Statement of Purpose

Hantao Lou *Ph.D. Applicant*

With AI systems racing ahead in capability, how can we make them safe and aligned with human intentions so they can become helpful tools and reliable agents? My research approaches this by rethinking supervision signals: all AI systems are ultimately evaluated by such signals, and most are optimized directly against them. Therefore, **how to provide and utilize supervision signals for scalable AI alignment?**

I have been attempting to answer these questions, as an undergraduate researcher at Peking University with Prof. **Yaodong Yang**, and later a research intern at Beneficial AI Foundation with Prof. **Max Tegmark**. My effort mainly focuses on: (1) utilizing supervision signals through improved **alignment algorithms**; (2) deriving supervision signals from model internals using **mechanistic interpretability**; and (3) constructing precisely specified supervision signals with **formal verification**. The following part will demonstrate my experience, expertise, and potential directions in research.

Alignment Algorithms

Most existing supervision signals are created by human or human-level intelligence; therefore, they are unstable and underspecified. At the beginning of my undergrad research, I worked for Prof. Yaodong Yang at Peking University, where I worked on designing alignment algorithms to increase the granularity of supervision signals and induce robust and efficient alignment.

Towards decomposing supervision signals, I co-authored **Aligner** [1], an alignment algorithm that decomposes supervision signals from binary preference to residuals between unpreferred and preferred responses, and learns these residuals with a smaller correction model to correct model outputs. **Aligner** topped Stanford Alpaca benchmark, successfully proved that detailed supervision signals could help alignment, and later **became an oral presentation in NeurIPS 2024**. Based on the observations from Aligner, I led **Stream Aligner** [2] paper, which increases the granularity of alignment during generation by decomposing supervision signals from response-level correction residuals to sentence-level ones, further demonstrating the potential of fine-grained supervision signals. This paper is recognized by **AAAI 2025 AI Alignment track**.

After working towards breaking down supervision signals in natural language space, I attempted to apply the same technique to multimodal scenarios, where higher complexity makes supervision signals easily underspecified. I co-led **Align Anything** [3], an end-to-end pipeline for aligning *all-modality* models using natural language critique as supervision signals, unifying different modalities under a more detailed universal language interface. I'm also the core contributor of the **Align-Anything toolkit** [4], a codebase supporting various alignment algorithms for *all-modality* models, providing both academic-level modularity and industry-level efficiency, and having **more than 4.5K stars on Github**. Working with Align-Anything, I'm **the first** to support the alignment of frontier *all-modality* models such as Chameleon and Janus.

Mechanistic Interpretability

Previous experience convinced me that external supervision signals, even when decomposed into rather high granularity, are unreliable. I began to use interpretability to extract supervision signals from model internals, without using any unreliable feedback source. Based on the intuition that alignment-related signals are compressed within the model, I started with information theory and discovered that the alignment-related information is less compressed and thus is easier to remove during unlearning. I co-authored **Language Model Resist Alignment** [5] paper, which later **won the best paper of ACL 2025**.

Limited to the difficulty of specifying safety and alignment in theory, I turned my eyes to **mechanistic interpretability**, which **investigates model internals to map features and their compositions into algorithms**. I learned about Sparse Autoencoder (SAE) and its potential for extracting features from language models during my participation in **MATS 6.0 program**. My effort on SAE alignment-related feature dynamics turned into a lesswrong blog [6]. Later, I attempted to select and utilize SAE features as supervision signals, and led the paper **SAE-V** [7], where I extended the current SAE paradigm to VLMs, built an unsupervised pipeline to locate alignment-related features, and used these features as supervision signals to assist alignment. This paper **got accepted as a poster in ICML 2025**.

Formal Verification

Interpretability provided a way to *find* supervision inside models, but also revealed how fragile those signals can be. To make supervision signals stable and precise, I turned to **formal verification**, which **integrates formal programming languages to build executable, provable, and scalable supervision signals**. During my internship with Prof. **Max Tegmark**, I worked with Lean4 to build automatic pipelines and formalize unverified codebases into specifications to provide supervision signals with mathematical proof. The production, **NumpySpec** [8] has become a part of the **Vericoding dataset & benchmark** [9], **the first large-scale, multi-lingual, multi-domain dataset and benchmark of LLM-generated verified code**. Moving on, I'm investigating IEEE754 float and formalizing it into **FloatSpec** [10], with much higher difficulties, much larger size, and a larger quantity of automatically generated verified signals. I'm also working on applying formal verification-based supervision signal to agent safety with Dr. **Jie Fu** at Shanghai AI Laboratory.

Future Paths

My past experiences and current research interests lie in the intersection of alignment algorithms, mechanistic interpretability, and formal verification. In the future, I'm excited to integrate my past research to build and utilize supervision signals for scalable AI alignment, to continue my research journey in academia, and to make greater contributions to the community.

References

[1] Jiaming Ji*, Boyuan Chen*, **Hantao Lou**, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[2] **Hantao Lou**, Jiaming Ji, Kaile Wang, and Yaodong Yang. Stream aligner: Efficient sentence-level alignment via distribution induction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27500–27508, 2025.

[3] Jiaming Ji*, Jiayi Zhou*, **Hantao Lou***, Boyuan Chen*, Donghai Hong*, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, et al. Align anything: Training all-modality models to follow instructions with language feedback. *In review*, 2024.

[4] Jiayi Zhou*, **Hantao Lou***, Boyuan Chen*, Donghai Hong*, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, et al. Align anything: A unified framework for all-modality alignment. `https://github.com/PKU-Alignment/align-anything`, 2024. GitHub repository.

[5] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Changye Li, **Hantao Lou**, Jiayi Zhou, Josef Dai, and Yaodong Yang. Language models resist alignment. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[6] **Hantao Lou** and Hubinger Evan. Automating llm auditing with developmental interpretability, 2024. URL `https://www.lesswrong.com/posts/uBZcQLmBPiiSyXsc9/automating-llm-auditing-with-developmental-interpretability`.

[7] **Hantao Lou***, Changye Li*, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment. In *Forty-second International Conference on Machine Learning*, 2025.

[8] **Hantao Lou***, Alok Singh, and Max Tegmark. Numpyspec: Formalizing the numpy library in lean 4. `https://github.com/Beneficial-AI-Foundation/NumpySpec`, 2024. GitHub repository.

[9] Sergiu Bursuc, Theodore Ehrenborg, Shaowei Lin, Lacramioara Astefanoaei, Ionel Emilian Chiosa, Jure Kukovec, Alok Singh, Oliver Butterley, Adem Bizid, Quinn Dougherty, et al. A benchmark for vericoding: formally verified program synthesis. *In review*, 2025.

[10] **Hantao Lou**, Alok Singh, and Max Tegmark. Floatspec: A formally verified ieee-754 floating-point specification in lean 4. `https://github.com/Beneficial-AI-Foundation/FloatSpec`, 2024. GitHub repository.