





YU ZHAO

✉ yu.zhao@ed.ac.uk  Semantic Scholar  GitHub  Twitter  LinkedIn

EDUCATION

University of Edinburgh Ph.D. candidate, supervised by Prof. Pasquale Minervini and Prof. Mirella Lapata	2023/09 - now
Harbin Institute of Technology, Shenzhen M.Sc. Computer Science, supervised by Prof. Xiaolong Wang and Prof. Baotian Hu	2020/09 - 2022/12
Hefei University of Technology B.Eng. Computer Science and Technology	2016/09 - 2020/06

EXPERIENCE

Xiaomi AI Lab, Beijing Natural Language Processing Engineer, Full-Time	2023/03 - 2023/08
--	-------------------

PUBLICATIONS

- Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering
Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, Pasquale Minervini
Preprint, 2024
- Analysing the Residual Stream of Language Models Under Knowledge Conflicts
Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, Pasquale Minervini
Foundation Model Interventions Workshop @ NeurIPS 2024
- A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression
Alessio Devoto*, **Yu Zhao***, Simone Scardapane, Pasquale Minervini (* denotes equal contribution)
EMNLP 2024 Main Conference, *Oral Presentation*
- Analysing The Impact of Sequence Composition on Language Model Pre-Training
Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, Pasquale Minervini
ACL 2024 Main Conference, *Oral Presentation*
- An Efficient Memory-Augmented Transformer for Knowledge-Intensive NLP Tasks
Yuxiang Wu, **Yu Zhao**, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel
EMNLP 2022 Main Conference, *Best Poster Award @ NeurIPS ENLSP*
- Medical Dialogue Response Generation with Pivotal Information Recalling
Yu Zhao*, Yunxin Li*, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, Min Zhang
KDD 2022
- Are We Done with MMLU?
Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, **Yu Zhao**, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, Pasquale Minervini
Preprint, 2024
- The Hallucinations Leaderboard – An Open Effort to Measure Hallucinations in Large Language Models
Giwon Hong*, Aryo Pradipta Gema*, Rohit Saxena*, Xiaotang Du*, Ping Nie*, **Yu Zhao***, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Pasquale Minervini*
Preprint, 2024
- Structured Packing in LLM Training Improves Long Context Utilization
Konrad Staniszewski*, Szymon Tworowski*, Sebastian Jaszczur, **Yu Zhao**, Henryk Michalewski, Łukasz

Kuciński, Piotr Miłoś
Preprint, 2023

10. Leveraging Pre-Training and Distillation Method for Title Generation with Limited Resource
Tianxiao Xu, Zihao Zheng, Xinshuo Hu, Zetian Sun, **Yu Zhao**, Baotian Hu
ICASSP 2023 SPGC Shared Task
11. MSDF: A General Open-Domain Multi-Skill Dialog Framework
Yu Zhao*, Xinshuo Hu*, Yunxin Li, Baotian Hu, Dongfang Li, Sichao Chen, Xiaolong Wang
NLPCC 2021

AWARDS AND SCHOLARSHIPS

The University of Edinburgh, CDT in NLP, 4-year PhD Scholarship	2023
Best Poster Award @ NeurIPS ENLSP 2022. Award: 750.00 USD	2022
Baidu Inc. Expert of Language Understanding and Generation Evaluation	2021
The First Prize Scholarship, Harbin Institute of Technology	2021
The First Prize Scholarship, Hefei University of Technology	2017

COMPETITIONS

ICASSP 2023 SPGC Shared Task: Title Generation with Limited Resource Won the 1st prize (1/47)	2023
2022 Language and Intelligence Challenge: Knowledge Grounded Dialogue Won the 3rd prize (4/610) in human evaluation and 1st place in automatic evaluation	2022
2021 Language and Intelligence Challenge: Multi-Skilled Dialogue Won the 3rd prize (6/750) in human evaluation and 8th place in automatic evaluation.	2021

SKILLS AND TOOLS

Programming Languages	Python, C/C++, Bash, SQL
Deep Learning Frameworks	PyTorch, JAX
Tools	Linux, git, Docker, L ^A T _E X, Scrapy, Django
Natural Languages	Mandarin, Eastern Min, English

ACADEMIC SERVICES

Program Committee Member/Conference Reviewer
EMNLP 2022, 2023; ACL 2023; CoNLL 2023, 2024; ACL SRW 2024; ICLR 2025

RESEARCH INTERESTS

Foundation Model Pre-Training
Efficient Training and Inference
Mechanism Interpretability