

TIANYU HU

✉ hutianyu@stu.pku.edu.cn · 📞 (+86) 18810665771

🎓 EDUCATION

University of Wisconsin-Madison

Expected Spring 2027

Incoming PhD student in Computer Science, Advisor: Prof. Yiqiao Zhong

Peking University (PKU), Beijing, China

2021 – 2026

B.S. in Computer Science, AI Honors Program (Zhi Class)

👤 EXPERIENCE

Dartmouth TRIP lab Research Intern

May. 2025 - Present

Model Diagnosis, pruning, bayesian neural network, random matrix theory

- Investigated the interpretability of Neural Networks by analyzing the spectral properties of their weight matrices
- Established a causal link between prior distribution and the model's reasoning ability

Yale GersteinLab Research Intern (Remote)

Mar. 2024 - Dec. 2024

Agent (memory/planning), in context learning, prompt engineering, RAG

- ChemAgent: Self-updating Memories in Large Language Models Improves Chemical Reasoning

Algorithm Intern, LiblibAI Inc.

Jan. 2024 – Apr. 2024

Classification, fine-tuning, multimodal, text2image

- Image classification based on OpenCLIP and style description generated by GPT4
- High resolution stable diffusion

🎓 PUBLICATIONS

ChemAgent: Self-updating Memories in Large Language Models Improves Chemical Reasoning (ICLR 2025 Poster)

Xiangru Tang, Tianyu Hu*, MUYANG YE*, Yanjun Shao*, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, Mark Gerstein*

Bayes Pruning Priors for Soft Spectral Editing in Controlled Knowledge Conflicts (In Submission)

John J. Guerrerio, Tianyu Hu*, Yuehao Xu*, Zihang Liu, Dongxia Wu, Pu Ren, Hengrui Luo, Yaoqing Yang*

📌 SELECTED RESEARCH PROJECTS

Base/Chat model diffing

Model Diffing, Mechanistic interpretability, Sparse Autoencoder

- This project focusing on analyzing the difference between chat model and base model, using Anthropic's multi-stage model diffing strategy.
- Github Repo Link

Prompts that Optimize Difference Between Two Models

Model Diffing, Reinforcement Learning, AI security, Sleeper Agent

- Using RL method (PRBO) to find prompts that optimize the difference between two models' responses. This project mainly focuses on diffing Sleeper Agent and normal models.
- Short Presentation Link

⚙️ SKILLS

- Programming Languages: Python > C++ > C » Haskell
- Skilled in Pytorch, TransformerLens
- **Featured Coursework:** Continuous-Time Diffusion Processes in Machine Learning Theory; Efficient Computing of Deep Neural Networks; Convex Analysis and Optimization Methods; Cognitive Reasoning

♡ HONORS AND AWARDS

Merit Student in Peking University

Sep. 2022

HuaTai Scholarship

Sep. 2022

📖 MISCELLANEOUS

- Languages: English - Fluent (TOEFL 105 / speaking 23)
- **Personal Website:** thewaylost.github.io