

Jiajie Li

School of Electrical and Computer Engineering, Cornell University
jl4257@cornell.edu | [Personal Website](#)

EDUCATION

Cornell University | Ithaca, New York, USA

Sep 2021 – Present

Ph.D. in Electrical and Computer Engineering

GPA: 4.00/4.00

M.S. in Electrical and Computer Engineering

Sep 2021 – Dec 2025

Courses: Applications of Parallel Computers (A+), High-Level Digital Design Automation (A+), Advanced Compilers (A), Advanced Topics in Hardware Acceleration for Deep Learning (A), Advanced Machine Learning Systems (A)

Tsinghua University | Beijing, China

Sep 2016 – Jun 2020

Bachelor of Engineering in Electronic Engineering

GPA: 3.81/4.00

INDUSTRY EXPERIENCES

AMD Research and Advanced Development | Research Associate | Mentors: Erwei Wang, Samuel Bayliss

An Agentic System for Efficient LLM Deployment on AIE NPUs

Jan 2026 – Present

- Built an agentic system that guides coding agents through kernel development and model composition for end-to-end LLM deployment on AMD XDNA™ 2 NPUs, combining a kernel database, reusable NPU engineering skills, and staged feedback on top of the open-source MLIR-AIR compiler flow. An early version was accepted by MLArchSys (co-located with ISCA 2026).
- Autonomously deployed multiple previously unsupported LLMs end-to-end, achieving significant energy-efficiency gains over an iGPU baseline and large prefill and decode speedups over the coding agent alone, with less agent time and lower API cost.
- Extended single-token decode megakernels to support DFlash speculative decoding through the agentic workflow, achieving 1.7–2.2× decode speedups over a production inference solution's autoregressive decoding.

ByteDance, MLSys Team in Seed Foundation | Research Scientist Intern | Mentors: Wenlei Bao, Li-Wen Chang

Characterizing High-Performance LLM Kernels on NVIDIA GPUs

Aug 2024 – Dec 2024

- Benchmarked GEMM and grouped GEMM across cuBLAS, Triton, and CUTLASS over a range of problem sizes, data types, and GPU architectures, using the CUTLASS profiler and Triton autotuning to find each library's best configuration.
- Identified warp specialization combined with the Tensor Memory Accelerator (TMA) on Hopper GPUs as the key to reaching peak performance, and analyzed why this design lets kernels overlap data movement with Tensor Core computation.

NVIDIA, End-to-End Training Perf Team | Deep Learning Training Perf Intern | Mentors: Rachit Garg, Burc Eryilmaz

LLM Training Toolbox: Memory Footprint Analyzer & Configuration Autotuner

May 2024 – Aug 2024

- Built a memory footprint analyzer that collects and analyzes memory snapshots for peak-memory debugging and memory-leak detection in LLM training, and integrated it into NeMo and the MLPerf training pipeline; the tool helped teammates debug memory issues with CUDA Graphs.
- Built an autotuner that searches the large training-configuration space (TP, PP, CP sizes, TP-overlap configs, etc.) for the best setup given a model and hardware, delivering a 1.1% end-to-end speedup over the heavily hand-tuned MLPerf submission config.

Alibaba DAMO Academy, Computing Technology Lab | Research Intern | Mentor: Yuanwei Fang

Neural Program Sampling for Fast Microarchitecture Simulation

Jul 2020 – Jan 2021

- Contributed to Neural Program Sampling (NPS), a GNN-based framework that learns execution embeddings from code structure and runtime state to select representative program intervals for simulation, outperforming SimPoint by up to 63%.
- Built the NPS-gem5 evaluation testbed by extending gem5 to emit detailed-simulation statistics at specified instruction intervals, enabling fast and flexible evaluation of sampling accuracy.

RESEARCH PROJECTS

Cornell University, Computer Systems Lab | Ph.D. Candidate | Advisor: Prof. Zhiru Zhang

Sep 2021 – Present

Agentic Kernel Generation Across Hardware Targets

Mar 2025 – Dec 2025

- Prototyped an LLM agentic workflow that generates and tunes kernels for CPUs, GPUs, and the AMD AI Engine through a feedback loop of correctness checking, error debugging, and iterative optimization, grounded in curated AI Engine API documentation; this work laid the foundation for the agentic NPU deployment system developed at AMD.

HeuriGym: An Agentic Benchmark for LLM-Crafted Heuristics in Combinatorial Optimization

Feb 2025 – Jun 2025

- Co-developed HeuriGym (ICLR'26, core contributor), an open-source agentic benchmark in which LLMs iteratively write, execute, and refine heuristics for nine combinatorial optimization problems, with a continuously updated public leaderboard.
- Contributed the technology-mapping problem from the electronic design automation (EDA) domain, including its formalization and expert solution; it remains among the hardest problems in the suite, where even the latest frontier models score lowest.

Rapid GPU-Based Pangenome Graph Layout

Nov 2022 – Apr 2024

- Developed the first GPU implementation of path-guided SGD for pangenome graph layout, achieving a $57.3\times$ average speedup on an NVIDIA A100 over a 32-thread CPU baseline and reducing layout time for 24 human whole-chromosome pangenomes from hours to minutes without quality loss. The work was accepted by SC'24 and integrated into the ODGI pangenome analysis toolkit.
- Characterized the workload as memory-bound with highly irregular access patterns, then designed CUDA optimizations including a cache-friendly data layout, coalesced random-number-generator states, and warp merging to reduce branch divergence.
- Proposed sampled path stress, a sampling-based layout-quality metric that reduces evaluation cost from quadratic to linear time complexity, and used it to explore performance-quality trade-offs, gaining a further $1.5\times$ speedup via warp-level data reuse.

