Yixiong Hao

yixiong_hao@outlook.com | (470)-457-3887 | Website |LinkedIn| GitHub | Google Scholar

EDUCATION

Georgia Institute of Technology — Atlanta, Georgia

Bachelor of Science in Computer Science & BSMS. **GPA: 4.0** — *Jan 2024 – May 2027*

- **Coursework:** DSA II, OOP, Objects & Design, Systems & Networks, Linear Algebra, Calculus III, AI, Deep Learning.
- **Key commitments:** AI Safety Initiative co-director, prev. Startup Exchange director of events.

PROFESSIONAL EXPERIENCE

Research Fellow | University of Chicago Existential Risk Lab — *Jun 2025 – Aug 2025*

- Prototyped a novel threat model involving **agentic misalignment in RLVR style post training** in collaboration with the Center for AI Safety - pre-disclosure work, to be published.
- Reproduced initial steps of RLLM's o1 level math performance by implementing **Dr GRPO in VeRL for distributed reinforcement learning** on DeepSeek-distilled Qwen-1.5B.

Research Engineer, Robotic Foundation Model Safety | People, AI and Robotics Lab @ GT — *Sep 2024 – Now*

- Leading early mechanistic interpretability work on low level actuators policies like Vision-Language-Action models.
- Designed **variational preference learning** methods to capture personal and multi-objective human preferences to improve the **alignment of large language models** through RLHF.
- Reproduced then-SOTA latent reasoning in LLMs via parallel meta token generation as 'thinking' (quietSTAR)

Quantitative Finance Research Intern | China Industrial Securities — *Jun 2024 – Aug 2024*

- Trained the **deep momentum** strategy in the Chinese stock market, achieving a **1.80 Sharpe ratio** with a long-short portfolio in back-testing. Used TensorFlow to build a deep neural network that classified stocks into expected return deciles based on momentum factors and implemented downstream statistical processing.
- **Optimized the proposed model to ~20% the original size** with grid search and no significant loss of performance.

RESEARCH PUBLICATIONS

Patterns and Mechanisms of Contrastive Activation Engineering — *ICLR 2025*

- Human-AI co-evolution, Bi-directional Alignment, and Building Trust in LLM workshops.
- **Contrastive activation engineering steers LLMs at inference time** towards desired behavior by manipulating hidden states; we provide critical data for future practitioners. I wrote the steering and evaluation pipeline and led a team of 4 to analyze CAE techniques across 10 features in Anthropic's MWE dataset on the Llama 3 family.

Interpreting large text-to-image diffusion models with dictionary learning — *CVPR 2025*

- Mechanistic interpretability for Vision workshop.
- Applied **sparse autoencoders** and **activation decomposition** to Flux diffusion models and extract interpretable features that can be used to steer generated images. Optimized hyper-parameters via grid search.

Language Models for Open-ended Wargames — *EMNLP 2025*

- Wordplay workshop.
- Synthesized 100+ studies to find significant opportunities for high player and adjudicator creativity wargames. I outline critical safety considerations and open research problems for using LLMs to simulate open ended serious wargames.

PROJECTS

BuzzBot – Georgia Tech virtual advisor — *Aug 2024 – Jan 2025*

- Designed and built comprehensive search engine to better answer students' questions for all things Georgia Tech.
- The search feature is implemented with **elastic search** encompassing lexical and semantic search. Natural language search and question answering are served by a LangChain **language model agent with RAG and tool use.**
- **Stack:** Docker, HuggingFace, transformers, LangChain, Pandas, Python.

Life-sized Quadruped robot — *2023*

- Planned, designed, and built a fully functional robot dog with 3 DoF limbs and custom 3D printed body.
- **Stack:** ESP32, Adafruit 16 channel controller, ToF sensors, C in Arduino IDE, inverse kinematics, trajectory planning.

TECHNICAL SKILLS

Programming Languages and tools: Python, C, Java, bash scripting, Git, Slurm, Docker

AI & ML: PyTorch, TensorFlow, HuggingFace, NumPy, Pandas, LangChain, SubmitIt, VeRL, deep learning, evolutionary algorithms, reinforcement learning, NLP, data visualization, transformers, weights & biases.

Full stack development: Django, Next.js, agile methodologies, CI/CD, SQL, API, HTML/CSS