Qizheng Zhang

Contact

Email: qizhengz@stanford.edu
Website: alex-q-z.github.io
GScholar: scholar.google.com/citations?user=xnt6d5oAAAAJ

Interests

Ph.D. candidate in Computer Science at Stanford University. My primary research interests lie in systems and algorithms for post-training and inference-time adaptation of large language models (LLMs) and agents. My recent work focus on building self-improving AI systems that close the loop between model inference, reflection, and systems optimization. I design frameworks like *ACE (ICLR 2026)*, *FlowRL (ICLR 2026)*, *APC (NeurIPS 2025)* and *Caravan (OSDI 2024)* that integrate scalable systems design with learning-based adaptation, enabling large models to reason efficiently and evolve autonomously.

Education

Stanford University
Ph.D. in Computer Science — Sep. 2022 - Dec. 2026 (expected)
M.S. in Computer Science — Sep. 2022 - Jan. 2025
Advisor: Prof. Kunle Olukotun
Research Area: Language Models, Machine Learning Systems

University of Chicago — Sep. 2018 - Jun. 2022
B.S. in Computer Science (with Honors)
B.S. in Mathematics, B.A. in Statistics

Selected Publications

1. *Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models*
Qizheng Zhang*, Changran Hu*, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, Kunle Olukotun
International Conference on Learning Representations (ICLR), 2026

2. *Agentic Plan Caching: Test-Time Memory for Fast and Cost-Efficient LLM Agents*
Qizheng Zhang, Michael Wornow, Kunle Olukotun
Conference on Neural Information Processing Systems (NeurIPS), 2025

3. *Caravan: Practical Online Learning of In-Network ML Models with Labeling Agents*
Qizheng Zhang, Ali Imran, Enkeleda Bardhi, Tushar Swamy, Nathan Zhang, Muhammad Shahbaz, Kunle Olukotun
USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2024
SRC JUMP 2.0 Best Paper Award

4. *CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion*
Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, **Qizheng Zhang**, Kuntai Du, Shan Lu, Junchen Jiang
ACM European Conference on Computer Systems (EuroSys), 2025
EuroSys Best Paper Award

5. *CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving*
Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, **Qizheng Zhang**, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, Junchen Jiang
ACM Special Interest Group on Data Communication (SIGCOMM), 2024

6. *FlowRL: Matching Reward Distributions for LLM Reasoning*
Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, **Qizheng Zhang**, Lin Chen, Fanghao Shao, Bo Xue, Yunchong Song, Zhenjie Yang, Ganqu Cui, Ning Ding, Jianfeng Gao, Xiaodong Liu, Bowen Zhou, Hongyuan Mei, Zhouhan Lin
International Conference on Learning Representations (ICLR), 2026

Research Experience

Stanford Pervasive Parallelism Lab Mar. 2023 - Now
Research Assistant, Agentic AI and Systems
Advisor: Kunle Olukotun Collaborators: James Zou, Muhammad Shahbaz

- Designed *ACE* (ICLR 2026), an agentic framework that enables language models to self-improve through generation–reflection–curation cycles, introducing scalable context evolution for long-term memory and interpretability.
- Developed *Agentic Plan Caching* (NeurIPS 2025), a test-time memory system that reuses structured reasoning traces to amortize agent planning cost across related queries.
- Developed *Caravan* (OSDI 2024), the first system enabling real-time in-network model adaptation using LLM-based labeling agents, with FPGA-hosted models supporting on-the-fly updates for autonomous drift detection and retraining.

LMCache Lab Nov. 2023 - Now
Research Assistant, Efficient Language Model Inference
Advisor: Junchen Jiang Collaborators: Ganesh Ananthanarayanan, Shan Lu

- Extended and benchmarked `vLLM` to support diverse KV cache compression algorithms for large language model inference.
- Co-developed *CacheGen* (SIGCOMM 2024) and *CacheBlend* (EuroSys 2025 Best Paper), research that led to `LMCache`, an open-source library optimizing the KV cache layer in large language model inference with 5.9K Github stars.

University of Chicago Networked Systems Group Sep. 2019 - Sep. 2022
Research Assistant, Video Analytics Systems
Advisor: Junchen Jiang Collaborators: Ravi Netravali

- Built video streaming pipelines that advanced the cost–accuracy frontier in large-scale vision analytics, leading to publications at venues such as SIGCOMM, NSDI, and MLSys.
- Optimized `x264` and `x265` codecs for adaptive, neural inference–aware video compression.

Industry Experience

Microsoft Research Redmond Jun. 2025 - Sep. 2025
Research Intern, Model Routing for Efficient Agentic AI
Mentors: Sharad Agarwal, Alec Wolman

- Investigated model routing for agentic workloads with tool-call awareness, uncovering key trade-offs between inference cost and performance in large-scale language model inference.
- Achieved a 28% reduction in inference cost compared to a production-grade model router at Microsoft Azure AI Foundry.

Honors and Awards

- **Best Paper Award:** EuroSys (2025), SRC JUMP 2.0 (2024)
- **Distinguished Artifact Award:** ISCA (2024)
- **Scholarship:** Soong Ching Ling Foundation ($12.5K), Jeff Metcalf Fellowship
- **Academic Honor:** Phi Beta Kappa

Skills

- **Programming**: C, C++, CUDA, Python, Javascript, Bash, SQL, MATLAB, R
- **ML/DL Frameworks**: PyTorch, HuggingFace Transformers
- **LLM Inference Frameworks**: vLLM, LMCache
- **Agentic AI Frameworks**: LangChain, LangGraph, DSPy

Service

Reviewer for major AI and systems venues (NeurIPS, ICML, ICLR, AAAI, AISTATS, EuroSys).