

EDUCATION

Zhejiang University <i>Undergraduate Student</i>	Hangzhou, China 2022 - 2026 (<i>expected</i>)
<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

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.

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:

(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.

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.

RESEARCH EXPERIENCES

Summer Research Program Beijing, China	2025.07-2025.08
<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) 	
UW-Madison Madison, USA	2025.03-2025.11(remote)
<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 	
Rutgers University New Jersey, USA	2024.07-2025.02(remote since 2024.9)
<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. 	
Shanghai Jiao Tong University Shanghai, China	2023.05-2024.09(remote)
<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)

- **Summary:** Triton Implementation of Fused RoPE and Exploration of Contextual Improvement of RoPE(Co-RoPE)
- **Link:** <https://github.com/Superposition09m/RoPE-CoRoPE>

PUBLICATIONS
AND
PREPRINTS

1. 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, 2025. NeurIPS 2025 Workshop @ResponsibleFM Accepted. ICLR 2026 Under Review.(8,6,4,4)
2. Yang Xu, Yi Wang, Hao Wang. *Tracking the Feature Dynamics in LLM Training: A Mechanistic Study.* arXiv preprint arXiv:2412.17626, 2024.
3. 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.
4. 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.

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, ICLR 2026.*