# Hantao Lou

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

## 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), Formal Verification, and other potentially scalable methods. My research interests and past work revolve around the following question:

- *How to build reliable and scalable alignment methods for advanced, complex AI systems?*

## Education

2022–2026 **Yuanpei College, Peking University**.
B.E. Student in Artificial Intelligence, a member and the monitor of the Tong Class (an honorary pilot class in AI)

## Research Experience

2023 – **Undergrad Researcher at PAIR Lab: PKU Alignment and Interaction Research Lab**.
Currently working on 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, et al.

2024 **Aligner: Achieving efficient alignment through weak-to-strong correction**, *NeurIPS 2024 Oral (0.5%)*.
Jiaming Ji*, Boyuan Chen*, **Hantao Lou**, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Yaodong Yang

2024 **Stream Aligner: Efficient Sentence-Level Alignment via Distribution Induction**, *AAAI 2025 AI Alignment Track Poster*.
**Hantao Lou**, Jiaming Ji, Kaile Wang, Yaodong Yang

2024 **Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback**, *Arxiv Preprint*.
Jiaming Ji*, Jiayi Zhou*, **Hantao Lou***, Boyuan Chen*, Donghai Hong*, et al.

2025 **SAE-V: Interpreting Multimodal Models for Enhanced Alignment**, *Arxiv Preprint*.
**Hantao Lou***, Changye Li*, Jiaming Ji, Yaodong Yang

### Opensource Projects

2024 **Align-Anything**, *Github repo*.
An open-source framework for multimodal alignment, with 2.6k+ stars. 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 the MATS 2024 summer cohort supervised by Evan Hubinger.