Weihan Fei

feiwh2400@mail.ustc.edu.cn | Hefei, China | wesleyfei1.github.io | github.com/wesleyfei1

Education

B.S. in Artificial Intelligence, University of Science and Technology of China (USTC) Expected 2028
School of the Gifted Young Hefei, China
GPA: 3.87/4.3 Rank: 14/103

Selected Core Coursework: Discrete Mathematics (100), Probability and Mathematical Statistics (92), Linear Algebra B(1) (90), Data Structures A (90).
Selected Advanced Coursework (Self-Directed): Stanford CS229 (Machine Learning), CS230 (Deep Learning), CS224n (Natural Language Processing), MIT 6.S184 (Generative AI Foundations), Berkeley CS61B (Data Structures).
Languages: Mandarin (native), English (TOEFL 89/120).

Research Interests

Generative recommendation, long-context memory for LLM agents, and efficient reasoning/retrieval mechanisms, especially methods that are mathematically intuitive, structurally concise, and grounded in empirical behavioral findings.Interested in verifiable compiled languages for AI system development and the intersection of AI and software engineering.

Research Experience

Research Intern, Alpha-Lab, USTC Nov 2025 – Present
Advisor: Prof. An Zhang
Adaptive-Thinking for Generative Recommendation
Studied a consistent behavioral gap between "think" and "not-think" inference in generative recommendation, especially their differences in predictive entropy, popularity bias, and downstream recommendation quality. Based on these findings, developed an adaptive-thinking framework that selectively invokes reasoning only when uncertainty is high, aiming to balance effectiveness and inference cost. Work in preparation for submission to NeurIPS 2026.

Research Intern, USTC Mar 2026 – Present
Advisor: Prof. Xiang Wang
QQMem: Hierarchical Query-to-Query Retrieval for Long-Context Agent Memory
Developing a memory retrieval framework for LLM agents motivated by the observation that direct episode retrieval is often semantically brittle in long-context settings. QQMem replaces episode-level matching with query-space alignment, using structured intermediate queries as semantic anchors to support more stable retrieval and grounded generation. Work in preparation for submission to NeurIPS 2026.

Honors & Awards

- **Undergraduate Research Opportunities Program (UROP)**, Research on Generative Recommendation Systems based on Large Language Models, advised by Prof. An Zhang Dec 2025 – Present
- **First Prize (Provincial Level)**, The 17th Chinese Mathematics Competitions (Non-Math Major, top 20) Oct 2025
- **Silver Prize**, Outstanding Undergraduates Scholarship Sept 2025
- **Bronze Prize**, Outstanding Student Scholarship Dec 2024

Skills

Research: Literature review, experimental design, empirical analysis, and end-to-end implementation for machine learning research.
Programming: Python, C, Java, Shell/Bash.
ML Frameworks: PyTorch, Hugging Face, vLLM, verl.
Model Training: Supervised fine-tuning (SFT), preference optimization (DPO/PPO), parameter-efficient tuning (LoRA), inference-time prompting and reasoning.
Tools: Linux, Git, Conda, tmux, nvitop, Weights & Biases.
Engineering: Rapid prototyping with LLM-assisted coding workflows(vibe coding) and modern web tooling.