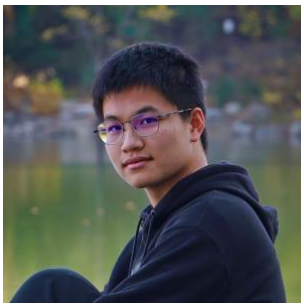
Hantao Lou

Peking University; Yuanpei College; Undergrad Student (‘26)
Email: hantaolou.htlou@gmail.com / lht_pku@stu.pku.edu.cn
Homepage: htlou.github.io



About Me

Hantao Lou is a final-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 provide and utilize supervision signals for scalable AI alignment.

Hantao has participated in **6 papers** (including **3 first/co-first author ones**), which are published in top conferences like ICML, ACL, NeurIPS and AAAI, with **650+** 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 **4,500+** stars, and he participated in the alignment, deployment and the open-source of Hong Kong Government HKGAI-104B model.

Hantao’s research on AI safety, interpretability and ethics has been cited by Meta AI, OpenAI, CMU, and MILA, and has been featured in MIT Tech Review. Notably, Hantao authored Chapter 4 of the groundbreaking "AI Alignment: A Contemporary 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.

Research Experience

2023.07 – 2025.02 Peking University, Alignment and Interaction Lab | **Alignment of Large Language Models**
2024.06 – 2024.10 Machine Alignment & Theory Scholars Program | **Interpretability of Large Language Models**
2025.02 – Now Shanghai AI Lab | **Formal Verification for LLM Safety**

Honors and Awards

2025 Fall Peking University, Institute of Aritificial Intelligence, Dean’s Scholarship
2023 Fall Peking University Freshman Scholarship

Grants

2025 - 2026 PI, Beijing Natural Science Foundation Undergrad Research Grant | **Formal Verification for AI Alignment**

Publication

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. **[ACL 2025 Best Paper]** Jiaming Ji*, Kaile Wang*, Tianyi Qiu*, Boyuan Chen*, Jiayi Zhou, Changye Li, **Hantao Lou**, Yaodong Yang. Language Models Resist Alignment: Evidence From Data Compression.
4. **[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
5. **[ACM Computing Survey 2025]** 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 Contemporary Survey.
6. **[In Review]** 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.

Open-Source

- **PKU-Alignment/align-anything, Stars: 4.5K+**
  - A highly modular open-source multimodal alignment framework, supporting multiple architecutres and models
  - Repository creator and the core developer
- **Beneficial-AI-Foundation/NumpySpec**
  - A verified numpy-compatible library in Lean 4
  - Repository creator and the core developer
- **Beneficial-AI-Foundation/FloatSpec (Ongoing)**
  - A fully verified IEEE754 floating-point library in Lean 4, with code and specification
  - Repository creator and the core developer

Courses (Overall GPA: 3.45/4)

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

Professional Service

- **Talks**

2025 Spring	Guaranteed Safe AI Summit
2025 Spring	Wisemodel Community
2025 Spring	Peking University Interpretability Seminar

- **Review**

2025	ICLR 2026 Conference
2025	AAAI 2026 Conference (Main Track, AI Alignment Track)
2025	NeurIPS 2025 Conference
2025	ICCV 2025 Conference
2025	ICML 2025 Conference
2024	AAAI 2025 Conference AI Alignment Track
2024	ICLR 2025 Conference