XINRAN ZHAO

Email: xinranz3@andrew.cmu.edu | GitHub: colinzhaoust | Website: colinzhaoust.github.io

EDUCATION

Carnegie Mellon University *2023 - now, PA*
- Second-year PhD. student at Language Technologies Institute (LTI), advised by Sherry Tongshuang Wu.

Stanford University *2021 - 2023, CA*
- Master of Computer Science (GPA=3.95/4), advised by Shikhar Murty and Prof. Christopher Manning.

The Hong Kong University of Science and Technology *2016 - 2020, HK*
- Bachelor of Computer Science (GPA=3.96/4), advised by Dr. Hongming Zhang, Prof. Yangqiu Song, and Prof. Dit-Yan Yeung.
- Academic Achievement Medal (Top 1%), Continuous Undergraduate Scholarship (Top 5%), First Class Honors, Dean's List.

Cornell University *2018 - 2019, NY*
- International Exchange Program in Computer Science, advised by Dr. Esin Durmus and Prof. Claire Cardie.

SELECTED PUBLICATIONS

Conferences and Journals

Year 2024

1. Improving Large Language Model Planning with Action Sequence Similarity (Link)
To appear in **ICLR 2025**, **Xinran Zhao**, Hanie Sedghi, Bernd Bohnet, Dale Schuurmans, and Azade Nova

- *Brief*: We propose GRASE-DC, an iterative exemplar re-sampling method to boost LLM planning performance using in-context learning (ICL). Considering plan-side action similarity and diversity, it achieves up to 11-40 absolute points of accuracy improvement with 27.3% fewer exemplars on various tasks and different LLMs.

2. Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models (Link)
In Findings of **ACL 2024**, **Xinran Zhao**, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen

- *Brief*: We propose to analyze how prompting methods influence model confidence calibration. To mitigate the over-confidence issue revealed by the above analysis, we design Fact-and-Reflection to improve the model honesty by eliciting known facts, which reduces the calibration error by 23.5 percent on question-answering tasks.

3. GeoHard: Towards Measuring Class-wise Hardness through Modelling Class Semantics (Link)
In Findings of **ACL 2024**, Fengyu Cai, **Xinran Zhao**, Hongming Zhang, Iryna Gurevych, Heinz Koeppl

- *Brief*: We propose to extend the current advances on instance-level data hardness to class-wise hardness. We observe consistent hardness across models and methods on the classes of the same datasets. To measure the hardness and utilize the understanding to improve the model learning, we propose an empirical metric GeoHard to measure it and achieve 59% more correlation than baseline methods.

4.Beyond Relevance: Evaluate and Improve Retrievers on Perspective Awareness (Link)
In Proceedings of **COLM 2024**, **Xinran Zhao**, Tong Chen, Sihao Chen, Hongming Zhang, Tongshuang Wu

- *Brief*: we propose a novel retrieval benchmark, PIR, to study if and how current retrievers can handle nuanced perspective changes in user queries from real-world scenarios, e.g., find a supporting vs. opposing document for a claim. We present how current retrievers lack perspective awareness and design a projection-based method to improve them without extra training.

5. Dense X Retrieval: What Retrieval Granularity Should We Use? (Link)
In Proceedings of **EMNLP 2024**, Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, **Xinran Zhao**, Hongming Zhang, Dong Yu

- *Brief*: We systematically study the design choice of retrieval unit for dense retrievers solving open-domain NLP tasks. We discover that such choices will impact both retrieval and downstream tasks. We introduce proposition-based retrieval and show its empirical advances over passage and sentence-based methods.

6. "Merge Conflicts!" Exploring the Impacts of External Distractors to Parametric Knowledge Graphs (Link)
In Proceedings of **COLM 2024**, Cheng Qian, **Xinran Zhao**, Sherry Tongshuang Wu

- *Brief*: We systematically analyze the behavior and inner mechanism of large language models (LLMs) when the external knowledge provided in context is in conflict with LLMs' parametric knowledge. Experiments on both black-box and open-source LLMs demonstrate that LLMs tend to deviate from their internal knowledge when encountering direct conflicts or confounding changes.

7. MixGR: Enhancing Retriever Generalization for Scientific Domain through Complementary Granularity (Link)
In Proceedings of **EMNLP 2024**, Fengyu Cai, **Xinran Zhao**, Tong Chen, Sihao Chen, Hongming Zhang, Iryna Gurevych, Heinz Koeppl

- *Brief*: We introduce MixGR to alleviate the challenges in domain-specific retrieval with complex query-document relationships. MixGR improves dense retrievers' awareness of query-document matching across various levels of granularity using a zero-shot approach that achieves on average +24.7% and 9.8% on nDCG@5 with unsupervised and supervised retrievers on various domains.