Analysis and Optimization of GNN-Based Recommender Systems on Persistent Memory

Sep 2021 – Nov 2022

- Showed that Intel Optane persistent memory enables single-machine training of GNN-based recommender systems on billion-edge graphs, outperforming distributed training by over $30\times$ for deep GNN models at less than half the memory cost.
- Built and tuned the distributed training baseline with DistDGL on a six-server CPU cluster, running NGCF and LightGCN on graphs with up to one billion edges for an end-to-end comparison against single-machine Optane training.

UCLA, VAST Lab | Research Intern | Advisor: Prof. Jason Cong

Jun 2019 – Sep 2019

HeteroHalide: An End-to-End Compilation System from Image Processing DSL to Efficient FPGA Acceleration

- Proposed HeteroHalide, an end-to-end compilation flow from Halide to FPGA accelerators, including a Halide-to-HeteroCL code generator in C++ and scheduling extensions that emit lower-level primitives at the spatial-architecture backend. The work was accepted by FPGA'20.
- Showed that the generated accelerators outperform both a 28-core CPU ($4.15\times$ speedup, $11.7\times$ energy efficiency) and the state-of-the-art Halide-to-FPGA compiler ($2\text{--}4\times$ throughput), while requiring only moderate changes to Halide schedules.

Tsinghua University, NICS-EFC Lab | Research Assistant | Advisor: Prof. Yu Wang

Nov 2018 – Jun 2019

Hardware-Friendly Neural Network Compression

- Studied hardware-friendly model compression for efficient deployment, quantifying how low-bit quantization, pruning, and knowledge distillation trade off model size against accuracy.

PUBLICATIONS

- **Jiajie Li**, Erwei Wang, Zhiru Zhang, Samuel Bayliss, "From Human Guidance to Autonomy: Agent Skill System for End-to-End LLM Deployment on Spatial NPUs", MLArchSys Workshop (co-located with ISCA 2026).  
- Xiaoyu Liang, Elaine Cao, Chun-Ning Kao, David Lin, **Jiajie Li**, Zhiru Zhang, "Deploying SmolVLA Model on an AMD XDNA NPU", FCCM 2026. 
- Hongzheng Chen*, Yingheng Wang*, Yaohui Cai*, Hins Hu*, **Jiajie Li***, Shirley Huang, Chenhui Deng, Rongjian Liang, Shufeng Kong, Haoxing Ren, Samitha Samaranyake, Carla P. Gomes, Zhiru Zhang, "HeuriGym: An Agentic Benchmark for LLM-Crafted Heuristics in Combinatorial Optimization", ICLR 2026.   (*core contributors)
- **Jiajie Li**, Jan-Niklas Schmelzle, Yixiao Du, Simon Heumos, Andrea Guarracino, Giulia Guidi, Pjotr Prins, Erik Garrison, Zhiru Zhang, "Rapid GPU-Based Pangenome Graph Layout", SC 2024.  
- Simon Heumos, Andrea Guarracino, Jan-Niklas M. Schmelzle, **Jiajie Li**, Zhiru Zhang, Jörg Hagmann, Sven Nahnsen, Pjotr Prins, Erik Garrison, "Pangenome graph layout by Path-Guided Stochastic Gradient Descent", Bioinformatics, Vol. 40, Issue 7, July 2024.  
- **Jiajie Li**, Yuze Chi, Jason Cong, "HeteroHalide: From Image Processing DSL to Efficient FPGA Acceleration", FPGA 2020.  
- Erwei Wang, Ephrem Wu, Victor J. B. Jung, **Jiajie Li**, Andre Rosti, Joseph Melber, Samuel Bayliss, "Programming AMD XDNA NPUs with Open-source Compiler Tools: A FlashAttention Case Study", arXiv:2609.21264, 2026. 
- Shihan Fang, Hongzheng Chen, Niansong Zhang, **Jiajie Li**, Han Meng, Adrian Liu, Zhiru Zhang, "Dato: A Task-Based Programming Model for Dataflow Accelerators", arXiv:2509.06794, 2025. 
- Yuanwei Fang, Zihao Liu, Yanheng Lu, Jiawei Liu, **Jiajie Li**, Yi Jin, Jian Chen, Yenkuang Chen, Hongzhong Zheng, Yuan Xie, "NPS: A Framework for Accurate Program Sampling Using Graph Neural Network", arXiv:2304.08880, 2023. 
- Yuwei Hu, **Jiajie Li**, Zhongming Yu, Zhiru Zhang, "Analysis and Optimization of GNN-Based Recommender Systems on Persistent Memory", arXiv:2207.11918, 2022. 

AWARDS & HONORS

2nd Place, Reconfigurable Computing Challenge (RCC), FCCM 2026	2026
DAC'23 Young Fellow	2023
Tsinghua – Zheng Geru Scholarship (awarded to 8 Tsinghua undergraduates for academic excellence)	2019
Scholarship for Academic Excellence, Tsinghua University	2018, 2019
Scholarship for Academic Progress (top 10%), Tsinghua University	2017

TEACHING & SERVICES

Teaching Assistant: ECE 2300 “Digital Logic and Computer Organization”, Cornell University	Aug 2023 – Dec 2023
Reviewer: IEEE TCAD 2025; MLArchSys @ ISCA 2026; MATH-AI @ NeurIPS 2025, 2026; RCC @ FCCM 2026	
Student Volunteer: FCCM'22	May 2022

SKILLS

Coding Agents : Claude Code, OpenAI Codex, Cursor

Languages : C/C++, Python

GPU Programming : CUDA, Triton, CUTLASS

Compilers & DSLs : MLIR, Halide, High-Level Synthesis C/C++

Parallel Programming : OpenMP, MPI

Profiling Tools : NVIDIA Nsight Systems, Nsight Compute, Intel VTune