YANG XU UNDERGRADUATE STUDENT

https://superposition09m.github.io/
yangxu09m@gmail.com

EDUCATION

Zhejiang University Hangzhou, China
Undergraduate Student 2022 - 2026 *(expected)*

- **Major**: Computer Science and Technology.
- **Minor Program**: Advanced Honor Class for Engineering Education(only 50 science&engineering students selected in each Grade), Chu Kochen Honors College, Zhejiang University.
- **GPA**: 3.98/4.3

RESEARCH INTERESTS

My research interests lie broadly in deep learning, particularly in **interpreting models** and **developing theoretical foundations** for them. My long-term goal is to understand the mechanisms of neural networks, transforming deep learning from a "black box" approach into a **rigorous science**. I am also interested in topics such as **alignment** and **trustworthiness** in AI systems, with the aim of ensuring that AI technologies are beneficial to society.

Currently, I am focusing on **mechanistic interpretability** in large language models (LLMs), a rapidly growing area of research that seeks to reverse-engineer the internal operations of neural networks to uncover how they function. Unlike learning theory, which emphasizes formal and mathematical frameworks, mechanistic interpretability takes an approach closer to the "physics" of LLMs. I believe that both formal theoretical approaches and mechanistic discoveries—viewed through a physical lens—are essential for achieving a comprehensive understanding of neural networks.

PUBLICATIONS AND PREPRINTS

1. Xu Cheng*, Lei Cheng*, Zhaoran Peng, **Yang Xu**, Tian Han, Quanshi Zhang. *Layerwise Change of Knowledge in Neural Networks*. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235:8038-8059, 2024.
2. Qihan Ren*, Junpeng Zhang*, **Yang Xu**, Yue Xin, Dongrui Liu, Quanshi Zhang. *Towards the Dynamics of a DNN Learning Symbolic Interactions*. Neural Information Processing Systems (NeurIPS), 2024.
   - Originally second author for theoretical contributions; authorship adjusted after merging experimental paper's first author.
3. **Yang Xu**, Yi Wang, Hao Wang. *Tracking the Feature Dynamics in LLM Training: A Mechanistic Study*. arXiv preprint arXiv:2412.17626, 2024.
4. **Yang Xu***, Xuanming Zhang*, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, Sharon Li. *Simulating and Understanding Deceptive Behaviors in Long-Horizon Interactions*. arXiv preprint arXiv:2510.03999, 2024.

INTERNSHIPS

Shanghai Jiao Tong University | Shanghai, China 2023.05-2024.09(remote)

- **Remote Research Intern** in the John Hopcroft Center for Computer Science, School of electronic information and electrical engineering.
- **Advisor**: Prof. Quanshi Zhang.
- **Study**: Interpretability of Neural Networks and Deep Learning Theory. **2 conference papers** were published.

Rutgers University | New Jersey, USA 2024.07-2025.02(remote since 2024.9)

- **Visiting Student and Research Intern** in the Department of Computer Science.
- **Advisor**: Prof. Hao Wang.
- **Study**: Mechanistic Interpretability Study of LLMs.

UW-Madison | Madison, USA 2025.03-Present(remote)

- **Remote Research Intern** in the School of Computer, Data & Information Sciences
- **Advisor**: Prof. Sharon Li.

Awards and Honors

- **2st Scholarship** in Zhejiang University.
- **Win 1st Prize for twice(2022, 2023)** in Zhejiang Division of National Mathematics Competition for College Students.

English Proficiency

TOEFL iBT: 102 (Listening: 30, Reading: 28, Speaking: 20, Writing: 24) *March 2024*

Activities: Member of ZJUFLA (Zhejiang University Foreign Language Association), English Corner Organizer for 2 semesters.

Sept. 2023 – June 2024

Academic Services

Reviewer for: *International Conference on Learning Representations (ICLR) 2025, North American Chapter of the Association for Computational Linguistics (NAACL) 2025.*