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. in Electrical Engineering (In Progress)

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

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

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

PUBLICATIONS

- **[NeurIPS 2025](To appear)** Jiayi Yuan*, **Hao Li***, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, Zirui Liu. *Give Me FP32 or Give Me Death? Challenges and Solutions for Reproducible Reasoning.* 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
- Overview: Developed a package for elastic KV cache memory management to enable dynamic GPU sharing and efficient multi-LLM serving. GitHub Repo.
- My Role: Assisted in the open-sourcing process, implemented kvcached support for tensor-parallel serving, developed benchmarks and pytests for several kvcached components.

Nondeterminism of LLM Reasoning | *LLM Reasoning, Reliability* **Feb. 2025 – Jun. 2025**
Prof. Zirui Liu, UMN
- Overview: Studied nondeterminism in LLM reasoning, developed a memory-efficient method with enhanced determinism.
- My Role: Primary contributor in all phases – problem setup, codebase, experiments, evaluation, paper writing.

KV Cache Quantization for Long-Context LLMs | *Efficient ML, Quantization* **Jan. 2025 – Mar. 2025**
Prof. Zirui Liu, UMN
- Overview: Proposed a hardware-aware algorithm that uses integer tensor cores for efficient KV cache quantization.
- My Role: Conducted simulation of quantization algorithms, helped build our codebase that implements PagedAttention.

Chiplet-based Accelerators for MoE Models | *Hardware Architecture, Chiplets* **Sept. 2024 – Nov. 2024**
Prof. Yang Zhao, UMN
- Overview: Studied algorithms-hardware co-design opportunities in chiplet and MoE models.
- My Role: Built performance modeling for ViT MoE workload in a chiplet simulator.

In-Network Acceleration for DLRMs | *Microarchitecture, Interconnects* **Oct. 2022 – Sept. 2023**
Prof. Yuan Xie, Prof. Yufei Ding, UCSB
- Overview: Proposed in-network computing architecture to accelerate embedding layers of deep learning recommendation models.
- My Role: Helped design in-network embedding cache, implemented microarchitecture simulator in Gem5/Garnet, conducted simulation experiments, helped with paper writing for the evaluation part.

Hybrid Memory Management Unit | *Microarchitecture, Memory Systems* **Aug. 2020 – Aug. 2021**
Prof. Paul Gratz, TAMU
- Overview: Explored a hybrid memory architecture comprising DRAM and NVM.
- My Role: Studied microarchitecture simulation, developed performance modeling for HMMU in ChampSim simulator.

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*