nov, 302
Novikova, 80
Nowak, 198
Nozza, 17, 327
Nunna, 199
Nussbaum, 86
Nutanong, 281
Nwatu, 212
Nwesri, 114

O'Connell, 208
O'Donoghue, 318
O'Neil, 286
O'Regan, 220
Oba, 72
Ochieng, 326
Ofitserov, 399
Oflazer, 240
Ogayo, 77
Ogier, 182
Ogueji, 89
Ogundepo, 152, 173, 284
Ogunremi, 104
Oguz, 233, 285
Oh, 17, 180, 217, 241, 335, 402, 426
Ohashi, 241
Ohko, 416
Ojha, 119
OJO, 113
Ojo, 89, 113
Oka, 248
Okabe, 336
Okazaki, 144, 370, 423
Oktay, 200, 225
Okumura, 346
Oladipo, 152, 173, 284
Olariu, 210
Olatunji, 281
Olausson, 277
Olivier, 18
Olsen, 112
Olteanu, 295
Omar, 111
Omirah, 110
Ong, 86, 202, 240, 246
Ontanon, 308
Onwuegbuzia, 284
Opedal, 278
Opoku, 218, 284
Opper, 72, 227
Orasan, 20, 98, 254
Orbach, 406
Orlikowski, 194
Ormazabal, 213
Ormerod, 301
Orwig, 292
Osei, 218, 281, 284
Oseki, 72
Oshika, 312
Oshin, 119
Osmelak, 302
Ostapenko, 153
Ostermann, 233
Osuchukwu, 281
Otani, 131
Otao, 356
Otiende, 284

Otto, 407
Ou, 309
Oumer, 113
Ounis, 292
Ousidhoum, 218, 376
Ouyang, 126, 153, 270, 278, 298, 324, 338, 395
Oved, 305
Overbay, 231
Overwijk, 358, 374
Owan, 72
Owodunni, 152, 284
Oyama, 211, 310
Oyamada, 330
Ozler, 383

Paccosi, 82
Pacheco, 106
Packer, 321
Padfield, 347
Padia, 18
Padmakumar, 56, 148
Padró, 341
Padua, 318
Padó, 320, 364
Pahari, 105
Pahwa, 227
Pai, 107, 256
Painkra, 279
Pakala, 227
Pakray, 97
Pal, 80, 94, 97, 99
pal, 73
Palaskar, 20
Palmer, 79, 185, 218, 395
Palmero Aprosio, 106
Pan, 72, 138, 144, 152, 171, 187, 221, 247, 273, 279–281, 285, 303, 328, 330, 351, 357, 407, 418
Panagiotopoulos, 411
Panchenko, 98, 262, 352
Pandey, 428
Pandya, 413
Pang, 107, 126, 138, 151, 200, 219, 387, 417, 425, 427
Panigrahi, 328
Panigrahy, 264
Panov, 262
Pantazopoulos, 148
Paola Garcia Perera, 20
Paolini, 196
Papadimitriou, 197, 201
Papakostas, 411
Papangelis, 17
Paperno, 203
Papi, 127
Papotti, 335
Pappas, 134, 342
Paquette, 296
Paradise, 423
Parapar, 373
Parashar, 133
Parde, 19
Parece, 412
Parekh, 148
Parikh, 18, 271
Parisien, 396
Park, 17, 20, 72, 99, 147, 188, 207, 221, 222, 224, 230, 231, 233, 235, 247, 250, 261, 268, 270, 313, 317, 349, 357, 365, 376, 383, 389, 395, 404, 409, 412, 417, 425
Parović, 147
Parra, 89
Parthan, 422
Parthasarathi, 72, 393
Parvez, 343
Passonneau, 175
Pastor, 82
Pasunuru, 168
Pasupat, 419
Patange, 401
Patel, 18, 93, 213, 345, 407
Pathak, 267
Patil, 208, 378, 428
Patras, 236
Patti, 148
Pattichis, 149
Patton, 324
Paturi, 238
Patwa, 104
Patwardhan, 113, 114, 400
Paul, 17, 168, 173, 227, 236, 270
Paulheim, 352
Pauli, 313
Pavlick, 151, 323, 412
Pavlov, 352
Pavlova, 110
Payne, 171
Pecina, 111
Pei, 139, 151, 243, 329, 382, 390, 397
Peitz, 19
Pellegrain, 355
Pelrine, 328, 424
Penedo, 110

Peng, 56, 89, 130, 152, 156, 158, 162, 166, 186, 197, 199, 212, 221, 230, 249, 252, 262, 269, 303, 325, 334, 339, 351, 366, 368, 369, 375, 385, 389, 403, 404, 406, 414, 421, 428
Pengpun, 80
Pennisi, 106
Pentland, 108, 426
Pereira, 19
Perez-Rosas, 419
Peris, 398
Perlitz, 303
Perry, 405
Pershin, 321
Pertseva, 238
Pesaran zadeh, 231
Peskine, 335
Peskoff, 18, 414
Peter, 78
Petersen, 242
Pethe, 136
Petit, 243, 350
Petrak, 241
Petrakov, 262
Petricek, 399
Petrick, 77
Petrov, 196, 258, 342
Petrushkov, 77
Petryk, 272
Petty, 301
Peyrard, 149, 365
Pezzelle, 148, 149, 322
Pfeiffer, 17, 130, 136, 173
Pfister, 164, 317, 420
Pham, 193, 312, 317, 351
Phan, 104, 107, 213, 283
Phang, 348
Phatthiyaphaibun, 85
Phiri, 284
Phung, 184, 424
Pi, 221
Pial, 136
Piano, 72
Piantanida, 184, 355
Piao, 98
Piazza, 157
Piccardi, 175, 350
Piergentili, 274
Pietquin, 243
Piktus, 173, 301
Pikuliak, 303, 391
Pilehvar, 20, 415
Pimentel, 18, 149, 154, 156, 215, 278
Ping, 268
Pintard, 274
Pinter, 20, 145, 205
Pinto, 325
Piontkovskaya, 361
Piraviperumal, 405
Pires, 100
Pirhadi, 90
Piskorski, 275
Pivovarova, 20
Plank, 146, 154, 270, 300, 334
Plaza del Arco, 327
Plekhanov, 140
Ploeger, 297
Plummer, 219
Podolskiy, 361
Podroužek, 391
Polpanumas, 85
Pombal, 98, 427
Ponti, 17, 147, 319
Ponwitayarat, 281
Ponzetto, 311
Poon, 165, 200, 225, 430
Poore, 186
Popa, 298
Popel, 76
Popescu, 221
Popovic, 19
Porada, 72
Poria, 312
Pospíšil, 111
Post, 77, 85, 86, 98, 126
Potamianos, 162
Poth, 173
Potluri, 321
Potthast, 173, 340, 348
Potti, 260
Potts, 185, 244, 331, 354, 403
Pougué-Biyong, 284
Pourreza, 368
Pouw, 94
Pozzobon, 227, 333
Prabhu, 228
Pradeep, 358
Pramanick, 127, 293
Prange, 233
Prasad, 72, 273
Prasse, 140, 261
Pratapa, 325, 386
Prato, 393
Pratt-Hartmann, 221
Preotiuc-Pietro, 19, 323
Presani, 170

Pressimone, 180
Preum, 300
Primadhanty, 251
Priya, 216
Prokhorov, 227
Prokic, 83
Prud'hommeaux, 19
Pruthi, 368
Pryzant, 175, 324
Przewozny-Desriaux, 345
Pu, 130, 132, 216, 372, 375
Pudota, 107
Puduppully, 180, 285
Pujara, 18, 169, 343, 364, 389
Puranik, 297
Purkayastha, 173, 179
Purver, 128
Purwarianti, 240
Pyatkin, 287, 384
Pyysalo, 301
Pérez, 373
Pérez-García, 200, 225

Qachfar, 111, 112
Qadar, 99
Qaddoumi, 114
Qasemi, 424
Qazi, 429
Qi, 79, 80, 126, 134, 147, 178, 188, 198, 268, 286, 304, 311, 347, 362, 394, 414
Qian, 130, 138, 158, 178, 179, 232, 353, 359, 389, 412, 416
Qiang, 270
Qiao, 98, 128, 148, 289, 293, 400
Qin, 110, 138, 158, 188, 194, 200, 214, 222, 231–233, 239, 253, 269, 278, 282, 302, 309, 317, 355, 363, 383, 396
Qiu, 231, 237, 260, 263, 264, 266, 293, 307, 309, 319, 325, 359, 371, 394
Qorib, 370
Qu, 108, 182, 203, 206, 209, 233, 257, 293, 394, 405, 427
Quadrianto, 310
Quan, 17, 139, 224, 260, 264
Quattoni, 251
Quintero, 383
Quochi, 19

