Hantao Lou

✉ lht_pku@stu.pku.edu.cn
🏠 htlou.github.io

○ Github   G Google Scholar   🐦 Twitter

Education

2022–2026 : **Tong Class, Yuanpei College, Peking University**.
B.S. Student in Artificial Intelligence

Publications

Arxiv Articles

2023 **AI Alignment: A Comprehensive Survey**, *Arxiv Preprint*.

Jiaming Ji*, Tianyi Qiu*, Boyuan Chen*, Borong Zhang*, **Hantao Lou**, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, Wen Gao

2024 **Aligner: Achieving efficient alignment through weak-to-strong correction**, *Arxiv Preprint*.

Jiaming Ji*, Boyuan Chen*, **Hantao Lou**, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Yaodong Yang

Research Experience

2023 – present **PAIR Lab: PKU Alignment and Interaction Research Lab**.
Currently working on Alignment (Especially Value Alignment & Safety Alignment) and Interpretability of Language Models.

Advisor : **Dr. Yaodong Yang**, *Deputy director of Centre for AI Safety and Governance at the Institute for AI, Peking University.*(*Personal Web-page*)

Fellowships & Awards

2022 Peking University Freshman Scholarship (¥10000 RMB)