Lijie(Derrick) Yang

Email: ly3223@princeton.edu | Website: https://derrickylj.github.io/

Education

Princeton University July 2025 - Present
PhD in Computer Science. Advisors: Tri Dao & Ravi Netravali *King Peh Kwoh Fellowship*

Carnegie Mellon University Sept 2021 - May 2025
B.S. in Computer Science. Advisors: Zhihao Jia & Tianqi Chen *University & College Honors (GPA: 4.0/4.0)*

Research Interests
My research has been focusing on developing more efficient and accurate AI models with co-design of hardware and software. We explore new methodologies and system designs that push the capability boundary of model in most challenging tasks.

Publications

First-author papers

- **(Under Review at COLM '26) Thought-Level Beam Search for Reasoning.** Lijie Yang, Hongyin Luo, Jiawei Zhao, Tri Dao, Ravi Netravali. [pdf]
- **(ICML '26) Less Is More: Fast and Accurate Reasoning with Cross-Head Unified Sparse Attention.** Lijie Yang*, Zhihao Zhang*, Arti Jain, Shijie Cao, Baihong Yuan, Yiwei Chen, Zhihao Jia, Ravi Netravali. [pdf]
- **(ICLR '25) TidalDecode: Fast and Accurate LLM Decoding with Position Persistent Sparse Attention.** Lijie Yang*, Zhihao Zhang*, Zhuofu Chen, Zikun Li, Zhihao Jia. [pdf]
- **(IEEE IoT '25) From Machine Learning-Based to LLM-Enhanced: An Application-Focused Analysis of How Social IoT Benefits from LLMs**. Lijie Yang, Runbo Su. [pdf]
- **(LCN '24) Blocking-Waived Estimation: Improving the Worst-Case End-To-End Delay Analysis in Switched Ethernet.** Lijie Yang, Théo Docquier, Ludovic Thomas, Ye-Qiong Song. Lijie Yang, Runbo Su. [pdf]
- **(Undergraduate Research Thesis) Accurate and Efficient Sparse Attention Systems for Long-Context LLMs: A Study with TidalDecode.** Lijie Yang. [pdf]

Collaborations

- **(ArXiv) Geometry Guided Self-Consistency for Physical AI**
  Yinwei Dai, Zhuofu Chen, Lijie Yang, Ravi Netravali. [pdf]
- **(ArXiv) Axe: A Simple Unified Layout Abstraction for Machine Learning Compilers**
  Bohan Hou, Hongyi Jin, Guanjie Wang, Jinqi Chen, Yaxing Cai, Lijie Yang, Zihao Ye, Yaoyao Ding, Ruihang Lai, Tianqi Chen. [pdf]
- **(Technical Report) Beyond Context Limits: Subconscious Threads for Long-Horizon Reasoning**
  Hongyin Luo, Nathaniel Morgan, Tina Li, Derek Zhao, Ai Vy Ngo, Philip Schroeder, Lijie Yang, Assaf Ben-Kish, Jack O'Brien, James Glass. [pdf]
- **(Technical Report) CaveAgent: Transforming LLMs into Stateful Runtime Operators**
  Maohao Ran, Zhenglin Wan, Cooper Lin, Yanting Zhang, Hongyu Xin, Hongwei Fan, Yibo Xu, Beier Luo, Yaxin Zhou, Wangbo Zhao, Lijie Yang etc. [pdf]
- **(MLSys '26) Event Tensor: A Unified Abstraction for Compiling Dynamic Megakernel**
  Hongyi Jin, Bohan Hou, Guanjie Wang, Ruihang Lai, Jinqi Chen, Zihao Ye, Yaxing Cai, Yixin Dong, Xinhao Cheng, Zhihao Zhang, Yilong Zhao, Yingyi Huang, Lijie Yang, Jinchen Jiang, Gabriele Oliaro, Jianan Ji, Xupeng Miao, Vinod Grover, Todd C. Mowry, Zhihao Jia, Tianqi Chen. [pdf] 🏆 Oral Presentation
- **(ICML '24) Accelerating Iterative Retrieval-augmented Language Model Serving with Speculation**
  Zhihao Zhang, Alan Zhu, Lijie Yang, Yihua Xu, Lanting Li, Phitchaya Mangpo Phothilimthana, Zhihao Jia. [pdf]
- **(ASPLOS '24) SpecInfer: Accelerating Large Language Model Serving with Tree-based Speculative Inference and Verification.** Xupeng Miao*, Gabriele Oliaro*, Zhihao Zhang*, Xinhao Cheng*, Zeyu Wang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, Zhihao Jia. [pdf] 🏆 Cited 300+ times

(*Equal Contribution)

RESEARCH EXPERIENCE

DaoAI & System4AI Lab July. 2025 - Present
PhD Researcher with Prof. Tri Dao and Prof. Ravi Netravali *Princeton University*

LessIsMore: Training-Free Sparse Attention Model for Efficient Reasoning
- Developed accurate sparse attention model for reasoning tasks with unified selection and fixed recency window.
- Implemented customized kernels for token selection and system on LLMs backed with GQA mechanism.
- Achieving up to 1.6× end-to-end speedup and 2× fewer attended tokens without accuracy loss on reasoning benchmarks.