Rügamer, 94
Rabbany, 328, 338, 424
Rabby, 401
Rabus, 83
Radev, 254, 311, 329
Radevski, 181
Radhakrishnan, 392
Radharapu, 396
Radinsky, 355
Rafailov, 197
Raffel, 289, 365
Rafiei, 368
Rafsan, 118
Raganato, 20
Raghu, 241
Raghuveer, 181
Raheja, 405
Rahimi, 380, 383
Rahman, 118, 119, 161, 401
Rahmani, 160
Rahouti, 118, 119
Rai, 216
Raihan, 119
Raj, 142, 218, 274
Rajabi, 78
Rajkomar, 299
Rajmohan, 400
Rajpurkar, 208, 225
Rajpurohit, 106, 226, 237, 286
Rakshit, 90
Ramakrishna, 423
Ramakrishnan, 340
Raman, 326, 368
Ramanathan, 184
Ramaneswaran, 136
Ramasamy, 312
Rambow, 253, 360
Ramesh, 326
Ramnath, 363
Ramponi, 157
Ramprasad, 417
Rana, 401, 405
Ranaldi, 18
Ranasinghe, 98, 119, 353, 388
Ranathunga, 89
Ranchordás, 320
Rangreji, 414
Rangwala, 352
Rani, 107, 157, 227, 430
Ranjbar Alvar, 71
Rank, 330
Rao, 77, 196, 287, 349, 411
Rashkin, 258
Rasiah, 107
Rastogi, 241
Ratan, 86
Rathore, 149
Rauf, 76

Raunak, 98, 341, 388
Ravfogel, 198, 284
Ravi, 378
Ravichander, 18, 363, 424
Rawte, 267
Ray, 199
Ray Choudhury, 94
Raz, 186
Raza, 429
Razniewski, 17, 284
Razumenko, 108
Razumovskaia, 143
Razzhigaev, 352
Rebedea, 396
Reddy, 81, 213, 225, 227, 287, 333, 350, 380, 399
Regev, 215
Rei, 19, 97, 98, 424, 427
Reich, 140, 261
Reichart, 18, 305
Reichel, 83
Reichstein, 130
Reif, 81, 195
Reilly, 243
Reina, 184
Reinanda, 390
Reis, 208
Reitter, 258
Rekabsaz, 370
Rekathati, 223
Rekesh, 105
Reksoprodjo, 328
Remy, 331
Ren, 55, 56, 99, 108, 132, 140, 182, 204, 206, 214, 243, 248, 263, 266, 267, 283, 298, 312, 332, 348, 357, 363, 373, 377, 380, 390
Renduchintala, 340
Renje, 83
Rennard, 289, 338
Renner, 202
Resnicow, 419
Resnik, 367, 413
Restaino, 18
Retkowski, 351
Reusens, 235, 302
Revi, 114
Reymond, 80
Rezaee, 382
Rezaei, 383
Rezagholizadeh, 72, 272
Rho, 17
Ribeiro, 97, 296, 311
Richardson, 321
Ricoul, 243
Ridley, 215
Riedl, 89
Rieser, 148, 308
Rietsche, 209
Rifat, 119
Rigau, 306, 337
Rigutini, 344, 403
Rijhwani, 271
Rikters, 76
Riley, 100
Rinott, 413
Risch, 351
Risini, 154
Ritter, 18, 143
Rivera, 284
Rivera-Soto, 346
Rizk, 242, 287
Ro, 417
Robinson, 77, 82, 396
Roca, 350
Rocholl, 347
Rodriguez, 345
Rodriques, 318
Roegiest, 108
Rogers, 318
Roh, 263
Rohatgi, 231, 259
Rokach, 108
Rokhlenko, 250, 402
Roll, 93
Roller, 75
Romani, 99
Romanou, 270
Romberg, 194
Romeo, 19, 134
Ronanki, 397, 404
Rong, 187
Rony, 398
Ropers, 78, 219, 220, 417
Rosario, 89
Rosati, 411
Rose, 153, 199, 377
Rosenfeld, 200
Rosenthal, 174
Ross, 79, 86, 134, 311
Rosset, 19, 189, 358
Rossi, 217, 259, 384
Rosso, 89, 113, 311, 335
Rostami, 17, 163
Rostamkhani, 89
Rosá, 20

Roth, 72, 196, 237, 274, 276, 280, 314, 369
Rottger, 324
Roukos, 299, 354
Roussinov, 430
Rovatsos, 107, 108
Rovera, 106
Roy, 119, 134, 205, 222, 382
Rozanov, 241
Rozovskaya, 19
Ru, 231
Ruas, 170, 235
Rubinstein, 358
Rubungo, 284
Ruder, 151, 173, 218, 244, 284
Rudinger, 365
Rudman, 318
Rudra, 108
Rudzicz, 323
Rueda, 86
Ruiz-Dolz, 174
Rumshisky, 273
Ruosch, 354
Ruppenhofer, 279
Rush, 141, 142, 242, 352, 358, 429
Russo, 271, 281
Rust, 192, 357
Rustogi, 205
Rutunda, 218
Ryabov, 352
Rybinski, 291
Rychly, 76, 99
Ryu, 72, 107
Rücker, 224
Rücklé, 17
Ré, 374

S, 216
S a n d o v a l - C a s t a n e d a, 77
Saad, 17, 110, 197
Saad-Falcon, 354
Saadany, 254
Saakyan, 272, 357, 412
Sabeeh, 113, 114
Sabharwal, 18, 220, 321, 324
Sabir, 341
Sablé-Meyer, 216
Sabour, 183
Sabty, 112
Sachan, 18, 139, 261, 265, 313, 321, 329, 332, 350, 421
Sadat, 343
Sadler, 269
Saeidi, 373
Saenko, 219
Saenz, 208
Safdari, 330
Saffari, 160
Sagae, 20
Sagirova, 364
Sagot, 76, 219, 243
Saha, 17, 20, 118, 119, 156, 183, 197, 205, 245, 273, 373, 423
Sahak, 323
Sahlgren, 223
Sahoo, 80, 398
Sahu, 362, 411
Saini, 80
Sainz, 292
Sairanen, 301
Saito, 139, 252
Sajjad, 19
Sakaguchi, 17, 382
Sakai, 75
Sakib, 119
Sakota, 149
Sakshi, 136
Sakti, 240
Sala, 252
SalahEldin, 112
Salcido, 369
Saleh, 111, 377
Salehi, 150
Saleki, 118, 119
Salesky, 126, 192
Salman, 183
Samaradivakara, 428
Samaraweera, 89
Samardžić, 244
Samavedhi, 207
Samih, 19
Samir, 171
Samo, 81, 298
Samory, 361
Sampath, 166
Samuel, 94
San Martin, 284
Sanasam, 293
Sanayai Meetei, 99
Sancheti, 365
Sanchez, 174
Sandoval, 314
Sang, 130
Sanghai, 308, 309
Sanjabi, 138
Sankar, 236
Sankaran, 206

Sankarasubbu, 73
Sano, 236
Sansom, 334
Santhanam, 354
Santra, 331
Santy, 55
Saon, 226
Sap, 129, 141, 184, 259, 261, 411
Saparov, 107, 195
Saphra, 18
Saraf, 381
Sarasua, 354
Sarawagi, 226, 234, 327
Sarhan, 115
Sarkar, 19, 173, 227, 242, 267, 269, 293, 367
Sarker, 119
Sarma, 99
Sarthak, 398
Sartran, 218
Sasano, 312
Sasikumar, 428
Sathe, 234
Satoh, 132
Savoldi, 76, 274
Saxena, 179, 305, 408, 426
Saxon, 153
Sayeed, 82, 94
Saynova, 193
Scao, 301
Scardigli, 280
Scarton, 328, 339, 363
Schütze, 55, 56
Scharlau, 410
Scheffer, 261
Scheffler, 17
Scheible, 170
Scherp, 80, 81
Scherrer, 19, 83, 112, 290
Schick, 421
Schiff, 319
Schimanski, 129, 258
Schirmer, 284
Schlangen, 73, 269
Schlechtweg, 83
Schlegel, 98, 287
Schlichtkrull, 19, 203, 222, 376
Schloss, 322
Schlötterer, 250
Schmid, 380
Schmidhuber, 337, 364
Schmidt, 78, 86
Schneider, 72, 85, 398
Schockaert, 223, 345, 378
Schommer, 112
Schott, 319
Schottmann, 100
Schubert, 184
Schuetze, 240, 321, 329, 331, 333, 362, 380, 421
Schulhoff, 325, 414
Schulte im Walde, 71, 242
Schuster, 17, 299
Schwabe, 348
Schwaighofer, 225
Schwandt, 83
Schwartz, 17, 81
Schweter, 72
Schöffel, 256
Schölkopf, 421
Sclar, 259, 387
Seah, 180
Sedoc, 19, 206, 216, 314
Sedova, 369
Seegmiller, 300
Segal, 326
Seibold, 428
Seifert, 250
Sellam, 271
Sellat, 111
Selvaraj, 322
Sembium, 402
Semenov, 97
Semnani, 206, 238
Sen, 361, 373
Sen Sharma, 119
Sengamedu, 163
Sengupta, 342, 388, 410
Sennrich, 19, 99, 100, 220
Sensharma, 227
Seo, 166, 185, 221, 222, 320, 337, 338, 425, 430
Serra, 286
Seth, 262
Setiawan, 100
Setzu, 407
Sewak, 169
Shachar, 405
Shafran, 241, 383
Shah, 133, 225, 236, 347, 397
Shahaf, 205, 288, 372
Shahmohammadi, 147
Shahriar, 390
Shaikh, 17, 227
Shaitarova, 108
Shaker, 113, 204
Shakeri, 196

