Siqi Zhu

Homepage: https://zhusq20.github.io/ email: zhusq20@gmail.com

EDUCATION

Tsinghua University Beijing, China
Economics and Finance(B.Econ.)+ Computer Science and Technology(B.Eng.), GPA:3.80/4 graduate:Jun.2025

REASEARCH EXPERIENCES

[6] **Efficient LLM Scheduling by Learning to Rank**.
Yichao Fu, **Siqi Zhu**, Runlong Su, Aurick Qiao, Ion Stoica, Hao Zhang.
NeurIPS 2024 Poster. [arxiv]
Advised by Prof. Hao Zhang, UC San Diego Feb.2024 - May.2024

- Develop a LLM iteration-level scheduling policy based on predicted request generation length rankings, achieve 1/2.8 latency in chatbot serving and 6.5x throughput in synthetic data generation.

[5] **Efficiently Serving LLM Reasoning Programs Using Certitude**.
Yichao Fu, Junda Chen, **Siqi Zhu**, Zheyu Fu, Zhongdongming Dai, Aurick Qiao, Hao Zhang.
Under Review, OSDI 2025.
Advised by Prof. Hao Zhang, UC San Diego Sept.2024 - Dec.2024

- Develope a serving system for reasoning algorithms such as MCTS and O1-like reasoning. Reduce compute by up to 50% in batch processing and sustain 3.3× higher query rates or 4.7× tighter latency SLOs in online serving.

[4] **Cost-Effective Synthetic Data Generation for Post-Training using QWICK**.
Yichao Fu*, **Siqi Zhu***, Junda Chen, Hao Zhang.
Under Review, ICLR 2025.
Advised by Prof. Hao Zhang, UC San Diego Jun.2024 - Sept.2024

- Utilize budget-constrained bandits to develop a synthetic data generation algorithm for LLM rejection fine-tuning. Reduce cost by up to 50% while maintaining data quality.

[3] **mTuner: Accelerating Parameter-Efficient Fine-Tuning on Multi-GPU Servers with Elastic Tensor**.
Kezhao Huang, **Siqi Zhu**, Mingshu Zhai, Liyan Zheng, Kinman Lei, Yuyang Jin, Jidong Zhai.
Manuscript.
Advised by Prof. Jidong Zhai, Tsinghua University Sept.2023 - Jan.2024

- Develop an efficient and scalable system for parameter-efficient fine-tuning, achieving a 1.51× throughput improvement over state-of-the-art systems and enabling 70B LLM fine-tuning on 8-GPU server.

[2] **Directed Independent Research**
Advised by Prof. Xuehai Qian, Tsinghua University Jun.2024 - Present

- Evaluate the performance of various LLM inference systems and enhance the efficiency of serving multiple models on clusters by optimizing parallelism, scheduling strategies and resource utilization.

[1] **Stable Prediction via Random Partitioned Variable Decoupling**.
Yue He, Zimu Wang, **Siqi Zhu**, Renzhe Xu, Wenchao Zou, Peng Cui.
Under Review, AAAI 2025.
Advised by Prof. Peng Cui, Tsinghua University Jul.2023 - Oct.2023

- Develop an algorithm that disentangles features through random permutation in the latent space, enhancing model robustness to covariate shift under causal assumptions.

PROJECTS

[2] High Performance Computing Course Project: GPU-accelerated SpMM, Floyd-Warshall Algorithm. MPI Odd Even Sort, Ring Allreduce. [github]
[1] Software Engineering Course Project: Capybara Chat, an instant messaging system. 3k LoC in JavaScript and Python. [github]

INDUSTRY EXPERIENCES

Zhipu AI — Beijing, China
Engineer Intern — Jan.2024 - May.2024

- **LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs**.
  Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, **Siqi Zhu**, Lei Hou, Yuxiao Dong, Jie Tang, Juanzi Li.
  Under Review, ICLR 2025. [arxiv]

AWARDS

Research Travel Grant, Tsinghua University — Apr.2024
Academic Excellence Scholarship, Tsinghua University — Oct.2021, 2022
First Prize, Guangdong Provincial, National Olympiad in Informatics in Provinces — 2019

SKILLS

Programming Language: C++, Python, CUDA, Triton
LLM Frameworks: vLLM, SGLang, Megatron, NeMo
Language: English (TOEFL 108), Mandarin