

EDUCATION	<p>Zhejiang University Hangzhou, China <i>Undergraduate Student</i> 2022 - 2026 (<i>expected</i>)</p> <ul style="list-style-type: none"> • 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	<p>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.</p> <p>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.</p>
PUBLICATIONS AND PREPRINTS	<ol style="list-style-type: none"> 1. Xu Cheng*, Lei Cheng*, Zhaoran Peng, Yang Xu, Tian Han, Quanshi Zhang. <i>Layerwise Change of Knowledge in Neural Networks</i>. 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. <i>Towards the Dynamics of a DNN Learning Symbolic Interactions</i>. Neural Information Processing Systems (NeurIPS), 2024. <ul style="list-style-type: none"> • Originally second author for theoretical contributions; authorship adjusted after merging experimental paper's first author. 3. Yang Xu, Yi Wang, Hao Wang. <i>Tracking the Feature Dynamics in LLM Training: A Mechanistic Study</i>. arXiv preprint arXiv:2412.17626, 2024.
INTERNSHIPS	<p>Shanghai Jiao Tong University Shanghai, China 2023.05-Present(remote)</p> <ul style="list-style-type: none"> • 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. <p>Rutgers University New Jersey, USA 2024.07-Present(remote since 2024.9)</p> <ul style="list-style-type: none"> • Visiting Student and Research Intern in the Department of Computer Science. • Advisor: Prof. Hao Wang. • Study: Mechanistic Interpretability Study of LLMs.
AWARDS AND HONORS	<ul style="list-style-type: none"> • 2st Scholarship in Zhejiang University. • Win 1st Prize for twice(2022, 2023) in Zhejiang Division of National Mathematics Competition for College Students.

ENGLISH PROFICIENCY	<p>TOEFL iBT: 102 (Listening: 30, Reading: 28, Speaking: 20, Writing: 24) <i>March 2024</i></p> <p>Activities: Member of ZJUFLA (Zhejiang University Foreign Language Association), English Corner Organizer for 2 semesters.</p> <p style="text-align: right;"><i>Sept. 2023 – June 2024</i></p>
ACADEMIC SERVICES	<p>Reviewer for: <i>International Conference on Learning Representations (ICLR) 2025, North American Chapter of the Association for Computational Linguistics (NAACL) 2025.</i></p>