

# Shiyi Cao

☎ +1 (650) - 304 7475 • ✉ shicao@berkeley.edu • 🌐 shiyicao.com

## Educations

---

### UC Berkeley

- Ph.D. in Computer Science.
- Advisors: Prof. [Ion Stoica](#) and Prof. [Joseph E. Gonzalez](#).

Berkeley, California

Sept. 2023 - Present

### ETH Zürich

- M.S. in Computer Science.
- Advisor: Prof. [Torsten Hoefler](#).

Zürich, Switzerland

Sept. 2020 - June. 2023

### Shanghai Jiao Tong University

- B.S. in Computer Science.

Shanghai, China

Sept. 2016 - June. 2020

## Research Interests

---

- My general research interests lie in the fields of computer systems and high-performance computing, with a focus on understanding and accelerating emerging applications on heterogeneous systems. I am currently working on building efficient LLM inference/training systems/algorithms.

## Selected Publications

---

### MoE-Lightning: High-Throughput MoE Inference on Memory-constrained GPUs.

- [Shiyi Cao](#), Shu Liu, Tyler Griggs, Peter Schafhalter, Xiaoxuan Liu, Ying Sheng, Joseph E. Gonzalez, Matei Zaharia, Ion Stoica.
- *ASPLOS*, 2025.

### GraphPipe: Improving the Performance and Scalability of DNN Training with Graph Pipeline Parallelism

- Byungsoo Jeon\*, Mengdi Wu\*, [Shiyi Cao\\*](#), Sunghyun Kim\*, Sunghyun Park, Neeraj Aggarwal, Colin Unger, Daiyaan Arfeen, Peiyuan Liao, Xupeng Miao, Mohammad Alizadeh, Gregory R. Ganger, Tianqi Chen, Zhihao Jia.
- *ASPLOS*, 2025.

### Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models

- Ling Yang, Zhaochen Yu, Tianjun Zhang, [Shiyi Cao](#), Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, Bin Cui.
- *Neural Information Processing Systems (NeurIPS)*, 2024. [[Spotlight](#)]

### SGLang: Efficient Execution of Structured Language Model Programs

- Lianmin Zheng\*, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, [Shiyi Cao](#), Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, Ying Sheng\*
- *Neural Information Processing Systems (NeurIPS)*, 2024.

### Atlas: Hierarchical Partitioning for Quantum Circuit Simulation on GPUs.

- Mingkuan Xu, [Shiyi Cao](#), Xupeng Miao, Umut A. Acar, and Zhihao Jia.
- *SC*, 2024.

### Fairness in Serving Large Language Models.

- Ying Sheng, [Shiyi Cao](#), Dacheng Li, Banghua Zhu, Zhuohan Li, Danyang Zhuo, Joseph E Gonzalez, Ion Stoica.
- *OSDI*, 2024.

### S-LoRA: Serving Thousands of Concurrent LoRA Adapters

- Ying Sheng\*, [Shiyi Cao\\*](#), Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica.
- *MLSys*, 2024.

### Accelerating Data Serialization/Deserialization Protocols with In-Network Compute.

- [Shiyi Cao](#), Salvatore Di Girolamo, and Torsten Hoefler.
- *Workshop on Exascale MPI, ExaMPI@SC*, 2022.

## AdaM: An adaptive fine-grained scheme for distributed metadata management

- Shiyi Cao, Yuanning Gao, Xiaofeng Gao, and Guihai Chen.
- *International Conference on Parallel Processing (ICPP)*, 2019.

## Industry Experiences

---

### NVIDIA Research, Efficient AI Team

Managers: Prof. [Song Han](#)

- **Research Intern.** Efficient Vision-Language Model Fine-tuning.

Santa Clara, California

*May. 2024 - present*