Hantao Lou

✉ hantaolou.htlou@gmail.com
htlou.github.io
Github/htlou Google Scholar Twitter/@HantaoLou

Summary

I'm a second-year undergraduate from Peking University ('26), and I'm currently working on AI alignment. Specifically, I focus on Alignment Algorithms, Mechanistic Interpretability(or for short, mech interp), and other potentially scalable methods. My research questions are:

- How can the findings from mechanistic interpretability be effectively integrated into practical applications, including the alignment process?
- How can the fundamental nature of intelligence be uncovered through the interpretation of various models that exhibit intelligent behavior?

I'm still learning mech interp, so it is quite possible that I may revise these questions :)

Education

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

Fellowships & Awards

2022 Peking University Freshman Scholarship (¥10000 RMB)

Research Experience

2023 – **Visiting Student Researcher at PAIR Lab: PKU Alignment and Interaction Research Lab**.
Currently working on Alignment (Especially Value Alignment & Safety Alignment) and Interpretability of Language Models under the guidance from Dr. Yaodong Yang.

2024 Summer **Scholar at MATS(Machine Learning Alignment & Theory Scholars) Program**.
Working under Evan Hubinger's Mentorship

Projects

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, In Submission*.

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

2024 **Language Models Resist Alignment**, *Arxiv Preprint, In Submission*.

Jiaming Ji*, Kaile Wang*, Tianyi Qiu*, Boyuan Chen*, Jiayi Zhou, Changye Li, **Hantao Lou**, Yaodong Yang

Opensource Projects

2024 **Align-Anything**, *Github repo*.

An open-source framework for multimodal alignment. I am one of the main contributors.