ENYU ZHOU

eyzhou19@fudan.edu.cn · (+86)198-2181-0575 · Enyu Zhou - Google Scholar

EDUCATION

Fudan University, Shanghai *2023.09 – Present*
Ph.D student (Outstanding Doctoral Program) - Computer Science and Technology
Research Interests: Alignment, Reasoning; Advisor: Prof. Xuanjing Huang

Fudan University, Shanghai *2019.06 – 2023.06*
BS in Engineering - Intelligent Science and Technology (Honors Program)

SELECTED PUBLICATIONS

Enyu Zhou, et al. "Steer LLMs via Scalable Interactive Oversight." (**in submission**) [PDF]

- Studied the scalable oversight challenge in agentic LLM systems by introducing an interaction agent that decomposes vague user intent into structured, low-effort decisions for pre-execution alignment.

Enyu Zhou, et al. "RMB: Comprehensively Benchmarking Reward Models in LLM Alignment." (**ICLR'2025**) [PDF]

- Proposed a fine-grained reward model benchmark covering pairwise and Best-of-N paradigms, and demonstrated strong correlation with downstream alignment performance.

Shihan Dou*, **Enyu Zhou***, et al. "LoRAMoE: Alleviating World Knowledge Forgetting in Large Language Models via MoE-Style Plugin." (**ACL'2024**) [PDF]

- Explored challenges of world knowledge forgetting during large-scale fine-tuning and proposed LoRAMoE, an architecture combining LoRA and MoE to mitigate knowledge conflicts and enhance multitask capabilities of LLMs.

Enyu Zhou, et al. "RealBehavior: A Framework for Faithfully Characterizing Foundation Models' Human-like Behavior Mechanisms." (**EMNLP'2023**) [PDF]

- Developed a psychometrics-based framework for characterizing human-like LLM behaviors, with automated evaluation and analysis of alignment training effects.

Shihan Dou, Yan Liu, **Enyu Zhou**, Songyang Gao, et al. "Alleviating Shifted Distribution in Human Preference Alignment through Meta-Learning." (**AAAI'2025**) [PDF]

- Proposed a meta-learning method to address distribution shifts in RLHF by alternating optimization of the reward model to adapt to shifted samples and distributions.

Zhiheng Xi, et al. "The Rise and Potential of Large Language Model-Based Agents: A Survey." (**SCIS**, **1000+ citations**; **Co-First Author**) [PDF]

- Conducted a comprehensive survey on the rise and future of LLM-based agents, including construction, applications, and open challenges.
- Featured on GitHub Trending and PaperWithCode Top Trending Research.

INDUSTRY EXPERIENCES

Research Intern, China Qijizhifeng Ltd.Co, Shanghai *2024.12 – Present*

- Core contributor to the post-training of a agentic model, focusing on coding abilities. [Nex-N1]
- Led post-training of long-CoT models, achieving frontier performance on math benchmarks such as AIME.
- Investigated entropy collapse phenomena during post-training and designed mitigation algorithms.
- Explored scalable interaction and scalable oversight strategies for model alignment and training.

HONORS AND AWARDS

- Fudan University Huatai Securities Technology Scholarship *2024.12*
- Shanghai Outstanding Graduate Award *2023.06*
- National Scholarship for Undergraduates (Selected as an Outstanding Example) *2021.12*
- Top Ten Students", School of Information Science and Engineering, Fudan University *2021.12*