

# Genghan Zhang

zhang677.github.io | ghzhang19@gmail.com

## EDUCATION

### Stanford University

PhD Student in Computer Science

September 2023 - May 2028 (expected)

Stanford, USA

- Research Interests: Self-improving LLM Agents for Programming AI Accelerators

### Tsinghua University

Bachelor of Engineer in Electronic Information Science and Technology

August 2019 - June 2023

Beijing, China

- GPA: 3.94/4.00 (Top 3%)

## RESEARCH EXPERIENCE

### Research Assistant

Department of Computer Science, Stanford University

April 2024 - Present

Stanford, CA

- Advisor: Prof. Kunle Olukotun
- Designing self-improving LLM agents for AI accelerator kernel implementation and optimization.

### Research Assistant

Department of Computer Science, Stanford University

January 2024 - March 2024

Stanford, CA

- Advisor: Prof. Azalia Mirhoseini
- Proposed GPU kernel fusion techniques to accelerate FFN layers for LLM inference by utilizing the sparsity of activation.

Accepted by COLM 2024

### Research Assistant

Department of Computer Science, Stanford University

March 2022 - December 2023

Remote

- Advisor: Prof. Fredrik Kjølstad
- Designed an algorithm template and code generation algorithm for *sparse workspace* to solve the sparse scattering problem with a sparse tensor algebra compiler called TACO. Accepted by PLDI 2024.

## SELECTED PUBLICATIONS

### • AccelOpt: A Self-Improving LLM Agentic System for AI Accelerator Kernel Optimization

Genghan Zhang, Shaowei Zhu, Anjiang Wei, Zhenyu Song, Allen Nie, Zhen Jia, Nandita Vijaykumar, Yida Wang, and Kunle Olukotun.  
*In submission, 2026*

### • Adaptive Self-improvement LLM Agentic System for ML Library Development

Genghan Zhang, Weixin Liang, Olivia Hsu, Kunle Olukotun.  
*International Conference on Machine Learning (ICML), 2025*

### • Compilation of Modular and General Sparse Workspaces

Genghan Zhang, Olivia Hsu, Fredrik Kjolstad.  
*Programming Language Design and Implementation (PLDI), 2024*

### • CATS: Context-Aware Thresholding for Sparsity in Large Language Models

Donghyun Lee, Jaeyong Lee, Genghan Zhang, Mo Tiwari, Azalia Mirhoseini.  
*Conference on Language Modeling (COLM), 2024*

### • Sgap: Towards Efficient Sparse Tensor Algebra Compilation for GPU

Genghan Zhang, Yuetong Zhao, Yanting Tao, Zhongming Yu, Guohao Dai, Sitao Huang, Yuan Wen, Pavlos Petoumenos, Yu Wang.  
*CCF Transactions on High Performance Computing, 2023*

## WORK EXPERIENCE

### Applied Scientist Intern

Amazon

June 2025 - September 2025  
Santa Clara, USA

- Mentors: Shaowei Zhu and Zhenyu Song
- Self-improving LLM agentic system for AI accelerator kernel optimization

### Software Engineer

NVIDIA

June 2024 - September 2024  
Santa Clara, USA

- Mentor: Andrew Kerr
- Compiler for Tile IR

### Software Engineer

Infinigence Tech

May 2023 - July 2023  
Beijing, China

- Mentor: Prof. Xiuhong Li (PKU)
- Assembled an in-house GPU kernel library for LLM inference, which demos the company's first-generation product.

## SERVICE

### • Reviewer: ICML 2025, ICLR 2025 (Notable Reviewer), NeurIPS 2024, ICLR 2025 DL4C Workshop, NeurIPS 2024 Sys2-Reasoning Workshop, NeurIPS 2022 GLFrontiers Workshop

### • Artifact Evaluation Committee: ASPLOS 2025 summer, PLDI 2025

### • Program Committee: LATTE 2025

## TEACHING

- Stanford CS149 - Parallel Computing (Fall 2025)

## TECHNICAL SKILLS

### Programming Languages & Software Tools

- Most experienced: CUDA, Python, PyTorch, Matlab
- Some experience: C++, Rust, Verilog HDL, LtSpice