Shakhmatov, 352
Shakhnarovich, 77
Shakya, 383
Shalyminov, 377
Shan, 375
Shang, 76, 77, 162, 177, 207, 214, 289, 292, 338, 364, 366, 376, 386, 419, 423
Shani, 205, 372
Shankarampeta, 378
Shao, 126, 134, 155, 201, 219, 263, 295, 321, 374
Shapira, 307, 319, 335
Sharaf, 341
Shardlow, 231
Shareghi, 20, 72, 130
Sharifi, 243
Sharma, 108, 180, 197, 210, 225, 226, 234, 249, 277, 291, 422, 423
Sharoff, 430
Shatabda, 119
Shavarani, 173
Shavrina, 157
She, 182
Shea, 223
Shefer, 319
Sheffield, 184
Shehadi, 302
Shehata, 183
Shehu-Bello, 218
Sheinwald, 303
Shelmanov, 262
Shen, 17, 132, 141, 201, 219, 224, 229, 235, 246, 249, 260–262, 275, 304, 305, 307, 309, 312, 347, 351, 352, 363, 367, 372, 387, 395, 398, 401, 409, 417, 418, 427
shen, 201
Sheng, 267
Sherborne, 243
Sherman, 85
Sheshadri, 166, 274
Sheth, 226, 227, 267, 383
Shi, 17, 18, 77, 97, 98, 126, 136, 139, 140, 152, 155, 186, 188, 199, 204, 207, 220, 230, 235, 245, 246, 256, 257, 264, 266, 268, 282, 285, 289, 307, 333, 363, 365, 367, 375, 384, 394, 407, 417
Shim, 228, 247
Shimada, 105
Shimizu, 177, 228, 403
Shimodaira, 159, 211, 310
Shin, 145, 185, 188, 320
Shinbir, 114
Shindell, 99
Shinzato, 399
Shirai, 203
Shivagunde, 273
Shivasankaran, 251
Shlens, 219
Shliazhko, 157
Shmatikov, 141
Shmueli-Scheuer, 303
Shode, 284
Shoeybi, 268
Shojaee, 217
Shomer, 366
Shon, 20
Shou, 135, 354
Shrivastava, 76, 98, 376
Shtedritski, 318
Shu, 17, 134, 253, 390
Shuai, 137, 374
Shui, 107, 425
Shukla, 114
Shum, 329
Shumailov, 316
Shutova, 141, 376, 411
Shvartzshanider, 108
Shwartz, 131, 168, 180, 378
Shypula, 417
Si, 304, 325, 346, 402, 407
Sia, 90
Sibaee, 113, 114
Sicilia, 17
Siddhant, 98, 271
Siewert, 83
Sifat, 119
Signoroni, 99
Sihler, 80, 81
Sikasote, 284
Sil, 174, 354
Silberer, 93
Siledar, 428
Sileo, 166
Silfverberg, 171
Silva, 97
Simhi, 211
Simko, 181, 303, 391
Simperl, 203, 378
Sinapov, 18
Sinclair, 72
Sindhujan, 98
Singer, 130
Singh, 86, 89, 99, 132, 133, 139, 147, 186, 218, 236, 257, 262, 274, 277, 279,

281, 299, 358, 373, 377, 381, 412, 428, 429
Singha, 429
Singhal, 106, 251, 404
Singla, 133, 149, 347, 400
Sinha, 329, 413, 420
Sinkala, 284
Siro, 284
Sirts, 19
Sitaram, 285, 326
Skala, 262
Skiena, 136
Skobov, 418
Skopek, 72
Skórzewski, 232
Slamu, 186
Slavkovski, 401
Slobodkin, 198, 307, 333
Slonim, 289, 303, 349, 406
Sloto, 75
Small, 19, 325
Smith, 72, 78, 112, 168, 170, 220, 239, 265, 287, 417, 426, 428
Smith-Renner, 252
Smola, 367
Smoleň, 391
Snajder, 260
Snell, 391
So, 196
Soares, 151, 205, 361
Soba, 107
Sodhani, 393
Sohn, 107, 155, 228, 272, 430
Sojitra, 197
Sokolov, 353
Solar-Lezama, 277
Soldaini, 262, 324, 384
Soliman, 114
Solorio, 104
Soltau, 241
Som, 400
Somasundaram, 408
Sommerauer, 334
Somov, 81
Son, 135, 212, 251, 258
Sonar, 153
Song, 18, 77, 98, 135, 137, 140, 153, 155, 165, 176, 191, 195, 199, 204, 208, 231, 238–240, 246, 252, 257, 266, 267, 270, 273, 282, 289, 292, 303, 306, 308, 321, 326, 339, 342, 350, 355, 357, 360, 374, 377, 390, 415
Soni, 226, 354, 401
Sonkar, 383
Soon, 108, 293
Sorensen, 250
Soricut, 248
Sotolar, 272
Sottana, 266
Sotudeh, 19
Soubki, 253
Soulier, 379
Sourati, 277
Souza, 157
Sow, 77
Sowański, 232
Spanakis, 305, 320
Spangher, 368
Sperber, 100, 104
Sperlì, 19
Spliethoff, 83
Spruit, 93, 393
Sravani, 72
Srba, 303, 391
Sreedhar, 162, 252, 291, 396
Sridhar, 387
Srikumar, 390
Srinath, 175
Srinivas, 106
Srinivasa, 382
Srinivasan, 219, 238, 252, 321, 342, 352, 365
Srivastava, 89, 186, 298, 316, 385, 391
Stürmer, 107, 108
Stacey, 424
Stafylakis, 236
Stahl, 331
Stahlberg, 19, 422
Staiano, 271
Staliūnaitė, 282
Stammbach, 108, 129, 265, 426
Stamou, 348
Stanojević, 218
Stap, 76, 182
Stappart, 376
Starace, 411
Staufer, 278
Steedman, 340
Stefanovitch, 275
Steimel, 85
Steinert-Threlkeld, 80
Steingrimsson, 77, 351
Stengel-Eskin, 141, 244
Stepanova, 79
Stephan, 237
Stepputtis, 323, 414
Stern, 107

Sterner, 104
Sterz, 173
Steuer, 94
Stevenson, 342
Stewart, 97, 414
Stiglitz, 141
Stogiannidis, 410
Stolf, 295
Stolfo, 313, 321
Stollenwerk, 86
Storchan, 355
Storks, 272, 381
Stoyanov, 176
Stratos, 167, 374
Strong, 143
Strube, 375
Strubell, 18, 169, 349, 384, 393
Strum, 108
Strötgen, 362
Stürmer, 157
Su, 18, 98, 154, 161, 173, 176, 181, 188, 198, 202, 208, 209, 239, 246, 280, 294, 308, 309, 318, 332, 353, 371, 377, 381, 413, 418, 423
Subramaniam, 234
Subramanian, 108, 131, 272, 291
Subramonian, 104
Suchanek, 294
Suchocki, 218
Sudhi, 398
Sudoh, 20
Suess, 398
Suesserman, 107
Sugawara, 71, 219
Sugiura, 93
Sugiyama, 388
Suglia, 148
Suhr, 18, 239
Sui, 163, 190, 208, 360
Sujaya, 424
Sulem, 276
Sullivan, 111
Sultan, 169, 299, 354
Suman, 99
Sun, 18, 72, 94, 125, 132, 135, 143, 151, 159, 161, 163, 165, 170, 176, 178, 187, 191, 193, 197, 200–202, 204, 208, 214, 220, 222, 237, 238, 248, 253, 264, 265, 270, 278, 286, 290, 291, 307, 309, 312, 313, 317, 328, 332, 336, 354–356, 375, 377, 389, 395, 398, 399, 401, 406, 413, 415, 418, 420
Sundaresan, 188, 268
Sundriyal, 192, 332
Sung, 134, 278, 308, 335, 373, 430
Suntorntip, 85
Suominen, 301
Surana, 288
Surani, 236
Surdeanu, 107, 108
Suresh, 318, 380
Suri, 355
Surikuchi, 148
Suriyawongkul, 85
Sutawika, 104
Suwaileh, 112
Suwono, 191
Suzuki, 18, 76, 211, 228, 295, 419
Suárez-Paniagua, 275
Sučik, 262
Svete, 196, 198
Swamy, 209
Swayamdipta, 18, 239
Sycara, 323, 414
Syed, 340, 348
Synnaeve, 219
Szpektor, 248, 305, 313
Søgaard, 171

T.Y.S.S, 154, 278
T.y.s.s, 107, 108
Taboada, 413
Tack, 19
Taffa, 88
Tafjord, 321, 324
Tafreshi, 20
Taghizadeh, 253
Tagliabue, 325
Tagliasacchi, 243
Taheri, 119
Tahmasebi, 223
Tai, 238, 270
Takabi, 385
Takamura, 131
Takeda, 312
Takeoka, 330
Takezawa, 355
Talafha, 111
Talat, 17, 308, 336
Talby, 112
Talukdar, 99, 191
Tam, 224
Tambwekar, 206
Tami, 355
Tamilselvam, 265

