Haofei Yu

haofeiy@cs.cmu.edu | haofeiyu.me | github | linkedin

Education

Carnegie Mellon University 2022/08 – 2024/05 (Expected)
Master of Science in Intelligent Information Systems, GPA **4.14**/4.33 Pittsburgh, PA, USA
- Teaching Assistant for 11-777 Multimodal Machine Learning (2023 Fall)

Zhejiang University 2018/09 – 2022/06
Bachelor of Engineering in Computer Science and Technology (with Honors), GPA **3.96**/4.00 Hangzhou, China
- Rank: 7/134, Provincial Scholarship (top 5%)

Publications

[1] **Haofei Yu**, Uri Alon, Graham Neubig. *Racoon: Ranked Code Generation*. Under review.
[2] 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*. Under review.
[3] **Haofei Yu**, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency. *Mixture of Multimodal Interaction Experts*. NeurIPS UniReps workshop.
[4] **Haofei Yu***, Cunxiang Wang*, Yue Zhang, Wei Bi. *TRAMS: Training-free Memory Selection for Long-range Language Modeling*. Findings of EMNLP 2023.
[5] Cunxiang Wang*, **Haofei Yu***, Yue Zhang. *RFiD: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering*. Findings of ACL 2023.
[6] Chiyu Song*, Hongliang He*, **Haofei Yu**, Leyang Cui, Pengfei Fang, Zhenzhong Lan. *Uni-Encoder: A Fast and Accurate Response Selection Paradigm for Generation-Based Dialogue Systems*. Findings of ACL 2023.
[7] Leonie Weissweiler*, Valentin Hofmann*, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, **Haofei Yu**, Hinrich Schuetze, 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.

Research Experience

Carnegie Mellon University - Language Technologies Institute 2022/09 – Present
Research Assistant Pittsburgh, PA, USA
- Worked in progress on finetuning language agents for web navigation in a simulated web browser environment.
- Proposed a multi-task training framework of CodeT5+ for code generation on CoNaLa and CONCODE benchmark.[1]
- Focused on training language agents using ReST-based and PPO-based methods in a simulated social environment.[2]
- Studied multimodal interaction classification and proposed an MoE-based method to handle different interactions.[3]
- Analyzed LLM's linguistic ability on the wug test and compared it with finetuned BERT for morphology benchmark.[7]

Westlake University - School of Enginnering 2021/02 – 2022/02
Research Assistant Hangzhou, China
- Trained a Fusion-in-Decoder Open-domain question-answering model with classification loss that has SoTA performance on NaturalQuestions and TriviaQA and better cross-attention understanding across multiple documents.[5]
- Proposed a BERT-based state-of-the-art response ranking model which is faster and better than cross-encoder.[6]

Work Experience

Apple - Siri & Information Intelligence 2023/05 – 2023/08
Machine Learning Intern Seattle, WA, USA
- Delivered an LLM-driven hierarchical prompting system including 1x large-scale document retriever, Nx LLM summarizers, and 1x LLM 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 Senior Director *Robby Walker* (top 10 in SII).

Tencent - Tencent AI lab 2022/02 – 2022/08
Research Intern Shenzhen, China
- Designed an effective query-independent memory selection metric on reformulated attention for Transformer-XL.[4]
- Proposed a diffusion-based method for the Named Entity Recognition task with competitive results with SpanBERT.