# Wenjie (Andy) Li

Curriculum Vitæ

ShanghaiTech University
andyisok.00@gmail.com
For my **latest updates**, please visit my website at andyonwaves.top

Google Scholar: 9UMWTqMAAAAJ
LinkedIn: andythesailor
Blog: andyonwaves
GitHub: Ruby-a07

Hi, I'm Wenjie :)
I conduct research on building AI that is not only capable but reliably aligned with intended use - which I believe to be crucial for real-world deployment. To achieve this, I approach the problem from both the data side (what the model learns from) and the model side (how it learns and internally operates). Two projects I led on this topic embody this philosophy: Δ-Influence tackles data integrity against poisoning attacks by tracing model failures back to root-cause training samples through an observed phenomenon we term *influence collapse*, enabling targeted correction without prior attack knowledge. NeuronLLM enables precise behavioral control by revealing *functional antagonism* in LLMs - task performance is jointly determined by opposing "good" and "bad" neurons through their coordinated interaction. This discovery opens new possibilities for targeted model steering, such as suppressing harmful capabilities or enhancing task-specific performance.

## Education

**ShanghaiTech University*** — Expected 2026
Master of Science in Computer Science — Shanghai
Three-year program (extended to four years due to full-time research position)
Relevant Courses: Deep Learning, Machine Learning Algorithms, Algorithmic Game Theory
Major GPA: 3.57 / 4.0
Honors: Priority Admission with Interview Waiver, University Scholarship

**North China Electric Power University*** — 2018-2022
Bachelor of Science in Computer Science — Beijing
Relevant Courses: Linear Algebra, Introduction to Artificial Intelligence, Java Programming, Principles of Circuits
Major GPA: 3.96 / 4.0
Honors: University Scholarship, Merit Student

## Professional Experience

**Machine Learning Research Engineer (Full-Time)** — 2025
MaLA Lab, Singapore Management University — Singapore
Research leave from master's program to lead NeuronLLM (resulted in ICLR 2026 submission)

## Publications and Manuscripts

[A] **W. Li**, J. Li, C. S. de Witt, A. Prabhu, and A. Sanyal. *Delta-influence: Unlearning poisons via influence functions*. The 2nd Workshop on Attributing Model Behavior at Scale (**ATTRIB @ NeurIPS**), 2024, arXiv preprint arXiv:2411.13731.

[B] **W. Li**, G. Pang, D. Gao, and D. Lo. *NeuronLLM: Identifying Good and Bad Neurons for Task-Level Controllable LLMs*, In submission to ICLR 2026.

[C] A. Costarelli, M. Allen, R. Hauksson, G. Sodunke, S. Hariharan, C. Cheng, **W. Li**, J. Clymer, and A. Yadav. *GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents*. The Language Gamification Workshop (**LanGame @ NeurIPS**), 2024, arXiv preprint arXiv:2406.06613.

[D] S. Jiang, Y. Hu, **W. Li**, and P. Zeng. *DeepFRC: An End-to-End Deep Learning Model for Functional Registration and Classification*, In submission to ICLR 2026, arXiv preprint arXiv:2501.18116.

## Selected Distinctions

**Stanford Existential Risks Initiative Fellowship** — Stanford University
**Berkeley AI Safety Initiative Scholarship** — UC Berkeley
**Machine Learning Safety Scholarship (MLSS)** — Center for AI Safety
**Horizon Fellowship** — Effective Altruism Hong Kong

## Service

**Reviewer** — Safe Generative AI Workshop (SafeGenAI) @ NeurIPS 2024

*__Double First-Class__, a status granted by the Chinese government representing the top 5% of higher education institutions in China.