Haofei Yu

(+1) 412-537-0110
haofeiy2@illinois.edu
https://haofeiyu.me

Education

University of Illinois Urbana-Champaign — Champaign, IL
Ph.D. in Computer Science — 2024 - 2029 *(expected)*

- Advisor: Prof. Jiaxuan You
- Research area: Large Language Model, Language-based Agent, Multimodal Machine Learning

Carnegie Mellon University — Pittsburgh, PA
M.S. in Intelligent Information Systems — 2022 - 2024

- GPA: 4.00/4.00.
- TA for 11-777 Multimodal Machine Learning (2023 Fall)
- TA for 11-877 Advanced Topics in Multimodal Machine Learning (2024 Spring)
- A+ Coursework: Natural Language Processing, Advanced Topics in Multimodal Machine Learning, etc

Zhejiang University — Hangzhou, China
B.Eng. in Computer Science — 2018-2022

- GPA: 3.96/4.00. Rank 7/134.
- Outstanding Graduates. Provincial Government Scholarship 2019.
- A+ Coursework: Discrete Methematics and Application, Object-Oriented Programming, Computer Networks, etc

Publications

1. **Haofei Yu**†, Zhengyang Qi†, Lawrence Jang†, Ruslan Salakhutdinov, Louis-Philippe Morency, Paul Pu Liang. MMOE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts. *EMNLP 2024*.
2. **Haofei Yu**†, Cunxiang Wang†, Yue Zhang, Wei Bi. TRAMS: Training-free Memory Selection for Long-range Language Modeling. *Findings of EMNLP 2023*.
3. Ruiyi Wang†, **Haofei Yu**†, Wenxin Zhang†, Zhengyang Qi†, Maarten Sap, Graham Neubig, Yonatan Bisk, Hao Zhu. SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents. *ACL 2024*.
4. Cunxiang Wang†, **Haofei Yu**†, Yue Zhang. RFiD: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering. *Findings of ACL 2023*.
5. Chiyu Song, Hongliang He, **Haofei Yu**, Pengfei Fang, Leyang Cui, Zhenzhong Lan. Uni-Encoder: A Fast and Accurate Response Selection Paradigm for Generation-Based Dialogue Systems. *Findings of ACL 2023*.
6. Pengrui Han, Peiyang Song, **Haofei Yu**, Jiaxuan You. In-Context Learning May Not Elicit Trustworthy Reasoning: A-Not-B Errors in Pretrained Language Models. *Findings of EMNLP 2024*.
7. Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, **Haofei Yu**, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents. *ICLR 2024 (top 5% submission, Spotlight)*.
8. Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, **Haofei Yu**, Hinrich Schütze, Kemal Oflazer, David R Mortensen. Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model. *EMNLP 2023*.
9. Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, **Haofei Yu**, Ruslan Salakhutdinov, Louis-Philippe Morency. HEMM: Holistic Evaluation of Multimodal Foundation Models. *NeurIPS 2024 Dataset and Benchmark Track*.

Open-source Projects

Sotopia: an Open-ended Social Learning Environment
Core contributor and maintainer for the release of sotopia package sotopia ★ 147 ⑂ 18

Research Town: Simulator for Human Research Community
Leader and Creater for the release of research_town package research-town ★ 75 ⑂ 6

Industrial Experience

MIT-IBM Watson Lab | Boston, MA 2024.05 - 2024.08
- Conducted a comprehensive evaluation on open-sourced and close-sourced LLMs on SWE-bench.
- Proposed a fine-grained analysis on the skill difference between open-sourced LLMs and close-sourced ones.

Apple AIML | Seattle, WA 2023.05 - 2023.08
- Delivered an LLM-driven hierarchical prompting system including one document retriever, multiple summarizers, and one QA model to disambiguate and accurately respond to challenging real-world Siri user queries.
- Enhanced satisfaction on internal user data and chosen to present to Siri Senior Director *Robby Walker* (top 10 in SII).

Tencent AI Lab | Shenzhen, China 2022.02 - 2022.08
- Designed a training-free memory selection metric in Transformer-XL, gaining 0.19 perplexity drop on WikiText-103.
- Proposed a diffusion-based approach for NER, achieving results comparable to SpanBERT on CoNLL03 and OntoNotes.

Academia Experience

University of Illinois Urbana-Champaign | Urbana, IL 2024.08 - Now
- Proposed a benchmark for automatic research with large language models and built a multi-agent environment for research community simulation.

Carnegie Mellon University | Pittsburgh, PA 2022.08 - 2024.05
- Developed web shopping agents using Mistral-7B trained on Mind2Web, achieving near GPT-4 WebArena success rates.
- Proposed contrastive ranking loss for multi-task training on code generation, enhancing pass@100 by 6.65% on ODEX.
- Implemented LLM training on the SOTOPIA benchmark using ReST, reaching a 37.8% boost in social goal achievement.
- Introduced multimodal MoE to handle 3 self-defined multimodal interations, achieving a state-of-the-art MUStARD performance.
- Analyzed LLM's morphologicl ability in 4 languages and compared few-shot prompting with sub-word-level models.

Westlake University | Hangzhou, China 2021.02 - 2022.02
- Proposed Rational Fusion-in-Decoder, augmented with passage rationale training, achieving a 2.8% increase in accuracy on NaturalQuestions, a 0.9% rise on TriviaQA, and a 15.4% enhancement in understanding across multiple documents
- Proposed a SoTA response selection model with a 2.9% R10@1 improvement and 4x faster inference on Ubuntu-v2.