Yilin Wang

+86 13148035071 | yilinw376@gmail.com | GitHub | Personal Website

Education

2023.09-Present (Now fourth-year student) **Xi'an Jiaotong University**-School of Computer Science (Honors Science Program) Xi'an China

- **Academic performance:** grade point ranking **8/45** (first year), **12/40** (second year), **11/37** (over three years); credit-weighted score over six semesters: **92.00**; GPA: **3.95/4.3**.
- **Main Couresework:** Computer Programming 100, Discrete Mathematics 97, Probability Statistics and Random Processes 100, Fundamentals of Optimization Methods 98, Linear Algebra and Analytical Geometry 98, Machine Learning 98, Mathematical Physics Equations 97, operating system 95, etc.
- **English:** IELTS: **Overall 7.5** (Listening: 8.0, Reading: 8.0, Writing: 6.5, Speaking: 6.5), CET-4: 622, CET-6: 536

Research

Co-lead the undergraduate research group of Professor Minan Luo (LUD) at Xi'an Jiaotong University with research outputs:

2024.05-2025.02 **Continuously Steering LLMs Sensitivity to Contextual Knowledge with Proxy Models** **EMNLP 2025 Main (Oral)**

- TL;DR: We propose the CSKS framework to address LLMs' difficulty in appropriately adjusting their belief in parametric knowledge when it conflicts with retrieved context in RAG. CSKS leverages the output distribution difference between two small models to guide the large model's predictions. Experiments show that CSKS significantly outperforms mainstream baselines on knowledge conflict handling. Code is open-sourced [Link].
- Contribution: **FIRST AUTHOR,** independently responsible for project design, coding implementation, and paper writing.

2025.03-2025.12 **Why Knowing Both Hops Is Not Enough: Understanding Two-Hop Generalization in Language Models** **EMNLP 2026 Fingdings**

- TL;DR: We develop a symbolic intervention framework and applied mechanistic interpretability to show that LLMs fail out-of-distribution in two-hop reasoning due to high-level functional mismatch, and proposed a recurrent training scheme that aligns layer-wise representations to achieve robust multi-hop reasoning across distributions. Code is open-sourced [Link].
- Contribution: Served as both **CO-FIRST AUTHOR** and **MENTOR.** I proposed the methodology, built the code framework, and led paper writing, while mentoring a collaborator to complete the remaining tasks.

Advised by Prof. Wen Muning at SAI, Shanghai Jiao Tong University; research on designing RL algorithms to enhance LLM reasoning.

2026.01-2026.03 **Think About LLM as Pi**

- TL;DR: We model the forward inference process of LLMs as an MDP and aim to exploit internal model signals for algorithmic optimization.
- Contribution: Research intern; **PROPOSED the project's methodology**.

Intern at Westlake NLP group; Co-supervised by Prof. Yue Zhang**,** Dr. Michalis (Cambridge fellow) and Zhizhang Fu (PhD).

2026.06-Present **Latent LLM-as-a-Judge**

- TL;DR: We aim to build a more rubust LLM judge from a latent perspective. Specifically, we investigate how evaluation criteria such as toxicity are encoded in internal representations, extract these latent features using techniques such as SAEs, and study whether these features can be used to understand and improve judge's behavior.
- Contribution: I **LEAD the work** and have done thorough preliminary exploratory work. You can look through my work here.

Projects

2025.08-2026.02 Completed Stanford CS336 and most course assignments independently.

- Contribution: Independently implemented most assignments with heavy, engineering-focused code, gaining solid practical and theoretical understanding of LLMs; code is open-sourced [Link].

2026.01-2026.03 To learn the popular RL framework verl, I started developing a lightweight verl-style framework nanoverl

- Contribution: Independently developing the project; a complete framework is now in place, supporting FSDP-based distributed RL training. The code is open-sourced [Link].

Honors

- First Prize (National Level) and Innovation & Entrepreneurship Award, 18th National College Student Information Security Competition
- Huawei Scholarship for Undergraduate Students
- First Prize, National College Student Artificial Intelligence Security Competition
- Nomination Award, Cybersecurity Innovation and Entrepreneurship Competition
- Third-class Scholarship (University-level)
- Campus Culture and Spiritual Civilization Academic and Scientific Research Award (University-level)

FYI

- In my spare time, I enjoy playing badminton and volunteering, and served as Head of the Empty-Nest Outreach Department in the XJTU Yangfan Student Society during my first three years.
- Now, I co-lead LUD lab, an undergraduate research group @ XJTU