8. HiMemFormer: Hierarchical Memory-Aware Transformer for Multi-Agent Action Anticipation (Link)
In VLM @ **NeurIPS 2024**, Zirui Wang, **Xinran Zhao**, Simon Stepputtis, Woojun Kim, Tongshuang Wu, Katia P. Sycara, Yaqi Xie

- *Brief*: We aim to achieve human-like action forecasting within multi-agent environments by presenting the Hierarchical Memory-Aware Transformer (HiMemFormer), as a transformer-based model for online multi-agent action anticipation. HiMemFormer integrates and distributes global memory that captures joint historical information across all agents and out-performs state-of-the-art models.

Before 2024

1. Thrust: Adaptively Propels Large Language Models with External Knowledge (Link)
In Proceedings of **NeurIPS 2023**, **Xinran Zhao**, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen

- *Brief*: We propose to improve the cost-efficiency and performance of the retrieval-augmented generation paradigm by scoring the queries at the instance level based on the model familiarity and rejecting potentially noisy and unnecessary external knowledge. We observe on average 26% improvement in 88% cases with various classification and open-domain question-answering tasks.

2. Video State-changing Object Segmentation (Link)
In Proceedings of **ICCV 2023**, Xiang Li, Jiangwei Yu, **Xinran Zhao**, Hongming Zhang, and Yu-Xiong Wang

- *Brief*: We identify learning about object state changes in Video Object Segmentation as a crucial task for understanding and interacting with the visual world and propose to extend the conventional VOS benchmark on static objects to a weakly supervised benchmark, VSCOS, focusing on identifying the state-changing objects in videos.

3. Towards Reference-free Text Simplification Evaluation (Link)
In Findings of **ACL 2023**, **Xinran Zhao**, Esin Durmus, and Dit-Yan Yeung

- *Brief*: We present *BETS* as a lightweight reference-free text simplification (TS) metric that leverages BERT and large-scale paraphrasing datasets to evaluate input-output pairs directly. Experiments show that *BETS* correlates better than existing metrics with human judgments. Controllable coefficients and reference-free properties further improve the applicability and transferability of TS models.

4. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (Link)
In **Transactions on Machine Learning Research (TMLR)**, a large-scale collaborative work

- *Brief*: We create two tasks based on WinoWhy, aiming to evaluate the language model's ability to conduct commonsense reasoning, which is combined and selected as one of the 24 featured tasks in BIG-bench Lite to provide a canonical measure of model performance.

5. On Measuring the Intrinsic Few-Shot Hardness of Datasets (Link)
In Proceedings of **EMNLP 2022**, **Xinran Zhao***, Shikhar Murty*, and Christopher Manning (*: equal contribution)

- *Brief*: We first show that few-shot hardness is empirically an intrinsic property of datasets by demonstrating the correlation among various methods. We then design an efficient metric that achieves better measurement of such intrinsic hardness than previous ones.

6. PCR4ALL: A Comprehensive Evaluation Benchmark for Pronoun Coreference Resolution (Link)
In Proceedings of **LREC 2022**, **Xinran Zhao**, Hongming Zhang, and Yangqiu Song

- *Brief*: We propose a novel benchmark to bridge the gap between document- and sentence-level PCR tasks, and evaluate the real-world robustness of systems through analysis from different angles including varying knowledge sources, domains, frequency, bias, and etc.

7. Weakly Supervised Text Classification Using Supervision Signals from a Language Model (Link)
In Findings of **NAACL 2022**, Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, **Xinran Zhao**, and Yangqiu Song

- *Brief*: We propose to improve weakly supervised learning by estimating label word distribution via probing models with various prompts. Empirical results show that involving the relations between such distribution and pre-defined categories leads to consistent gain.

8. Leveraging Topic Relatedness for Argument Persuasion (Link)
In Findings of **ACL 2021**., **Xinran Zhao**, Esin Durmus, Hongming Zhang, and Claire Cardie

- *Brief*: Previous study on argument persuasion mainly focuses on audience and language style factors. In this work, we model the relatedness among controversial topics and individuals' stances and leverage such to incorporate topic semantics in predicting persuasiveness.

9. Probing Toxic Content in Large Pre-Trained Language Models (Link)
In Proceedings of **ACL 2021**, Nedjma Ousidhoum, **Xinran Zhao**, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung

- *Brief*: We propose to quantify the potentially harmful content in large multilingual language models at scale through building a large set of probing queries generated by pattern-matching on Knowledge Graphs with respect to various potentially vulnerable social groups.

