

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: 3.99/4.3
RESEARCH INTERESTS	<p>Recently, my research philosophy has undergone a significant evolution from Deconstruction to Construction. Previously, I focused on Interpretability and Trustworthy ML, driven by a desire to decipher the“physics” of neural networks and ensure their controllability.</p> <p>Now, my research interests have shifted toward new model architectures (e.g., Sparse/Linear Attention, DeltaNet) and new learning paradigms (e.g., Continual Learning, Test-time Learning), grounded in two core beliefs:</p> <p>(1) Understanding through construction, not just deconstruction. As Richard Feynman famously said: “What I cannot create, I do not understand.” (2) The next paradigm won’t emerge from reverse-engineering Transformers alone. Relying solely on that path risks trapping our understanding in the local minima of existing frameworks.</p> <p>While this represents a substantial shift in direction and I am still in an exploratory phase, I believe my previous research experience provides valuable priors for this new terrain. My long-term goal is to build models with true agency and efficiency.</p>
RESEARCH EXPERIENCES	<p>Summer Research Program Beijing, China 2025.07-2025.08</p> <ul style="list-style-type: none"> • Mentored by Jingyang Yuan (Researcher at Deepseek, the best paper author of ACL 2025) • Focus: Triton-based kernel optimization; Efficient attention mechanisms (linear and sparse attention) <p>UW-Madison Madison, USA 2025.03-2025.11(remote)</p> <ul style="list-style-type: none"> • Remote Research Intern in the School of Computer, Data & Information Sciences • Advisor: Prof. Sharon Li. • Focus: Evaluation of LLM Trustworthiness, Collaboration with Amazon <p>Rutgers University New Jersey, USA 2024.07-2025.02(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. • Focus: Mechanistic Interpretability Study of LLMs. <p>Shanghai Jiao Tong University Shanghai, China 2023.05-2024.09(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. • Focus: Interpretability of Neural Networks and Deep Learning Theory.

PROJECTS	Fused RoPE & Co-RoPE (Exploration)
	<ul style="list-style-type: none"> • Summary: Triton Implementation of Fused RoPE and Exploration of Contextual Improvement of RoPE(Co-RoPE) • Link: https://github.com/Superposition09m/RoPE-CoRoPE
	Stanford CS336: Language Modeling from Scratch - A1
	<ul style="list-style-type: none"> • Summary: A modular, from-scratch PyTorch implementation featuring modern LM architectural components: BPE tokenization, MHA, RMSNorm, SwiGLU activations, RoPE, and custom AdamW optimizer. Validated on TinyStories (loss 1.34) and OpenWebText (loss 3.78). • Link: https://github.com/Superposition09m/CS336-Mine(course website: https://stanford-cs336.github.io/spring2025/)
PUBLICATIONS AND PREPRINTS	<ol style="list-style-type: none"> 1. Yang Xu*, Xuanming Zhang*, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, Sharon Li. <i>Simulating and Understanding Deceptive Behaviors in Long-Horizon Interactions</i>. ICLR 2026. 2. Yang Xu, Yi Wang, Hao Wang. <i>Tracking the Feature Dynamics in LLM Training: A Mechanistic Study</i>. arXiv preprint arXiv:2412.17626, 2024. 3. 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. 4. 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.
AWARDS AND HONORS	<ul style="list-style-type: none"> • 2nd 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	Reviewer for: <i>International Conference on Learning Representations (ICLR) 2025, North American Chapter of the Association for Computational Linguistics (NAACL) 2025, ICLR 2026.</i>