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:** 4.03/4.3

RESEARCH INTERESTS

My research interests lie broadly in deep learning, spanning both theory and applications. I am particularly interested in understanding the **mechanisms** behind deep learning models. Additionally, I am interested in **empirical methodology**, believing that intelligence can be understood through construction.

Regarding the working principles of deep learning models, I am currently focusing on **mechanistic interpretability** of LLMs—a rapidly growing research area that aims to reverse-engineer the internal operations of neural networks to uncover how they function. On the empirical side, I plan to explore aspects such as **RL** and **foundation models** in the future.

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.

INTERNSHIPS

Shanghai Jiao Tong University | Shanghai, China 2023.05-Present(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-Present(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.

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*.