

Hao Kang

[my github](#)
[google scholar](#)

[personal link](#)
[my email](#)

Education

Georgia Institute of Technology

PhD. Student in Computer Science

Zhejiang University

Bachelor in Computer Science with CKC Honor

Aug. 2023 – Present

Advisor: Tushar Krishna

Aug. 2019 – June. 2023

Experience

Visiting Researcher at MIT

ML architecture and distributed learning

- Working on LLM Agentic Efficiency

Sep. 2025 – Now

Mentor Prof Song Han

Research Intern at TogetherAI

LLM Agentic Efficiency and Agentic RL

- Leading project of ThunderAgent, the first program-aware agentic LLM infra.

Feb. 2026 – Apr. 2026

Mentor Simran Arora

Research Intern at MSR

efficient LLM and model compression

- a paper accepted by MLsys 2025

May. 2024 – Aug. 2024

Mentor Srikant Bharadwaj

Graduate Researcher at GT

Efficient machine learning and LLM agent

Aug. 2023 – Now

Advisor Prof. Tushar Krishna

Undergrad Researcher at UCLA

dataset distilling

- a paper accept by ICML 2024

Aug. 2022 – Mar. 2023

Advisor Prof. Baharan Mirzasoleiman

Undergrad Researcher at MIT

model compression and edge ml

- a 4k+ star Github repo
- Deploy model on cell phone with TVM android and pytorch mobile

Feb. 2022 – Aug. 2022

Advisor Prof. Song Han

Publications

ThunderAgent: A Simple, Fast and Program-Aware Agentic Inference System

LLM agents, efficient ml

Hao Kang*, Ziyang Li*, Xinyu Yang*, Weili Xu*, Yinfang Chen, Junxiong Wang, Beidi Chen, Tushar Krishna, Chenfeng Xu, Simran Arora
ICML 2026 **Spotlight** Acquired by TogetherAI, Anyscale and Dynamo's product.

Win Fast or Lose Slow: Balancing Speed and Accuracy in Latency-Sensitive Decisions of LLMs

LLM agents, efficient ml, LLM for trading

Hao Kang, Qingru Zhang, Han Cai, Weiyuan Xu, Tushar Krishna, Yilun Du, Tsachy Weissman
Neurips 2025 **Spotlight**

OckBench: Measuring the Efficiency of LLM Reasoning

Reasoning efficiency, efficient ml

Zheng Du*, **Hao Kang***, Song Han, Tushar Krishna, Ligeng Zhu
Neurips 2025 Reasoning Efficiency Workshop

TURBOATTENTION: EFFICIENT ATTENTION APPROXIMATION FOR HIGH THROUGHPUTS LLMs

efficient ml, hardware

Hao Kang, Srikant Bharadwaj, James Hensman, Tushar Krishna, Victor Ruehle, Saravan Rajmohan
Mlsys 2025

GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative Inference of LLM

model compression, efficient ml

Hao Kang*, Qingru Zhang*, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, Tuo Zhao
NIPS ENLSP 2025 **Best Paper Candidate**, AiStats2026

Slim-mux: Orchestrating small language models for reasoning

efficient ml, multiagent system

Chenyu Wang*, Zishen Wan*, **Hao Kang**, Zhiqiang Xie, Vijay Janapa Reddi, Tushar Krishna, Yilun Du
In submission

Towards Sustainable Learning: Coresets for Data-efficient Deep Learning

dataset distilling, efficient ml

Yu Yang, **Hao Kang**, Baharan Mirzasoleiman
ICML2024

AI Metropolis: Scaling Large Language Model Agent Interaction with Out-of-order Execution

LLM agents, efficient ml

Zhiqiang Xie, **Hao Kang**, Ying Sheng, Tushar Krishna, Kayvon Fatahalian, Christos Kozyrakis
Mlsys 2025

Privatar: Enabling Privacy-preserving Real-time Multi-user VR via Secure Outsourcing

efficient ml, ai security

Jianming Tong, Hanshen Xiao, **Hao Kang**, Krishnakumar Nair, Ashish Sirasao, G. Edward Suh, Tushar Krishna.
USENIX Security 2025

LvM-compress-bench: Benchmarking the broader impact of large vision-language model compression

ML efficiency, Benchmark

Souvik Kundu, Anahita Bhiwandiwala, Sungduk Yu, Phillip Howard, Tiej Le, Sharath Nittur Sridhar, David Cobbley, Hao Kang, Vasudev Lal
NAACL 2025

Open-source Projects

THOP: PyTorch-OpCounter

a pytorch operator profiler which has over **4.8k** stars

ThunderAgent

Agentic Infra that has over **260** stars

GEAR

KV cache compression which has over **180** stars

Extracurricular

Research Interests

1. Agentic context management.
2. Agentic infra
3. Deploying LLM Agent to more practical use case like trading.