Hao Li

li003703@umn.edu https://nanomaoli.github.io GitHub Google Scholar

RESEARCH INTEREST

I am interested in **Machine Learning and Systems**. Recently, I have been working on:
- Determinism in LLM Inference and Training.
- Elastic KV Cache for Efficient GPU Memory Utilization.

EDUCATION

University of Minnesota Twin Cities **Sept. 2024 – Present**
Ph.D. studies in Computer Science (In Progress) *GPA: 3.75*

University of California, Santa Barbara **Sept. 2022 – Jun. 2024**
Ph.D. studies in Computer Engineering (Transferred) *GPA: 3.48*

Texas A&M University **Sept. 2019 – May 2022**
M.S. in Computer Engineering *GPA: 3.91*

Beihang University **Sept. 2015 – Jun. 2019**
B.Eng. in Instrumentation *GPA: 3.59*

PUBLICATIONS

- [**NeurIPS 2025**] Jiayi Yuan*, **Hao Li***, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, Zirui Liu. *Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference.* In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems.* Oral Presentation. (Acceptance Rate: 77/21575, ∼0.36%). arXiv Code
- [**ISCA 2025**] Guyue Huang, **Hao Li**, Le Qin, Jiayi Huang, Yangwook Kang, Yufei Ding, Yuan Xie. *TRACI: Network Acceleration of Input-Dynamic Communication for Large-Scale Deep Learning Recommendation Model.* In *Proceedings of the 52nd Annual International Symposium on Computer Architecture.* (Acceptance Rate: 127/570, ∼22.28%). Paper

RESEARCH EXPERIENCE

kvcached: Elastic KV Cache | *ML Systems, Open-Source* **Jun. 2025 – Sept. 2025**
Prof. Jiarong Xing, Rice
- Implemented a working tensor-parallel (TP) KV-cache mapping mechanism in kvcached. Ensured consistent KV-cache layout across different TP sizes and correctness for reproducible inference.
- Built synchronous worker and asynchronous broadcasting scheduler, reducing IPC overhead.
- Designed standalone benchmarks for evaluating IPC overhead and KV cache memory mapping. Designed and updated testing scripts into meaningful unit tests or fixtures compatible with updated APIs.
- Assisted in open-sourcing, writing review-friendly documentation and resolving issues. GitHub Repo

Nondeterminism of LLM Reasoning | *LLM Reasoning, Reliability* **Feb. 2025 – Jun. 2025**
Prof. Zirui Liu, UMN
- Identified the nondeterminism problem of LLM inference through hacking and experiments on reasoning models.
- Implemented the codebase for evaluating LLM inference nondeterminism, babysat all the experiments, collected and clean experimental results for further analysis.
- Assisted in designing the experiment plan and data analysis on the results.
- Wrote the experiment and evaluation sections of the paper, helped polish the whole paper.

KV Cache Quantization for Long-Context LLMs | *Efficient ML, Quantization* **Jan. 2025 – Mar. 2025**
Prof. Zirui Liu, UMN
- Helped build the codebase for KV cache quantization experiments which implements the PagedAttention algorithm.
- Conducted extensive simulation experiments for different quantization strategies across various LLMs.

In-Network Acceleration for DLRMs | *Microarchitecture, Interconnect* **Oct. 2022 – Sept. 2023**
Prof. Yuan Xie, Prof. Yufei Ding, UCSB
- Designed and implemented microarchitecture simulator in Gem5/Garnet for in-switch cache and reduction table.
- Conducted simulation experiments, processed and analyzed the simulation results.
- Assisted in paper writing for the experiment and evaluation part.

TEACHING

Teaching Assistant
- UMN, EE 1301 Introduction to Computing Systems (Fall 2025).
- UMN, EE 2015 Signals, Circuits and Electronics (Fall 2025).
- UCSB, ECE 180 Introduction to Deep Learning (Spring 2024).
- UCSB, ECE 10 Foundations of Analog and Digital Circuits & Systems (Fall 2023, Winter 2024).

Grader
- TAMU, ECEN 757 Distributed System (Spring 2021).
- TAMU, ECEN 350 Computer Architecture and Design (Fall 2021).
- TAMU, ECEN 765 Machine Learning with Networks (Spring 2022).

HONORS AND AWARDS

- **NeurIPS 2025 Oral (77 out of 21575 submissions), *by NeurIPS*** **Sept. 2025**
- **Professor Aldert van der Ziel Memorial Graduate Fellowship, *by UMN ECE*** **Feb. 2024**
- **Dr. Krzysztof K. Burhardt and April L. Spas Fellowship, *by UMN ECE*** **Feb. 2024**
- **Graduate Merit Scholarship, *by TAMU ECE*** **Aug. 2021**

SERVICES

Conference Volunteers: Hot Chips 2023/2024
Artifact Evaluation Committee Member: *MICRO 2024*, *HPCA 2025*