Tamine, 18
Tamiru, 89
Tamkin, 197
Tan, 18, 76, 104, 128, 132, 141, 155, 156, 170, 183, 187, 201, 202, 265, 271, 275, 285, 306, 321, 341, 359, 387, 394, 413
Tanaka, 112
Tanaya, 240
Tandon, 17, 274, 287, 378
Tang, 18, 88, 93, 128, 159, 180, 187, 224, 229, 252, 265, 280, 292, 311, 315, 328, 347, 355, 366, 385, 402, 415, 417
Tanguy, 345
Tanmay, 400, 411
Tanmoy, 119
Tanner, 421
Tao, 98, 99, 128, 182, 240, 253, 260, 268, 289, 342, 363
Tastet, 72
Tata, 260
Tatsuno, 129
Tay, 175, 196, 301, 308, 349
Taylor, 112
Tazi, 301
Tegnér, 392
Tekiroglu, 281
Telaar, 100, 104
Teleki, 413
Tenenbaum, 277
Teng, 131, 283, 291
Tensmeyer, 259
Teodorescu, 150, 384
Teslia, 76
Tetreault, 252, 283
Teucher, 398
Teufel, 104
Thai, 287
Thakkar, 191
Thalken, 141
Thapa, 113–115, 119
Thapliyal, 151, 248
Thawakar, 113, 204
Thawani, 343
The, 393
Thibault, 328
Thielk, 376
Thieme, 225
Thoma, 72
Thomas, 75
Thomason, 163, 239
Thompson, 75, 97, 238
Thorne, 19, 282, 295, 327, 335, 426
Thost, 180
Thounaojam, 99
Tian, 138, 197, 224, 235, 241, 242, 257, 265, 268, 269, 336, 337, 354, 359, 360, 363, 366, 404, 414
Tikhonov, 345
Tikhonova, 157
Tilli, 93
Tillmann, 174
Timiryasov, 72
Timkey, 288
Tinn, 225
Titov, 80, 81, 313, 321, 330, 334
Tits, 403
Tiwari, 114, 184, 199, 401
Tiwary, 400
Tizpaz-Niari, 106
Todd, 93, 195
Tomar, 234, 258
Tomeh, 110, 111
Tomlin, 334, 407
Tonelli, 82, 106
Toneva, 18
Tong, 90, 163, 423
Tonglet, 235
Tonja, 88, 281, 284, 381
Tonmoy, 226, 267
Toporkov, 172
Torii, 129
Torki, 115
Toshev, 219
Toshniwal, 377, 413
Touileb, 112
Tow, 184
Towle, 254
Toyin, 110
Trakic, 108, 293
Tran, 133, 196, 262, 349, 358, 405, 406
Trawick, 149
Treat, 86
Trenous, 99
Treviso, 98
Trikalinos, 146
Tripto, 407
Trischler, 295
Trivedi, 106, 174, 396
Troncy, 335
Truhn, 429
Trukhina, 236
Tsai, 359
Tsakalidis, 372
Tsarfaty, 167, 249, 339, 379
Tsatsaronis, 400

Tsipidi, 278
Tsunomori, 388
Tsur, 193
Tsvetkov, 56, 153, 265, 325, 376, 387
Tsvigun, 262
Tu, 19, 20, 75, 149, 156, 160, 192, 217, 239, 245, 268, 288, 293, 325, 347, 370, 386, 387, 389, 428
Tuck, 111
Tulpan, 193
Tun, 288
Turc, 258
Turcan, 324
Turchi, 97, 190, 193, 317
Turek, 71
Tutar, 399
Tuteja, 106
Tutek, 18
Tutubalina, 81
Tyagi, 136, 355
Tzimiropoulos, 236
Tzou, 405

Uban, 231
Uchendu, 259, 303, 407
Udagawa, 396
Uddin, 383
Udomcharoenchaikit, 80, 85, 281
Ugan, 351
Ullman, 412
Umapathi, 73
Unanue, 350
Ung, 170
Ungar, 85, 180, 206, 216, 314
Unni, 228
Unnithan, 231
Upadhyaya, 297
Urlana, 376
Urrutia, 324
Uryupina, 19
Usbeck, 88
Ushio, 288, 382
Usuyama, 200, 225
Uthus, 308
Utpala, 419

Vaduguru, 142
Vahtola, 301
Vaidya, 71
Valentini, 369
Valentino, 220
Vallejo, 282
Vallurupalli, 217
Vamvas, 99, 220
van de Loo, 400
van de Meent, 417
van der Goot, 20, 270, 334
van der Meer, 280
van der Plas, 20, 242
van Dijck, 305
van Dijk, 93, 393
Van Dorpe, 406
van Duijn, 93, 393
Van Durme, 108, 141, 178, 187, 226, 244
Van Esch, 19
van Genabith, 233, 256
van Niekerk, 130
van Rooij, 376
van Schijndel, 342
Van Stigt, 98
Vandenhirtz, 313
vanderPutten, 93
Vansh, 330
Varia, 206
Varoquaux, 294
Vashishth, 191
Vashishtha, 236, 284
Vashurin, 262
Vasilev, 262
Vaska, 206
Vasselli, 75
Vassos, 410
Vats, 108
Vaudaux, 108
Vaz, 97
Vazhentsev, 262
Vazirgiannis, 289, 338
Vazquez, 161
Vecchi, 20
Vechtomova, 362
Veeramachaneni, 170
Veeramani, 113–115
Veerubhotla, 326
Vejvar, 339
Velldal, 112
Veloso, 150, 296
Venezian, 289, 349
Venkatapathy, 343
Venkateswaran, 133, 242, 287
Venkatraman, 175
Venkit, 175
Venugopal, 208
Venugopalan, 19
Vepa, 395
Verberne, 93, 358
Verbruggen, 257, 429

Vergès, 108
Verga, 18
Verheul, 83
Verhoef, 200
Verkes, 341
Verma, 111, 112, 126, 259, 279, 334, 355
Veseli, 284
Vetzler, 340
Vezzani, 75
Vial, 108
Vianna, 411
Viaud, 301
Vicente Navarro, 75
Vidgen, 324
Vieira, 275, 278
Vig, 55
Vijay-Shanker, 113
Vijayakumar, 240
Vijayanarasimhan, 134
Vijayaraghavan, 404
Vijjini, 343
Vilar, 19, 78
Vilares, 20, 350
Vilarinho Lopes, 100
Villata, 20, 346
Villena, 89
Vincent, 233, 243
Vinh, 77
Vinzamuri, 343
Viskov, 98
Visokay, 414
Viswanathan, 396
Vives, 271
Vizcarra, 382
Vlachos, 143, 203, 282, 315, 373, 376
Vogel, 398
Vogler, 273
Vogt, 313
Voigt, 17, 130
Voita, 18, 219
Volodina, 223
von der Wense, 369
Vora, 390
Vosoughi, 164, 170, 349, 383
Voss, 150
Vossen, 280
Vreeken, 205
Vtyurina, 108
Vu, 19, 93, 111, 130, 179, 406, 422, 424
Vuli, 173
Vulić, 20, 136, 143, 147, 282, 391, 407, 410
Vuong, 107, 397
Vyas, 114, 196, 206, 264, 314
Vydiswaran, 19, 250
Vykopal, 391
Vylomova, 19, 82
Vásquez-Rodríguez, 231

W, 353
Wachsmuth, 331, 410
Wachspress, 414
Wada, 93, 284
Wagner, 258, 361
Waheed, 110, 111, 302
Wahle, 170, 235
Waibel, 351
Wal, 94
Waldendorf, 200
Walker, 347
Wallace, 94, 146, 184, 417
Wallbridge, 215
Walter, 111
Walther, 18, 386
Wambsganss, 209
Wan, 90, 130, 190, 200, 202, 211, 224, 249, 264, 266, 276, 291, 317, 326, 364, 375, 387, 410, 415
Wang, 18–20, 72, 75, 77, 78, 85, 88, 93, 98, 104, 107, 108, 110, 126, 128, 129, 131–133, 137–140, 144–146, 148, 151, 153, 155, 156, 158–164, 167–170, 172–182, 186–196, 198, 199, 201, 202, 204, 206–209, 212–215, 217, 218, 221, 224, 228–232, 234, 235, 239, 241, 243, 245–249, 252, 253, 257, 259, 260, 262–268, 270, 272, 274, 276, 278–280, 282–286, 288–292, 294–296, 298, 299, 301, 303–305, 308, 310, 312, 313, 315, 317–319, 322, 324, 325, 330, 332–337, 339–343, 345–350, 352, 356–359, 361–366, 368, 371–374, 376–379, 382–387, 391, 393–395, 397, 398, 400, 402–405, 407–410, 413, 414, 416–420, 423–427, 429, 430
Warikoo, 373
Warjri, 97
Warstadt, 215
Wasserblat, 271
Watanabe, 18, 75, 97
Watson, 150, 342
Way, 99, 351
Wazrah, 111
Webber, 161
Weber, 72, 80, 107, 369

