

Ji Zeng

zeng-ji@sjtu.edu.cn | ji-z.net | github.com/stargazerZJ | [Google Scholar](#)

EDUCATION

Shanghai Jiao Tong University(SJTU)

B.S. in Science, Zhiyuan College, ACM Honors Class

Sep. 2023 – Exp. June 2027

GPA Rank: 4/27

RESEARCH EXPERIENCE

Research Assistant

CX Research Group, Language Technologies Institute, Carnegie Mellon University

May 2026 – Present

Advisor: Chenyan Xiong

Research Assistant

Generative Artificial Intelligence Research Lab, SJTU, SH, China

July 2025 – May 2026

Advisor: Pengfei Liu

SELECTED PUBLICATIONS

daVinci-Dev: Agent-native Mid-training for Software Engineering

ICML 2026 Oral

Ji Zeng, Dayuan Fu, Tiantian Mi, Yumin Zhuang, Yaxing Huang, Xuefeng Li, Lyumanshan Ye, Muhang Xie, Qishuo Hua, Zhen Huang, Mohan Jiang, Hanning Wang, Jifan Lin, Yang Xiao, Jie Sun, Yunze Wu, Pengfei Liu

arXiv | Hugging Face

- Proposed the principle of agent-native data synthesis for software engineering mid-training
- Engineered a high-throughput distributed GoLang & Rust system to curate the 70B-token dataset
- Constructed a 10,000-core k8s cluster for parallel SWE benchmarking with stable cross-border connectivity
- Gained experience in 1000+ GPU LLM training and efficient evaluation of code LLMs
- Led coordinated data collection, model training, and evaluation; improved Qwen 2.5 32B/72B SWE-Bench scores to **56.1%/58.2%**

PROJECTS

3-9x Throughput for HF Tokenizers | [Issue Link](#) | [Production Implementation](#)

- 7-9x tokenization throughput on server CPU in realworld training dataset preparation workflow
- 3x on uniformed length random strings
- Sharding the tokenizer instance to improve cache locality
- Use a worker pool optimized for variable length input

Partial Rollout for VeRL | [Pull Request Link](#)

- 2x speed gain in long output, few GPU scenario
- Adopted by community users

Compiler | [Repo Link](#)

- Compiles Mx (a simplified C-Java mixture) into RISC-V Assembly
- 4k lines of concise and efficient Python & C++ code
- 2nd/27 in generated code execution efficiency benchmark with careful backend tuning

TEACHING EXPERIENCE

Teaching Assistant	Principles and Practice of Computer Algorithms	Summer 2025
Teaching Assistant	Data Structures	Spring 2025
Teaching Assistant	Programming Design	Fall 2024

AWARDS AND SCHOLARSHIPS

First-Class Cybersecurity Scholarship	2024
<i>Awarded for discovering security vulnerability in campus smart door system</i>	
Zhiyuan Honors Scholarship	2024, 2025
<i>Award for top students majoring in science</i>	