

Shen Fu

No. 100, Fuxing Rd., High-Tech District, Hefei, Anhui Province, China 230031

 sh.fu@outlook.com  Fr4nk1inCs

RESEARCH INTERESTS

LLM inference optimization, System for MoE.

RESEARCH PROJECTS

Parallelism Planning for MoE Inference with Dynamic Top-K Routing ADSL, USTC

Core Member

Mar 2025–Aug 2025

- An inference framework for dynamic top-k routing MoE models, which automatically plans parallelism strategies to maximize throughput on prefill-dominated workloads.
 - Participated in the implementation of the model profiler, adoption of dynamic top-k routing, pipeline parallelism enhancements, and the design of the parallelism planner.

PUBLICATIONS

Jin, Z., **Fu, S.**, Tang, C., Bai, Y., Wang, S., Zhu, J., Fang, C., Gong, P., & Li, C. (2026). SMIDT: High-Performance Inference Framework for MoE Models with Dynamic Top-K Routing. *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*.

EDUCATION

University of Science and Technology of China

Hefei, Anhui

M.E. in Computer Science and Technology

Sep 2024–Present

- Advisor: Prof. Cheng Li
 - GPA: 4.13/4.30

University of Science and Technology of China

Hefei, Anhui

B.E. in Computer Science and Technology

Sep 2020–Jun 2024

- School of the Gifted Young
 - GPA: 3.92/4.30, Rank: top 8%

HONORS & SCHOLARSHIPS

- Qiangwei “Yuanzhi” Scholarship (**Top 3%**) Oct 2023, USTC
 - Jianghuai & NIO Automobile Scholarship Jan 2023, USTC
 - Cheng Linyi Scholarship Jan 2022, USTC
 - Outstanding Freshman Scholarship, Grade 2 Sep 2021, USTC

MISCELLANEOUS

SERVICE

- USENIX ATC '25 Artifact Evaluation Committee

TEACHING

- T.A. for *Compiler Principles and Techniques* (Instructor: Prof. Cheng Li) 2023 Autumn, USTC

OPEN SOURCE CONTRIBUTIONS

- [sgl-project/sglang] feat: add dp attention support for Qwen 2/3 MoE models (#6121)

SKILLS

- **Languages:** Mandarin Chinese (Native), English (Fluent)
- **Programming:** Python, C/C++, Lua, Shell Script
- **Frameworks:** PyTorch, vLLM, SGLang