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**/4.00 Pittsburgh, PA, USA
- Teaching Assistant for 11-777 Multimodal Machine Learning (2023 Fall)

Zhejiang University 2018/08 – 2022/06
Bachelor of Engineering in Computer Science and Technology (with Honors), GPA **3.96**/4.00 Hangzhou, China
- Rank: 7/134, Outstanding Graduate, Provincial Scholarship (top 5%)

Publications

[1] Racoon: Ranked Code Generation
Haofei Yu, Uri Alon, Graham Neubig. Under preparation.
[2] SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents
Xuhui Zhou*, Hao Zhu*, Leena Mathur, Ruohong Zhang, **Haofei Yu**, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, Maarten Sap. Submitted to *ICLR*, 2024, under review.
[3] MMOE: Mixture of Multimodal Interaction Experts
Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency. . In *NeurIPS UniReps workshop*, 2023.
[4] TRAMS: Training-free Memory Selection for Long-range Language Modeling
Haofei Yu*, Cunxiang Wang*, Yue Zhang, Wei Bi. In *Findings of EMNLP*, 2023.
[5] RFiD: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering
Cunxiang Wang*, **Haofei Yu***, Yue Zhang. In *Findings of ACL*, 2023.
[6] Uni-Encoder: A Fast and Accurate Response Selection Paradigm for Generation-Based Dialogue Systems
Chiyu Song*, Hongliang He*, **Haofei Yu**, Leyang Cui, Pengfei Fang, Zhenzhong Lan. In *Findings of ACL*, 2023.
[7] Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model
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.. In *EMNLP*, 2023.

Research Experience

Carnegie Mellon University - Language Technologies Institute 2022/09 – Present
Research Assistant Pittsburgh, PA, USA
- 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 [1].
- Implemented LLM training on the SOTOPIA benchmark using ReST, reaching a 37.8% boost in social goal achievement [2].
- Introduced multimodal MoE to handle 5 self-defined multimodal interations, achieving a 2% MUStARD improvement [3].
- Analyzed LLM's morphologicl ability in 4 languages and compared few-shot prompting with sub-word-level models [7].

Westlake University - School of Enginnering 2021/02 – 2022/02
Research Assistant Hangzhou, China
- 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 [5].
- Proposed a SoTA response selection model with a 2.9% R10@1 improvement and 4x faster inference on Ubuntu-v2 [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 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 Senior Director *Robby Walker* (top 10 in SII).

Tencent - Tencent AI lab 2022/02 – 2022/08
Research Intern Shenzhen, China
- Designed a training-free memory selection metric in Transformer-XL, gaining 0.19 perplexity drop on WikiText-103 [4].
- Proposed a diffusion-based approach for NER, achieving results comparable to SpanBERT on CoNLL03 and OntoNotes.