

## EDUCATION

## University of Illinois Urbana-Champaign

Urbana, IL

*Ph.D. in Computer Science*2024 - 2029 (*expected*)

- GPA: 4.00/4.00.
- Advisor: Prof. Jiaxuan You
- Research area: Large Language Model, Language 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 Mathematics and Application, Object-Oriented Programming, Computer Networks, etc

## PUBLICATIONS

1. **Haofei Yu**<sup>†</sup>, Zhengyang Qi<sup>†</sup>, Lawrence Jang<sup>†</sup>, Ruslan Salakhutdinov, Louis-Philippe Morency, Paul Pu Liang. MMOE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts. *EMNLP 2024*.
2. **Haofei Yu**<sup>†</sup>, Cunxiang Wang<sup>†</sup>, Yue Zhang, Wei Bi. TRAMS: Training-free Memory Selection for Long-range Language Modeling. *Findings of EMNLP 2023*.
3. Ruiyi Wang<sup>†</sup>, **Haofei Yu**<sup>†</sup>, Wenxin Zhang<sup>†</sup>, Zhengyang Qi<sup>†</sup>, Maarten Sap, Graham Neubig, Yonatan Bisk, Hao Zhu. SOTOPIA- $\pi$ : Interactive Learning of Socially Intelligent Language Agents. *ACL 2024*.
4. Cunxiang Wang<sup>†</sup>, **Haofei Yu**<sup>†</sup>, 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	<b>Research Town: Simulator for Human Research Community</b> <i>Leader and creator</i>	🔗 research-town ★ 116 📄 15
	<b>Sotopia: an Open-ended Social Learning Environment</b> <i>Core contributor and maintainer</i>	🔗 sotopia ★ 173 📄 23
INDUSTRIAL EXPERIENCE	<b>MIT-IBM Watson AI Lab</b>   Boston, MA	2024.05 - 2024.08
	<ul style="list-style-type: none"> <li>Conducted a comprehensive evaluation on open-sourced and close-sourced LLMs on SWE-bench.</li> <li>Proposed a fine-grained analysis on the skill difference between open-sourced LLMs and close-sourced ones.</li> </ul>	
	<b>Apple AIML</b>   Seattle, WA	2023.05 - 2023.08
	<ul style="list-style-type: none"> <li>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.</li> <li>Enhanced satisfaction on internal user data and chosen to present to Siri Senior Director <i>Robby Walker</i> (top 10 in SII).</li> </ul>	
ACADEMIA EXPERIENCE	<b>Tencent AI Lab</b>   Shenzhen, China	2022.02 - 2022.08
	<ul style="list-style-type: none"> <li>Designed a training-free memory selection metric in Transformer-XL, gaining 0.19 perplexity drop on WikiText-103.</li> <li>Proposed a diffusion-based approach for NER, achieving results comparable to SpanBERT on CoNLL03 and OntoNotes.</li> </ul>	
	<b>University of Illinois Urbana-Champaign</b>   Urbana, IL	2024.08 - Now
	<ul style="list-style-type: none"> <li>Proposed a benchmark for automatic research with large language models and built a multi-agent environment for research community simulation.</li> </ul>	
	<b>Carnegie Mellon University</b>   Pittsburgh, PA	2022.08 - 2024.05
	<ul style="list-style-type: none"> <li>Developed web shopping agents using Mistral-7B trained on Mind2Web, achieving near GPT-4 WebArena success rates.</li> <li>Proposed contrastive ranking loss for multi-task training on code generation, enhancing pass@100 by 6.65% on ODEX.</li> <li>Implemented LLM training on the SOTOPIA benchmark using ReST, reaching a 37.8% boost in social goal achievement.</li> <li>Introduced multimodal MoE to handle 3 self-defined multimodal interactions, achieving a state-of-the-art MUsTARD performance.</li> <li>Analyzed LLM's morphological ability in 4 languages and compared few-shot prompting with sub-word-level models.</li> </ul>	
	<b>Westlake University</b>   Hangzhou, China	2021.02 - 2022.02
	<ul style="list-style-type: none"> <li>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</li> <li>Proposed a SoTA response selection model with a 2.9% R10@1 improvement and 4x faster inference on Ubuntu-v2.</li> </ul>	
AWARDS	<b>Outstanding Reviewer</b>   EMNLP 2024	2024.11