

Summary

I'm a third-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

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

Selected Projects

Papers

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**, *2024 NeurIPS Oral (0.5%)*.

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

Open-source Projects

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

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

Blog Posts

2024 **Automating LLM Auditing with Developmental Interpretability**, *Lesswrong post*.

An introduction of my work done in MATS 2024 summer cohort supervised by Evan Hubinger. I'm still working on it for more insights :)