Weber-Genzel, 270
Webersinke, 129
Webson, 412
Weerasooriya, 353
Wei, 76, 77, 164, 170, 175, 176, 186, 189, 193, 196, 203, 226, 228, 264, 270, 292, 301, 302, 314, 356, 376, 390, 399, 417, 418, 430
wei, 268
Weikum, 284
Wein, 263
Weinberger, 341
Weinreb, 134
Weinzierl, 346
Weischedel, 152
Weissweiler, 240, 329
Wekhof, 129
Welch, 20
Weld, 262
Welleck, 363, 425
Weller, 226
Wen, 125, 137, 139, 165, 190, 197, 223, 229, 234, 305, 310, 329, 357, 406, 420
Wen-Yi, 323
Wendt, 260
Weninger, 161
Wermter, 382
West, 129, 149, 160, 239, 250, 363, 365
Wetscherek, 225
Wettig, 183, 369
Weyers, 72
Whang, 290
White, 169, 178, 284
Whitehouse, 138
Wicke, 18
Wicks, 77
Widder, 393
Widiaputri, 240
Wiegand, 279
Wiegreffe, 18, 274, 324
Wieling, 83, 226
Wies, 108
Wiesner, 20
Wieting, 151, 197
Wijaya, 194, 273, 417
Wijnholds, 20, 72
Wilcox, 149, 154, 156, 215
Wilhelm, 262
Wilie, 134
Wilkens, 141, 274
Williams, 72, 78, 170, 399, 413
Williamson, 400
Wilson, 20, 80, 175, 207, 301
Winata, 104, 153, 302
Wind, 252
Winterstein, 406
Wintner, 302
Wirtz, 194
Wiseman, 167
Wisniewki, 19
Wisniewski, 77, 275
Woldeyohannis, 89
Wolf, 215, 260, 301
Wolfe, 222
Wolfson, 224
Wolovick, 90
Wong, 148, 180, 192, 200, 209, 233, 290–292, 328, 386
Wood, 220
Woodside, 280
Worku, 89
Woźniak, 252
Wright, 17, 369
Wright-Bettner, 395
WU, 353
Wu, 17, 18, 20, 55, 76, 77, 98, 108, 130–132, 136, 138–141, 146, 152, 153, 155–157, 162–165, 167, 169, 175, 177, 178, 180, 182–188, 194, 197, 202, 205, 206, 209, 212, 215, 228, 230, 233, 238, 239, 241, 244, 246, 250, 253, 254, 258, 260, 265, 266, 269, 278, 289, 298, 306, 307, 310, 312, 318, 320, 321, 329, 333, 334, 337, 339, 352, 354, 363, 366, 370–372, 383, 387–389, 394, 396, 398, 402, 422–424, 427–429
Wuebker, 19
Wullach, 225
Wumaier, 94

Xenos, 236
Xi, 251, 296
Xia, 17, 139, 234, 247, 266, 292, 335, 399, 410
Xian, 266
Xiang, 163, 194, 228, 357, 406
Xianlong, 276
Xiao, 19, 85, 94, 112, 132, 150, 151, 156, 159, 165, 195, 202, 214, 229, 234, 239, 248, 250, 268, 354, 392, 394, 395, 416, 420
Xie, 19, 76, 77, 85, 127, 145, 164, 170, 186, 200, 206, 209, 239, 248, 253, 260, 282, 289, 291, 307, 309, 323, 354, 360, 369, 373, 377, 406, 414, 418
Xin, 421

Xing, 107, 134, 307, 419
Xiong, 77, 132, 161, 176–178, 186, 206, 208, 231, 252, 254, 257, 288, 296, 305, 329, 336, 358, 374, 377, 386, 421
Xiu, 399
Xu, 18–20, 55, 71, 77, 78, 89, 125, 127, 132–135, 138, 140, 145, 148, 154, 163, 165, 172, 177, 178, 184, 187, 189, 190, 192, 195, 197, 199, 204, 206, 209, 214, 218, 219, 221, 222, 228, 231, 236–238, 241, 246, 253, 254, 257, 260, 266, 268, 274, 275, 278, 289–291, 296, 297, 299, 303, 304, 307, 308, 310–312, 330, 334, 338, 341, 351, 354, 358, 359, 362, 367, 379–381, 387, 392, 394, 397, 418, 419, 421, 430
Xuan, 18
Xue, 94, 248, 292, 325, 347, 368, 369, 427

Yadav, 400
Yadavalli, 281
Yagcioglu, 410
Yaghoobzadeh, 20, 410
Yalcin, 333
Yamada, 312, 315, 355, 356
Yamagiwa, 159, 310
Yamamoto, 300
Yamashita, 259, 303
Yan, 98, 133, 139, 149, 202, 207, 208, 214, 257, 272, 283, 307, 317, 332, 397, 409, 429
Yaneva, 19
Yang, 55, 76, 80, 86, 93, 94, 98, 99, 108, 125, 127, 128, 131, 135, 138, 139, 143–146, 149, 151, 158, 165, 170, 174, 175, 177, 182, 185–188, 190, 192–194, 200, 205, 211, 214, 217–219, 224, 225, 232, 234, 235, 237, 239, 245, 249, 260–262, 264, 276, 278, 289, 293, 295, 304, 307–309, 316, 319, 322, 325, 332, 334, 338, 354, 356, 357, 359, 362, 365–367, 369, 371, 375, 386, 389, 392, 397, 400, 401, 406, 415, 416, 420, 425, 428, 429
Yangarber, 179
Yanqing, 98, 99
Yao, 18, 140, 186, 188, 197, 206, 214, 221, 222, 238, 243, 262, 265, 268, 298, 315, 322, 351, 372, 377, 379, 380, 395, 397, 415
Yarom, 248
Yassine, 85
Yasunaga, 278
Yauney, 195
Yavuz, 254, 425
Yazdanbakhsh, 331
Ye, 17, 55, 56, 137, 177, 186, 189, 209, 211, 229, 248, 251, 268, 288, 306, 320, 328, 329, 338, 363, 371, 380, 381, 404, 424
Yeager, 322
Yeatman, 262
Yeganova, 75
Yen, 159, 246, 323, 385
Yenigalla, 401
Yeo, 240, 408
Yerukola, 184, 411
Yeung, 164, 337
Yi, 94, 170
Yifei, 314
Yih, 212, 377
Yildirim, 315
Yilmaz, 160, 381
Yimam, 88, 89, 218
Yin, 18, 55, 56, 152, 174, 196, 212, 214, 234, 263, 266, 269, 276, 305, 309, 310, 332, 342, 365, 386, 398, 406, 420
Yiu, 356
Yogatama, 356
Yokoi, 20, 159, 211, 420
Yon, 19
Yong, 104, 285
Yoo, 185, 409, 429
Yoon, 161, 164, 166, 188, 304, 408
Yoran, 224
Yosef, 288
Yoshida, 416
Yoshinaga, 399
Yoshino, 20, 297
You, 249, 251, 253, 328, 395, 409, 414
Young, 417
Youssef, 250
Yu, 17, 75–77, 99, 126, 128, 129, 134, 135, 137, 138, 165, 167, 169, 174, 181, 182, 186, 194, 198, 202, 208, 212, 220, 223, 226, 228, 230, 231, 234, 236, 238, 240, 243, 245, 252, 260, 267, 268, 279, 282, 283, 285, 291, 303, 306, 310, 318, 323, 334, 335, 345, 349, 357, 360, 381, 391, 397, 399, 401, 405, 410, 412, 414, 415, 418, 422, 424
Yuan, 94, 143, 144, 178, 190, 195, 200, 237, 254, 266, 270, 305, 337, 360, 372,

397, 399, 416, 419, 420, 423, 427
Yuanlong, 285
Yue, 159, 161, 190, 351, 372
Yuguchi, 297
Yun, 18, 146, 147, 151, 161, 201, 273, 380, 409
Yunès, 77
Yvon, 190, 331, 336, 350

