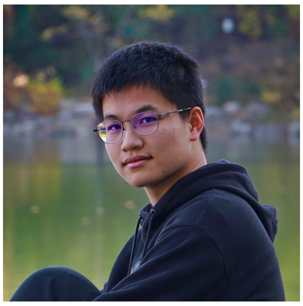

Hantao Lou

Peking University; Yuanpei College; Undergrad Student (graduating at 26)
Seeking internship/PhD positions in AI alignment & interpretability
Email: lht_pku@stu.pku.edu.cn
Homepage: htlou.github.io

About Me

Hantao Lou is a third-year undergrad student at Yuanpei College, Peking University, majoring in Artificial Intelligence, supervised by Prof. Yaodong Yang. His research mainly focuses on large language model alignment, aiming to develop safe, interpretable, ethical, and scalable AI systems. He has participated in **6 papers** (including **3 first/co-first author ones**), which are published in top conferences like ICML, NeurIPS and AAAI, with **300+** Google Scholar citations. He served as reviewer for top ML conferences including NeurIPS, ICLR, ICML, and AAAI. As an active contributor of open-source projects, Hantao is the repository creator and the core developer of multiple open-source projects with **3,400+** stars, and he participated in the alignment, deployment and the open-source of Hong Kong HKGAI-104B large models. His research on AI safety, interpretability and ethics has been cited by Meta AI, OpenAI, CMU, and Bengio, and has been featured in MIT Tech Review. He was invited to present his work at Wisemodel, a leading Chinese open-source community. Notably, Hantao authored Chapter 4 of the groundbreaking "AI Alignment: A Comprehensive Survey," focusing on evaluation and interpretability. This work has served as a key technical resource for global AI safety initiatives, including the Bletchley Summit, Beijing International AI Safety Consensus, and Venice AI Safety Consensus Conference.

Internship Experience

2024.06 – 2024.10	Machine Alignment & Theory Scholars Program | **Interpretability of Large Language Models**
2023.07 – Now	Peking University, Center of AI Safety and Governance, PAIR Lab | **Alignment of Large Language Models**

Courses

2024 Fall	Large Language Models: Foundations and Alignment	95
2024 Fall	Directed Research in AI Systems	94
2024 Fall	Natural Language Processing with Deep Learning	91
2023 Spring	Multi-agent Systems	95
2023 Spring	Mathematical Foundation for Artificial Intelligence	93

Open-Source

- **PKU-Alignment/align-anything, Stars: 3.4K+**
  - a highly modular open-source multimodal alignment framework, supporting multiple architecutres and models
  - the repository creator and the core developer

Publication (For a detailed list, please refer to Google Scholar)

1. **[ICML 2025] Hantao Lou***, Changye Li*, Jiaming Ji, Yaodong Yang; SAE-V: Interpreting Multimodal Models for Enhanced Alignment
2. **[AAAI 2025] Hantao Lou**, Jiaming Ji, Kaile Wang, Yaodong Yang; Stream Aligner: Efficient Sentence-Level Alignment via Distribution Induction
3. **[NeurIPS 2024 Oral]** Jiaming Ji*, Boyuan Chen*, **Hantao Lou**, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, Yaodong Yang; Aligner: Efficient Alignment by Learning to Correct

Arxiv

1. Jiaming Ji*, Jiayi Zhou*, **Hantao Lou***, Boyuan Chen*, Donghai Hong*, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, Yaodong Yang. Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback. CoRR abs/2412.15838 (2024)
2. Jiaming Ji*, Kaile Wang*, Tianyi Qiu*, Boyuan Chen*, Jiayi Zhou, Changye Li, **Hantao Lou**, Yaodong Yang. Language Models Resist Alignment. CoRR abs/2406.06144 (2024)

3. 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. AI Alignment: A Comprehensive Survey. CoRR abs/2310.19852 (2023)