Zhanhui Zhou

zhziszz.github.io / asap.zzhou@gmail.com / 86-186-2154-1215

Education

University of Michigan (dual degree), Ann Arbor, MI 2020 - 2022
B.S.E., Computer Science (Highest Honors)
GPA: 4.00/4.00

Shanghai Jiao Tong University (dual degree), Shanghai, China 2018 - 2020
B.S.E., Computer Engineering (Highest Honors)
GPA: 3.78/4.00

Professional Experience

Shanghai AI Lab (AI Alignment Team), Shanghai, China 2022 - Now
Research Engineer, full-time
- Led research on generative model alignment and AI safety.

Shanghai AI Lab (Embodied-AI Team), Shanghai, China 2022
Research Intern, part-time
- Led the development of a large-scale distributed reinforcement learning infrastructure for autonomous Minecraft agents.

Selected Publications

* indicates equal contribution

Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, Yu Qiao. 2024
Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models. *Preprint, under review.*

Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang, Wanli Ouyang, 2024
Yu Qiao. Emulated Disalignment: Safety Alignment for Large Language Models May Backfire! **Outstanding Paper Award ($<$ 1% of all submission)**. *Annual Meeting of the Association for Computational Linguistics (ACL).*

Zhanhui Zhou*, Jie Liu*, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli 2024
Ouyang, Yu Qiao. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. *Findings of the Association for Computational Linguistics ACL (ACL Findings).*

Zhanhui Zhou*, Man To Tang*, Qiping Pan*, Shangyin Tan, Xinyu Wang, 2022
Tianyi Zhang. INTENT: INteractive TENsor Transformation Synthesis. *Symposium on User Interface Software and Technology (UIST).*

Other Publications

* indicates equal contribution

Jie Liu*, **Zhanhui Zhou***, Jiaheng Liu, Xingyuan Bu, Chao Yang, Han-Sen Zhong, Wanli Ouyang. Iterative Length-Regularized Direct Preference Optimization: A Case Study on Improving 7B Language Models to GPT-4 Level. *Preprint, under review.* 2024

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, **Zhanhui Zhou**, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, Wanli Ouyang. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. *Annual Meeting of the Association for Computational Linguistics (ACL).* 2024

Yanan Wu, Jie Liu, Xingyuan Bu, Jiaheng Liu, **Zhanhui Zhou**, Yuanxing Zhang, Chenchen Zhang, Zhiqi Bai, Haibin Chen, Tiezheng Ge, Wanli Ouyang, Wenbo Su, Bo Zheng. ConceptMath: A Bilingual Concept-wise Benchmark for Measuring Mathematical Reasoning of Large Language Models. *Findings of the Association for Computational Linguistics ACL (ACL Findings).* 2024

Zhichen Dong*, **Zhanhui Zhou***, Chao Yang, Jing Shao, Yu Qiao. Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).* 2024

Academic Awards

College of Engineering - Dean's Honor List (Umich) All semesters

Second-Class Academic Excellence Scholarship (top 10%) (SJTU) All semesters

Service

NeurIPS Conference Reviewer 2024

ICLR Conference Reviewer 2024

Skills

Language: English (fluent), Chinese (native)

Programming Language: Python, C++, C

Skills: PyTorch, TensorFlow, LaTeX