Zablotskaia, 283
Zacharopoulos, 216
Zaghouani, 111
Zaheer, 287
Zahid, 221
Zaiem, 243
Zakir, 118
Zakizadeh, 415
Zalmout, 401
Zamai, 344
Zaman, 226, 316
Zamaninejad, 90
Zamarron, 262
Zampieri, 19, 119, 353
Zan, 352
Zantou, 89
Zaporojets, 331
Zaraket, 110, 111
Zaratiana, 110, 111
Zarrie, 130
Zarrieß, 83, 94
Zayats, 347
Zeghidour, 243
Zeinalipour, 110
Zeira, 302
Zekiye, 115
Zeldes, 17
Zelikman, 262
Zemel, 398
Zemlyanskiy, 308, 309
Zemánek, 111
Zeng, 76, 126, 145, 151, 164, 175, 192, 227, 230, 262, 294, 296, 321, 334, 354
Zeraati, 410
Zerva, 97
Zerveas, 370
Zesch, 19
Zettlemoyer, 78, 142, 212, 256, 287, 407
Zevallos, 411
Zhai, 274, 356, 370, 380
Zhan, 136, 185, 246, 264, 321
Zhang, 17–20, 71, 76, 78, 82, 89, 90, 94, 97–99, 104, 107, 108, 111, 113, 125–128, 130, 132–135, 137–141, 143–146, 148, 152–155, 158, 160–165, 167, 168, 170, 173, 174, 176–182, 185–187, 189–192, 197, 198, 200–202, 206, 207, 209, 214–216, 218, 219, 221, 224, 226, 228–232, 234, 236–240, 242, 243, 246–249, 251, 252, 256–258, 263–266, 268, 270, 272, 274, 276–286, 289, 291, 292, 296, 298, 299, 302–307, 309, 314–316, 319–321, 328, 329, 332–334, 336, 338, 342, 345, 347, 351, 352, 354, 356, 357, 359–364, 366–369, 371, 372, 374–376, 378, 380, 381, 383, 387–392, 395, 397, 399, 400, 402, 403, 405, 406, 408, 414–418, 420–422, 426, 430
Zhao, 17, 19, 71, 77, 89, 94, 98, 99, 107, 108, 126, 128, 137–140, 142, 156, 160–163, 165, 173, 186, 189, 190, 192, 198, 201, 206, 212, 218, 221, 223, 229, 230, 232, 241, 242, 249, 254, 257, 267, 274, 278, 282, 283, 287, 289, 292, 294, 305, 307, 310, 311, 314, 316, 317, 319, 328, 329, 332, 341–343, 346, 348, 349, 354, 355, 360, 370, 371, 376, 377, 394–396, 398, 400, 402, 404, 407, 409, 415, 418, 420, 422
Zheng, 77, 133–135, 152, 155, 164, 173, 189, 194–196, 207, 209, 219, 220, 229, 232, 235, 241, 247, 251, 253, 264, 268, 296, 305, 310, 317, 345, 363, 372, 381, 397, 419, 424
Zhifei Wang, 80
Zhiyi, 202
Zhong, 126, 178, 185, 189, 196, 214, 268, 270, 317, 324, 363, 365, 369, 391, 395
Zhongyi, 231
Zhou, 18, 75, 88, 89, 93, 97, 129, 137–139, 143, 144, 148, 152–154, 156, 157, 160, 165, 170, 171, 175, 184, 189, 192, 196–198, 204, 207, 210, 212, 213, 220, 222, 223, 227, 229, 231, 234, 243, 249, 251, 252, 254, 257, 259, 260, 264, 266, 267, 270, 273, 278, 282, 283, 288, 291, 294, 296, 300, 306, 307, 309, 313, 321, 323, 328, 332, 338, 341, 343, 346, 356, 357, 360, 367, 368, 370, 385, 387, 389, 395, 401, 403, 404, 410, 415, 416, 420, 421, 424, 425
Zhu, 76, 77, 85, 98, 107, 125, 129, 130, 139,

142, 144, 162, 167, 172, 174, 175, 178, 182, 186, 187, 194, 195, 198, 200, 208, 212, 214, 225, 228, 230, 231, 239, 240, 252, 253, 258, 262, 263, 265, 267, 270, 291, 295, 296, 307, 312, 323, 324, 334, 340, 343, 346–348, 355, 359, 363, 367, 387, 392, 397, 398, 400, 413, 414
Zhuang, 202, 358, 383
Zhuo, 108, 293, 294, 357, 397
Zhuocheng, 410, 426
Zhuravinskyi, 184
Ziems, 146, 365
Zimerman, 260
Zimina, 77
Ziser, 161, 319, 374
Zitouni, 114
Ziyadi, 423
Ziętkiewicz, 232
Zloch, 407
Zoicas, 231
Zong, 76, 77, 97, 160, 215, 316
Zope, 108
Zosa, 203
Zou, 129, 182, 207, 208, 225, 257, 266, 311, 361, 387
Zouhar, 97, 139, 261, 265
Zubiaga, 19
Zuccon, 202, 261
Zugarini, 310, 344, 403
Zuidema, 392
Zuluaga-Gomez, 238
Zuo, 89

GTCOM

**Global Tone Communication Technology Co., Ltd.,** established in 2013, is a world-leading big data and artificial intelligence company. It developed on its own key technologies of machine translation, big data, knowledge graph, artificial intelligence generation, human-machine interaction, and big model, serving key fields such as military and national security, scientific research data analysis, smart city and global strategic data, and helping global enterprise users collaborate across domains and make intelligent security decisions.

## Core Businesses

- Global defense and security
- Smart city
- Research data analysis
- Global strategic data

## Key technologies in machine translation

### GeWu big model

- Multi-lingual pre-training model
- Multi-modal pre-training model
- Multi-lingual machine translation big model
- Multi-lingual generative big model

### Cross-language big data governance

- Multi-fields big data algorithm models
- One-stop data governance analysis and solutions

### NLP

- **60+** languages
- **15** algorithm types
- **65+** algorithm services

### Knowledge graph

- Knowledge graph construction
- Knowledge modeling and training
- Knowledge graph open service
- Knowledge computing and reasoning
- Knowledge data management
- Knowledge operation management

### Machine translation

- Winning many international MT contests
- Supporting **60+** languages
- MT for finance, science, medical and other major vertical fields
- Multiple national key research projects
- High-quality bilingual parallel corpus with **8 billion** sentence pairs

Colossal-AI

# HPC-AI TECHNOLOGY

HELLO,

*HPC-AI Technology is a global company for **high-performance** computing and artificial intelligence, through self-developed technologies such as efficient multi-dimensional parallelism, heterogeneous memory management, large-scale optimization libraries, and adaptive task scheduling.*

***We have created Colossal-AI, a universal deep learning system for the era of large models.***

*UP TO ANY BUSINESS SCENARIO,*

*UP TO ANY LOW-COST SOLUTION,*

**& UP TO ANY LARGE MODEL**

## WE OFFER

### COMPLETE LIFECYCLE SERVICES

*Our service packages help to bring your AI acceleration goals to life. Benefit from Colossal-AI's state-of-the-art parallel processing, distributed computing and memory management capabilities.*

### INDUSTRY-LEADING TECHNOLOGY

- ***10x speedup***
- ***47x cost savings***
- ***>175B parameters***

*Realize a substantial ROI by using Colossal-AI to tap into the power of high performance computing for Machine Learning.*

### INTELLIGENT CLOUD PLATFORM

*Empower your AI journey with our cutting-edge Colossal-AI Platform – your gateway to stable AI model training, real-time model monitoring, and lightning-fast training inference acceleration, all on the cloud.*

## CONTACT US NOW

**Singapore**
2 Science Park Drive, Ascent, #01-05, Singapore 118222

**service@hpcaitech.com**

Github

Slack

www.hpc-ai.tech

github.com/hpcaitech/ColossalAI

***LET'S GROW WITH US***

LIVEPERSON®

# Supercharge enterprise growth and efficiency with generative AI-powered features

**We've been pioneering digital conversational technology for over 27 years. Today, our award-winning Conversational Cloud™ platform empowers hundreds of the world's leading brands to deliver Curiously Human™ experiences that drive extraordinary results.**

## Drive scientific innovation with Curiously Human LLMs

Discover the transformative power of our Curiously Human approach in maximizing LLMs for data science advancements and effective Conversational AI.

### DATA-DRIVEN EXCELLENCE

**Enhance personalization and relevance.**
Elevate your LLMs with the world's largest conversational dataset, sourced from over a billion monthly interactions.

*This wealth of data empowers our AI to understand your customers and provide uniquely tailored experiences.*

### AI WITH A HUMAN TOUCH

**Boost customer satisfaction and retention.**
Maintain grounded, factual, and industry-specific conversations with the support of over 350,000 skilled humans in the loop, who continuously refine our models.

*Our AI ensures your customer interactions are both accurate and engaging.*

### ACTIONABLE INSIGHTS

**Optimize performance and drive results.**
Harness the power of enterprise-level analytics and reporting that automatically delivers actionable insights.

*LivePerson's approach to conversational intelligence helps you make data-driven decisions to optimize customer experiences and drive results.*

### RESPONSIBLE AI

**Build trust and ensure compliance.**
Minimize the risk of bias and ensure ethical AI implementation by partnering with LivePerson, the founders of Equal AI.

