

Wanli Yang

Institute of Computing Technology, Chinese Academy of Sciences

✉ yangyywl@gmail.com • 🏠 Website • 🎓 Scholar • 🐙 GitHub

Education

Institute of Computing Technology, Chinese Academy of Sciences

Ph.D. Candidate in Cyberspace Security

2024–Present

Advisors: Prof. Fei Sun and Prof. Xinran Liu

College of Software, Nankai University

B.E. in Software Engineering

2020–2024

Rank: 5/140; CET-4: 616; CET-6: 614

Research Interests

I study how AI systems can **improve reliably**: safely updating parametric knowledge through **model editing**, eliciting **latent knowledge** through reinforcement learning, and more broadly investigating whether and how **autonomous** agents can improve the **AI R&D** process itself.

Research Contributions

- **Knowledge Updating.** Led a **four-paper program**: uncovered and explained editing-induced model collapse; exposed systematic overestimation in prevailing protocols and rebuilt evaluation around practical deployment; then developed a simple, efficient fine-tuning method, the first to sustain 100K edits and scale to 72B models.
- **Knowledge Recall.** Extended reliable model improvement from writing new facts to eliciting what models already know: reinforcement learning yields about 27% average relative gains in direct factual recall, primarily by redistributing probability mass toward existing knowledge rather than acquiring new facts.
- **Toward Recursive Self-Improvement.** Broadened the agenda from human-led model improvement to autonomous AI R&D: developed a process-aware evaluation beyond final scores that reveals how harness design shapes reliability, when experience helps or misleads, how rarely genuine novelty emerges, and where agent systems need improvement.

Research Internships

Recursive Self-Improvement, Meituan Inc.

Research Intern

2025–Present

Mentors: Dr. Hongyu Zang & Dr. Wenjie Shi

Research topics: RL for parametric knowledge recall; evaluation of autonomous AI R&D agents

Baidu Search, Baidu Inc.

Research Intern

2023–2025

Mentor: Dr. Xinyu Ma

Research topic: reliable model editing, spanning collapse analysis, practical evaluation, and scalable methods

Preprints and Publications

Beyond Final Scores: A Systematic Evaluation of Agents for Long-Horizon AI Research and Development

- Yiwei Li*, Wanli Yang*, Hexiang Tan*, Xiangzhou Huang, Zhengyu Chen, Ziran Li, Borun Chen, Shanglin Lei, Huaisheng Zhu, Hao Tian, Fei Sun, Xunliang Cai, Jingang Wang.

* Equal contribution.

- **Preprint, 2026**

Beyond Reasoning: Reinforcement Learning Unlocks Parametric Knowledge in LLMs

- Wanli Yang, Hongyu Zang, Junwei Zhang, Wenjie Shi, Du Su, Jingang Wang, Xueqi Cheng, Fei Sun.

- **Preprint, 2026**

BaseCal: Unsupervised Confidence Calibration via Base Model Signals

- Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, Xueqi Cheng.
- ACL 2026, Main Conference

Fine-tuning Done Right in Model Editing

- Wanli Yang, Rui Tang, Hongyu Zang, Du Su, Qi Cao, Jingang Wang, Huawei Shen, Xueqi Cheng, Fei Sun.
- ICLR 2026

The Mirage of Model Editing: Revisiting Evaluation in the WILD

- Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, Xueqi Cheng.
- ACL 2025, Main Conference **Best Paper Award, KnowFM Workshop @ ACL 2025**

The Fall of ROME: Understanding the Collapse of LLMs in Model Editing

- Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, Huawei Shen.
- EMNLP 2024, Findings

Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?

- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, Xueqi Cheng.
- ACL 2024, Main Conference

The Butterfly Effect of Model Editing: Few Edits Can Trigger Large Language Models Collapse

- Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, Xueqi Cheng.
- ACL 2024, Findings

Honors and Awards

Outstanding Graduate (Top 5%), Nankai University	2024
College Scholarship, University of Chinese Academy of Sciences	2023
Academic Excellence Scholarship (Top 5%), Nankai University	2022 & 2023
Arts and Sports Scholarship (Top 2%), Nankai University	2021 & 2022
Innovation Scholarship (Top 3%), Nankai University	2021

Technical Skills

Technical: Python; C/C++; Java; SQL; PyTorch; Git; $\text{\LaTeX}$; Linux

Hobbies

Sports: Fitness, swimming, cycling, and dragon boat racing (former paddler of Nankai University Dragon Boat Team)

Reading: Works by Haruki Murakami