Catalyst Lab Apr. 2023 - June 2025
Research Assistant with Prof. Zhihao Jia and Prof. Tianqi Chen *Carnegie Mellon University*

TidalDecode: Fast & Accurate LLM Decoding with Position-Persistent Sparse Attention
- Developed an efficient decoding algorithm using selective sparse attention, reducing memory usage and latency bottlenecks.
- Implemented efficient kernels for token selection, supporting both full attention and fixed-token sparse attention.
- Achieved up to a 2.1x reduction in decoding latency while preserving performance comparable to full-weight attention.

Axe: A Full-Granularity Distributed-Aware Tensor Programming Intermediate Language and Compiler
- Implemented new Blackwell datatypes and operators within the Intermediate Representation to facilitate buffer visualization and enhance CUDA code generation, committed to upstream GitHub repo TVM (12.5K stars) apache/tvm
- Optimized tiled-layout normalization and vectorization for tensor core ops, enabling efficient GEMM kernels on A100s.
- Conducted benchmarks on low-bit GEMM/V configurations to evaluate and compare performance across various libraries.

SpecInfer: Generative LLM Accelerator with Speculative Inference (1.7K GitHub stars)
- Optimized Docker configurations and workflows to support over ten CUDA versions for the deep learning system.
- Integrated multiple half-precision LLaMA model weights and facilitated their deployment on AWS for model testing.
- Prebuilt the Legion framework on Oracle Cloud Infrastructure, creating a release that decreased CI build latency by 75%.

RaLMSpec: Accelerating Iterative Retrieval-Augmented Language Model (RaLM) Serving with Speculation
- Developed a system to enhance RaLM pipeline latency by speculatively retrieving documents from a local cache.
- Implemented asynchronous verification and correction mechanisms in retrieval phase and validated model performance.
- Achieved a generic maximum speedup of 2.39x on iterative RaLM applications with all features enabled.

HAN Lab Dec. 2023 - Jun. 2024
Research Assistant with Dr. Wei-Chen Wang and Prof. Song Han *Massachusetts Institute of Technology*

TinyChatEngine: On-Device LLM Inference Library (600 GitHub stars)
- Engineered a Metal-accelerated system for TinyChatEngine on macOS/iOS devices, utilizing quantized Llama.cpp models.
- Developed and optimized Metal compute kernels for all stages of model generation, including matrix multiplication and activation functions.
- Attained execution speed and performance comparable to CUDA-implemented GEMM kernels when operating under mixed-precision settings.

Team Simbiot May 2022 - May 2024
Research Assistant with Prof. Ye-Qiong Song *Lorraine Research Laboratory*

Blocking-Waived Estimation: A Tighter Worst-Case E2E Delay Bound Estimation in Switched Ethernet
- Proposed 'Blocking-Waived Estimation' (BWE), a novel Response Time Analysis (RTA)-based approach for worst-case delay estimation in asynchronous frame-switching networks.
- Minimized redundant calculations for delay analysis between adjacent hops in full-duplex transmission links through recursive generalization.
- Validated BWE's effectiveness in a dual-switch industrial use-case, demonstrating up to a 34% reduction in estimated worst-case delay compared to standard RTA.

TEACHING EXPERIENCE

15-210: Parallel and Sequential Data Structures and Algorithms	S23, F23, S25
18-213/18-613: Foundations of Computer Systems	Summer 22, F22
18-100: Introduction to Electrical and Computer Engineering	S22

Awards

- *King Peh Kwoh Fellowship*, Princeton University 2026
- *Silver Reviewer Award*, ICML 2026
- *Top Reviewer Award*, NeurIPS 2025
- *Dean's List, High Honors*, Carnegie Mellon University All Semesters
- *Honorable Mention*, Allen Newell Award for Excellence in Undergraduate Research, School of Computer Science, Carnegie Mellon University 2025
- *Senior Leadership Recognition Award*, Carnegie Mellon University 2025
- *Phi Beta Kappa Early Initiation* (Top 1.9% of CMU Seniors) 2024
- *Summer Undergraduate Research Apprenticeship*, Carnegie Mellon University 2024

Service

Reviewer: COLM 26; ICML 26; ICLR 25, 26; NeurIPS 25, 26

Ambassador, SCS Undergraduate AI Mentoring Program Carnegie Mellon University