10. WinoWhy: A Deep Diagnose of Essential Commonsense Knowledge for Answering Winograd Schema Challenge (Link)
In Proceedings of **ACL 2020**, Hongming Zhang*, **Xinran Zhao***, and Yangqiu Song (*: equal contribution)

- *Brief*: Pre-trained language models show great performance on various benchmark on commonsense reasoning. However, we show that they are still far from solving such problems by challenging them with a transferable task requiring the identification of correct reasoning.

11. Learning Contextual Causality between Daily Events from Time-consecutive Images (Link)
In **Causality in Vision @ CVPR 2021**, Hongming Zhang, Yintong Huo, **Xinran Zhao**, Yangqiu Song, and Dan Roth

- *Brief*: We design a novel task on mining contextual causal knowledge from consecutive frames in videos and propose a Vision-Contextual Causal model as an effective way to represent the events in real-world images.

12. A Brief Survey and Comparative Study of Recent Development of Pronoun Coreference Resolution in English (Link)
In **CRAC @ EMNLP 2021**, Hongming Zhang, **Xinran Zhao**, and Yangqiu Song

- *Brief*: We survey on various models and datasets in the area of Pronoun Coreference Resolution and show that unfamiliar, out-of-domain, or knowledge-intensive examples and hyperparameter changes are still challenging problems for current models.

Other Projects

1. Can Pre-trained Language Models Understand Definitions?(Link)
CS 224N Final Project, Tina Li, Xiaoyuan Ni, **Xinran Zhao**

- *Brief*: Motivated by the psycholinguistic fact that humans show great capability in concluding and describing concepts, we propose to examine the PTLMs' ability to understand the definitions with a novel generalized word-sense matching task.

2. Seek to Embed ASER: A Large-scale Eventuality Knowledge Graph (Link)
In Proceedings of **WWW 2020**, An acknowledged contributor advised by Xin Liu and Yangqiu Song

- Designs a novel embedding model combining text embedding and graph embedding algorithms to learn the node representation for ASER, which has text as nodes and eventuality relations as edges. Provides useful signals for link prediction, unknown event resolution, and representation captioning on the graph, with a model utilizing BERT embeddings and LSTM.

RESEARCH AND WORK EXPERIENCE

Google DeepMind *May 2024 - Nov 2024, Mountain View*
Student Researcher. **Advisor:** Azade Nova and Hanie Sedghi

Tecent AI Lab @ Bellevue *June 2023 - Aug 2023, Bellevue*
Research Intern in NLP. **Advisor:** Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen

Stanford NLP Group *Oct 2021 - June 2023, Stanford*
Research Assitant. **Advisor:** Prof. Christopher Manning

Tecent AI Lab @ Bellevue *June 2022 - Sep 2022, Bellevue*
Research Intern in NLP. **Advisor:** Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen

HKUST ML Group *Sep 2020 - Feb 2021, HKUST*
Research Assitant. **Advisor:** Prof. Dit-Yan Yeung

HKUST KnowComp Group *June 2019 - August 2021, HKUST*
Research Assitant. **Advisor:** Prof. Yangqiu Song

INVITED TALKS AND ACADEMIC SERVICE

Invited Talks

Towards Building Retrieval Systems in Complex Realistic Scenarios: Samaya AI *Summer 2024*.

On Task Difficulty of Few-shot Learning: ICML 2022 Commonsense Tutorial, *Summer 2022*.

NLP as a Tool for Scientific Discovery: Department of Marketing @ the National University of Singapore, *Spring 2022*.

Service

Area Chair for ACL 2024, EMNLP 2024, NAACL 2025;

Reviewer for ECCV 2022, EMNLP 2022, AAAI 2023, ACL 2023, AACL 2023, EMNLP 2023, AAAI 2024, EACL 2024, NAACL 2024, NeurIPS 2024, and ICLR 2025.

SKILLS

Programming & Framework: Python/C++/Java/JavaScript. Familiar with Matlab, Rust, Racket, Lean; Pytorch, Tensorflow, Keras, AllenNLP, Spacy, NLTK, Amazon Turk, and etc.

Languages: English (Proficient, with TOEFL=112), Mandarin (Native).

Interests: Debate (HKUST Mandarin Debate Team: Championship for Bayarea Debate Invitational; Top 8 for the Sixth and Seventh International Mandarin Debate Invitational in Singapore), Archery, Gym, Seal Engraving, Basketball, Traditional Chinese Poem Writing