*We've been spearheading standards and certification for responsible, safe, and secure AI since 2019.*

**Discover the LivePerson advantage**

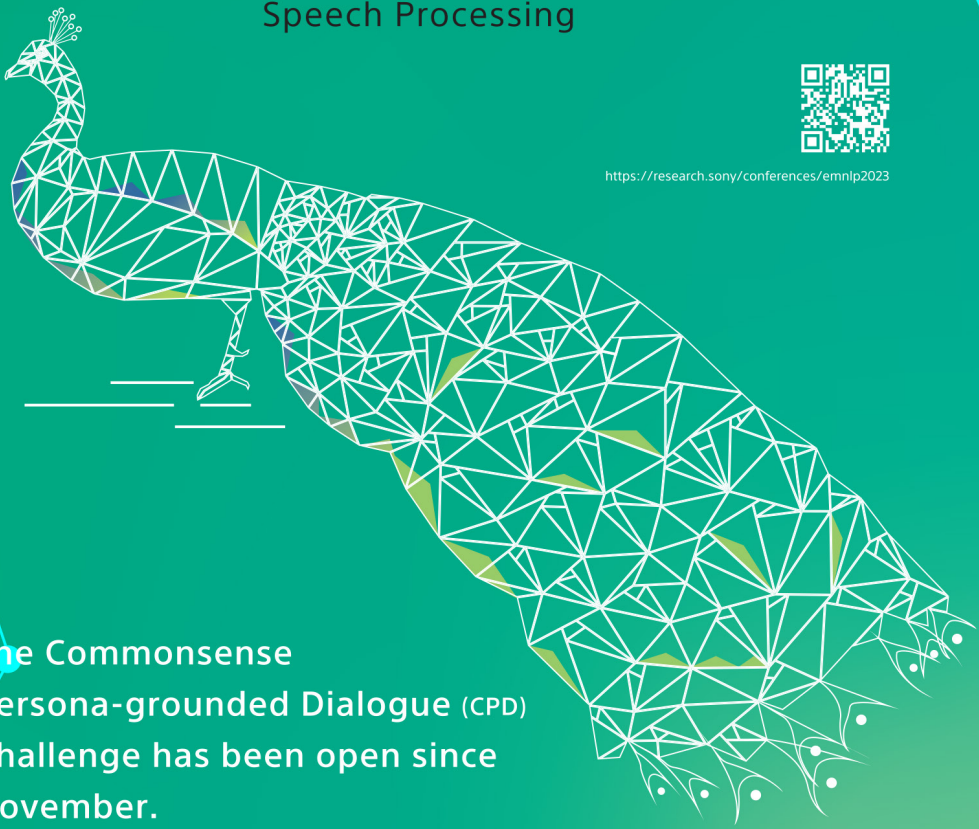
Visit our AI hub to learn more https://www.liveperson.com/ai/resources/

SONY

# We're looking for NLP researchers / engineers!

Vision-Language Pretraining, Text-to-Image/Sound,
Commonsense Knowledge Graphs,
Natural Language Generation, Dialogue Generation,
Speech Processing

https://research.sony/conferences/emnlp2023



The Commonsense Persona-grounded Dialogue (CPD) Challenge has been open since November.
Join the challenge!

Task 1 : Commonsense Dialogue Response Generation
Task 2 : Commonsense Persona Knowledge Linking

https://www.aicrowd.com/challenges/commonsense-persona-grounded-dialogue-challenge-2023

# Machine Learning
# Deep Learning
# Data Science
# Search
# LLM
# NLP
# IR
# CV
# AI

Think Common Crawl is big?
We're 100x bigger.

For over 10 years, Ahrefs has been crawling the web, storing and processing petabytes of data. Our web crawler ranks third globally after Bing and Google.

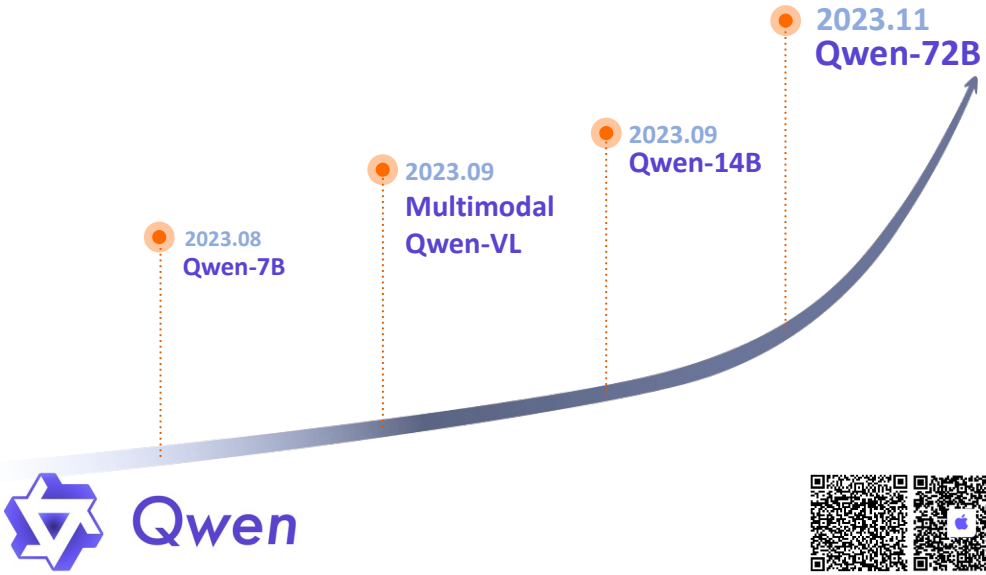
And now it’s powering our search engine, Yep.com, and our LLM too.

We have built Yep from the ground up to offer an unbiased, private search experience that rewards and compensates the makers behind the content.

ahrefs

# Qwen

A Continuously Evolving Large Language Model




Qwen

Android iOS

- **Strong Base Models in 7B, 14B, and 72B**

Stably pretrained for up to 3 trillion tokens of multilingual data with a wide coverage of domains, languages, achieving competitive performance on benchmark datasets.

- **Chat Models Aligned with Human Preference**

Chat, create content, extract information, summarize, translate, code, solve math problems, and so on.

- **Integrated Tool-using and Agent Abilities**

Be able to use tools like search engines and PDF interaction, play as agents, or even play as code interpreters, etc.

- **Easy Inference and Serving**

Supported by multi-scale quantization (int8, int4) for serving, inherent KV Cache, and multiple popular serving frameworks, such as vLLM and FastChat.

- **Quick Start on Finetuning and Customization**

Build your own finetuned Qwen with full parameters, LoRA, and Q-LoRA training without tears

## Open-source Resources

HuggingFace ModelScope Technical Report

## Contact

If you are interested to leave a message to either our research team or product team, join our Discord or WeChat groups!

GitHub Discord WeChat

**Email**

qianwen_opensource@alibabacloud.com

**Commercial Use**

https://dashscope.console.aliyun.com/openModelApply/qianwen

# Come build the future with us

At Amazon, we believe scientific innovation is essential to building Earth's most customer-centric company. Our scientists are conducting cutting-edge research in areas ranging from natural language processing to machine learning, operations, quantum computing, robotics, and more.

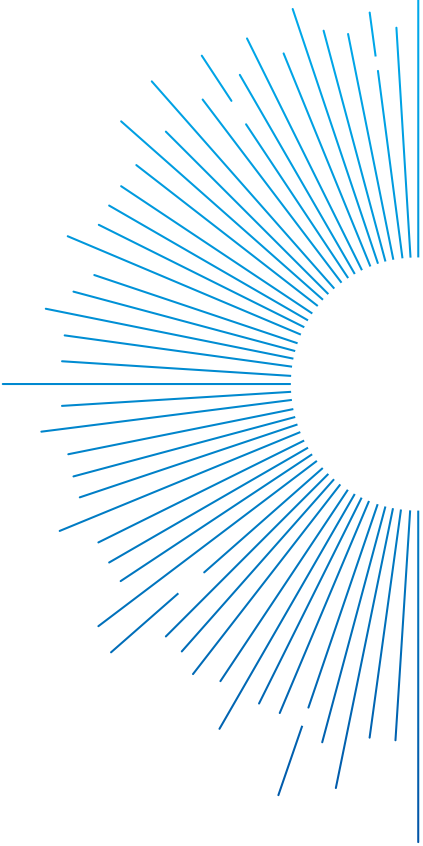
**Learn more about our research and papers published at EMNLP by visiting Amazon.Science. Want to drop us a note? Say hello at emnlp-conference-2023@amazon.com.**

amazon | science

# BAIDU NLP

## BAIDU NATURAL LANGUAGE PROCESSING

On a mission to enable machines to understand language and acquire intelligence so as to make the world better, Baidu NLP is dedicated to core NLP technologies, leading technology platforms and innovative products that are set to serve users across the globe and make the complex world simpler.



Baidu is the leading Chinese language internet search provider. Baidu aims to make the complicated world simpler through technology.

Email: nlp@baidu.com
Web: ai.baidu.com

ByteDance

Founded in 2012, ByteDance's mission is to inspire creativity and enrich life. With a suite of more than a dozen products, including TikTok, CapCut, Lark, and PICO, as well as platforms specific to the China market, including Toutiao, Douyin, Fanqie Novel and Xigua Video, ByteDance has made it easier and more fun for people to connect with, consume, and create content.

## NLP at ByteDance

ByteDance advances the state-of-the-art in all areas of AI, including Machine Learning, Knowledge and Data Mining, Natural Language Processing, etc. We build and improve the AI-systems that are the core of our products which are enjoyed by hundreds of millions of users worldwide. Here, you can join a team of world-class scientists, researchers and engineers to build large-scale, vastly powerful machine learning systems.

## Join Us

Connect with us to learn more about Bytedance NLP : lab-nr@bytedance.com

Make your inspiration infinite with a career at ByteDance.

Visit job.bytedance.com/en or scan the QR code to explore our opportunities!

