

Zhanhui Zhou

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

Education

University of Michigan (dual degree) , Ann Arbor, MI B.S.E., Computer Science (Highest Honors) GPA: 4.00/4.00	2020 - 2022
Shanghai Jiao Tong University (dual degree) , Shanghai, China B.E., Electrical and Computer Engineering (Highest Honors) GPA: 3.80/4.00	2018 - 2022

Professional Experience

Shanghai AI Lab (AI Alignment Team) , Shanghai, China Research Engineer, full-time ○ Led research on generative model alignment and AI safety.	2022 - Now
Shanghai AI Lab (Embodied-AI Team) , Shanghai, China Research Intern, part-time ○ Led the development of a large-scale distributed reinforcement learning framework for autonomous Minecraft agents.	2022

Selected Publications

* indicates equal contribution

Zhanhui Zhou , Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, Yu Qiao. Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models. <i>Neural Information Processing Symposium (NeurIPS)</i> .	2024
Zhanhui Zhou , Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang, Wanli Ouyang, Yu Qiao. Emulated Disalignment: Safety Alignment for Large Language Models May Backfire! Outstanding Paper Award (< 1% of all submissions) . <i>Annual Meeting of the Association for Computational Linguistics (ACL)</i> .	2024
Zhanhui Zhou* , Jie Liu*, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, Yu Qiao. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. <i>Findings of the Association for Computational Linguistics ACL (ACL Findings)</i> .	2024
Zhanhui Zhou* , Man To Tang*, Qiping Pan*, Shangyin Tan, Xinyu Wang, Tianyi Zhang. INTENT: Interactive Tensor Transformation Synthesis. <i>Symposium on User Interface Software and Technology (UIST)</i> .	2022

Other Publications

* indicates equal contribution

Zhixuan Liu*, Zhanhui Zhou* , Yuanfu Wang, Chao Yang, Yu Qiao. Inference-Time Language Model Alignment via Integrated Value Guidance. <i>Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)</i> .	2024
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. <i>Preprint, under review</i> .	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. <i>Annual Meeting of the Association for Computational Linguistics (ACL)</i> .	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. <i>Findings of the Association for Computational Linguistics ACL (ACL Findings)</i> .	2024
Zhichen Dong*, Zhanhui Zhou* , Chao Yang, Jing Shao, Yu Qiao. Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey. <i>Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)</i> .	2024

Academic Awards

College of Engineering - Dean’s Honor List (UMich)	All semesters
Second-Class Academic Excellence Scholarship (top 10%) (SJTU)	All semesters

Skills

Language: English (fluent), Chinese (native)

Programming Language: Python, C++, C

Skills: PyTorch, TensorFlow, \LaTeX