Shiyi Cao

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

Educations

UC Berkeley — **Berkeley, California**
Sept. 2023 - Present
- Ph.D. in Computer Science.
- Advisors: Prof. Ion Stoica and Prof. Joseph E. Gonzalez.

ETH Zürich — **Zürich, Switzerland**
Sept. 2020 - June. 2023
- M.S. in Computer Science.
- Advisor: Prof. Torsten Hoefler.

Shanghai Jiao Tong University — **Shanghai, China**
Sept. 2016 - June. 2020
- B.S. in Computer Science.

Research Interests

- My research lies at the intersection of computer systems and AI. I develop efficient and scalable infrastructure for LLM inference and training, while advancing self-evolving agents that can tackle real-world reasoning and engineering tasks. Broadly, I aim to close the loop between systems and models – building infrastructure that trains agents efficiently, and agents that can in turn analyze and optimize the systems they run on.

Open Source Projects

SkyRL: A Modular Full-stack RL Library for LLMs.
- Co-lead the project.
- *https://github.com/NovaSky-AI/SkyRL*

Selected Publications

S*: Test Time Scaling for Code Generation.
- Dacheng Li*, **Shiyi Cao***, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, Ion Stoica.
- *EMNLP Findings*, 2025.

Language Models Can Easily Learn to Reason from Demonstrations.
- Dacheng Li*, **Shiyi Cao***, Tyler Griggs*, Shu Liu*, Xiangxi Mo, Eric Tang, Sumanth, Hedge, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, Ion Stoica.
- *EMNLP Findings*, 2025.

Nvila: Efficient frontier visual language models.
- Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, **Shiyi Cao**, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, Yao Lu.
- *CVPR*, 2025.

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.

Selected Preprint

SkyRL-Agent: Efficient RL Training for Multi-turn LLM Agent .

- **Shiyi Cao***, Dacheng Li*, Fangzhou Zhao, Shuo Yuan, Sumanth R. Hegde, Connor Chen, Charlie Ruan, Tyler Griggs, Shu Liu, Eric Tang, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, Ion Stoica.
- Arxiv, 2025.

Industry Experiences

NVIDIA Research, Efficient AI Team — **Santa Clara, California**
Managers: Prof. Song Han — *May. 2024 - Dec. 2024*

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