For more info about ByteDance
https://www.bytedance.com/en/

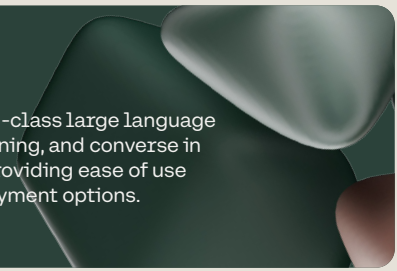
Let's Stay Connected | @ByteDancer | @ByteDance | @ByteDance

cohere

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 uses 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

## The Cohere Difference

**Customization (fine-tuning):** Cohere offers sophisticated fine-tuning tools and capabilities that enhance model performance for specific business tasks or domain knowledge, giving customers superior performance at industry-leading inference cost

**Scalability:** Cohere's models are packaged with inference engines that deliver better runtime performance at a lower cost than open-source equivalents

**Flexible deployment:** Models can be accessed through a SaaS API, cloud services (e.g., OCI, AWS SageMaker, Bedrock), and private deployments (VPC and on-prem)

**Privacy:** Customer data is never used in training base models, and customers have complete control over customization and model inputs/outputs

**Model integrity:** Models are trained from scratch from known, purchased, or public domain data sources and are subject to extensive adversarial testing and bias mitigation

To learn more about Cohere or discuss your business needs,
**contact us at: cohere.com/contact-sales**

# ABOUT US

Our mission is to empower people with better information to make their best decision.
Megagon Labs is an innovation hub within the Recruit Group, conducting top-notch research and building technologies in Natural Language Processing, Machine Learning, Data Management, Data Integratinon, Human-Computer Interaction, and Intelligent Visual Analytics.

For more information, visit www.megagon.ai

Megagon Labs

NEC Laboratories Europe

\Orchestrating a brighter world

NEC

# Building a better tomorrow

Combining explainable AI and NLP for human-AI collaboration.

Interpretability | Information Extraction | Knowledge Graphs | Complex Reasoning | AI Innovation

AI | Digital Health | ICT

neclab.eu

#NECLabs

# Welcome Future Engineers

Bloomberg
Engineering

At Bloomberg, the paths you can take are endless. We know because we're 21,000 unique individuals, each pursuing our own. Here, you can develop your skills, expand your experience, and think bigger about your career, all while always being your authentic self. Whatever you hope to make happen in your career, you can make it happen here.

Scan below to learn more.

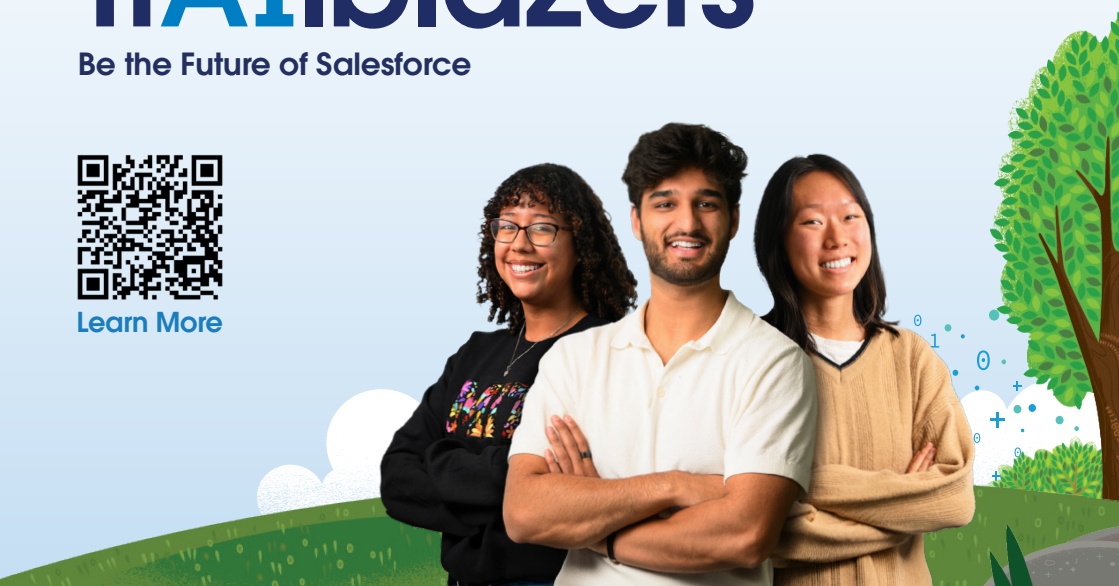
Make it happen here.

## Calling All TrAIlblazers

salesforce

Be the Future of Salesforce

Learn More



# JPMorgan Chase & Co.

**Visit Us at Booth #17**

MLCOE is a world class machine learning team which continually advances state-of-the-art methods to solve a wide range of real world financial challenges using our vast and unique datasets.

We actively partner and collaborate with business, data analytics, engineering and product teams across every function. From sales and trading desks to operations, digital, finance and risk, our MLCOE teams work with every JPMorgan Chase business.

## Our Businesses

- Corporate & Investment Banking
- Asset & Wealth Management
- Consumer & Community Banking
- Commercial Banking
- Corporate Functions

## Our Capabilities

- Large Language Models
- Natural Language Processing
- Speech Recognition
- Representation Learning
- Large Scale Computing
- Reinforcement Learning
- Time Series Analysis
- Recommender Systems
- Graph Analytics

## Your Opportunities

We're looking for problem-solvers with a passion for developing and applying innovative machine learning solutions.

MACHINE LEARNING Center of Excellence

jpmorgan.com/mlcoe

ANT GROUP

# 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

# NOAH'S ARK LAB OF HUAWEI TECHNOLOGIES

The Noah's Ark Lab is the AI research center for Huawei Technologies, located in Hong Kong, Shenzhen, Beijing, Shanghai, Xi'an, Nanjing, Hefei, London, Paris, Toronto, Montreal, Edmonton, etc.

The mission of the lab is to make significant contributions to both the company and society by innovating in artificial intelligence, data mining and related fields. Mainly driven by long term and big impact projects, research in the lab also tries to advance the state of the art in the fields as well as to harness the products and services of the company, at each stage of the innovation process.

As a world class research lab, we are pushing the frontier of research and development in all areas that we work in. We dare to address both the challenges and opportunities in this era of AI and big data, to revolutionize the ways in which people work and live, and the ways in which companies do business, through intelligentization of all processes, with the slogan 'from big data to deep knowledge'.

Research areas of the lab mainly include computer vision, natural language processing, search & recommendation, decision and reasoning ,AI theory and AI system engineering.

Founded in 2012, the lab has now grown to be a research organization with many significant achievements in both academia and industry. We welcome talented researchers and engineers to join us to realize their dreams.

## About Us

The Speech and Language Lab (of Noah's Ark Lab) dedicates to research and applications of speech and natural language processing and large-scale pre-trained models.

## Job Summary

We are hiring researchers in the related areas at all levels, including junior and senior positions. The researchers in the lab will conduct academic and applied research in the field of speech and natural language processing and deep learning, develop AI-enabled products and services with other groups in the company, as well as collaborate with world-class organizations in academia.

## Location

Positions available at the offices in Hong Kong, Shenzhen, Beijing, Montreal, London.

## Application

- Interested candidates should send application materials including resume to noahlab@huawei.com.
- More information about the lab is available at http://www.noahlab.com.hk

aiXplain

Build, diagnose, and improve AI systems
**continuously, efficiently, and effortlessly!**

aiXplain helps you create and maintain AI systems easily. You can design your own AI pipeline, benchmark your own model against others, monetize your own datasets, and accomplish much more with little to no effort.

aiXplain X yourself

We are hiring! Apply on **aixplain.com**

**Change the world, one word at a time**

Duolingo AI Research is a nimble and fast-growing group, revolutionizing language learning for more than 300 million people worldwide.

We're looking for creative ML/NLP researchers with interdisciplinary ideas to join our team. Help create the best language learning technology in the world for everyone, everywhere!

duolingo.ai

With Jenni AI | Before Jenni AI

Get published at
**EMNLP 2024**

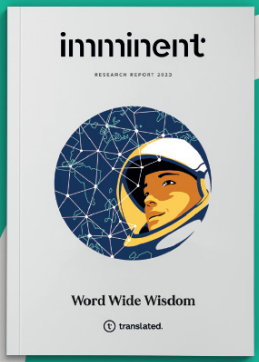
Write, edit, & cite with Jenni AI.
www.jenni.ai — scan & save 20%

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.



# EMNLP 2023 SPONSORS

## DIAMOND SPONSORS

LIVEPERSON

Google Research

Colossal-AI

GTCOM

SONY

مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language

## PLATINUM SPONSORS

amazon | science

Orchestrating a brighter world
NEC

Megagon Labs

Alibaba Cloud

ByteDance

ahrefs

cohere

Baidu 百度

## GOLD SPONSORS

SAP

HUAWEI

Bloomberg
Engineering

salesforce

JPMorgan Chase & Co.

ANT GROUP

## SILVER SPONSORS

aixplain

translated.

duolingo

jenni

## BRONZE SPONSORS

Adobe

NYONIC

Babelscape

mercari

面壁智能

## DIVERSITY & INLCUSION

Champion:

RIKEN AIP

Ally: amazon | science