Brian E. Zhang

(954) 600-7197
bez@cmu.edu
homepage
atomicapple0

EDUCATION

May.'23 – May.'24 **Carnegie Mellon University** Pittsburgh, PA
Masters of Science in Computer Science, Research Thesis (GPA: 4.0/4.0)
Built LithOS, an operating system for multi-tenant deep learning workloads on GPUs. Advised by Dimitrios Skarlatos & Todd Mowry.

Aug.'19 – May.'23 **Carnegie Mellon University** Pittsburgh, PA
Bachelors of Science in Computer Science (GPA: 3.8/4.0; $200,000 grant)
Coursework includes: Advanced Distributed & Operating Systems; Robot Localization & Mapping; Computer Graphics; Computer Vision; Compiler Design; HoT Compilation; . . .

WORK EXPERIENCE

Nov.'24 – Present **Modular - LLM Serving.** Software Engineer
Built large portions of batch scheduler and PagedAttention memory management system in the MAX LLM serving platform. Main DRI for Prefix Caching and KVCache CPU offloading. Also wrote some GPU kernels in Mojo and optimized various operator fusion passes within our C++ MLIR based Graph Compiler.

Jun.'22 – Aug.'22 **Meta - AI Infra.** Software Engineer Intern
Improved Starlight, an internal ML pipelining platform, with compile-time type checking for various native Python types to preemptively catch failures in user code before launching expensive training jobs.

Sep.'21 – Dec.'21 **NASA - Orion Backup Flight Software** Software Engineer Intern
Engineered software limits on rocket thruster firings to meet power usage requirements on the Orion spacecraft for the Artemis II mission. Built tooling to manage how bytes are packed in Orion's telemetry message structs.

Jun.'21 – Aug.'21 **Amazon - Search Relevance** Software Developer Intern
Extended Amazon's A/B testing library to track the impact of newly released Amazon search ranking features on key business metrics. Scheduled daily Spark jobs to clean, preprocess, and extract insights from petabytes of user data.

PROJECTS

May.'23 – Present **LithOS: An OS for GPUs** Lead Researcher & Developer
LithOS achieves best-in-class performance isolation and GPU utilization across many GPU sharing benchmarks. Required significant reverse engineering effort for NVIDIA GPU drivers. Written in Rust & CUDA. Work is a collaboration with Meta and a submission to SOSP'25.

Jan.'24 – May.'24 **SMoL: A SML to C Compiler** Developer
Implemented compiler passes including elaboration, hoisting, closure conversions, etc in the SML functional programming language. Includes cheney-scan semispace garbage collector.

Oct.'22 – Nov.'22 **Pebbles OS: A Preemptive Unix Kernel** Developer
Developed a Unix kernel from scratch in C & x86 assembly. Supports guest OSes with paravirtualization. Also wrote a POSIX-like user-space threading library on top of Pebbles.

Mar.'22 – May.'22 **RadarSLAM: Localization for Self-Driving Cars in Adverse Weather** Developer
Wrote first open-source implementation of SOTA RadarSLAM algorithm. Evaluated algorithm performance on real-world driving datasets. 30+ GitHub stars.

Made with ❤ using LaTeX

LEADERSHIP & SERVICE

Jul.'23 – Aug.'23 **Come On Out - Japan** Teacher
Taught English to Japanese middle and high school students for five weeks in Tokyo, Nagano, and Yamanashi.

May.'20 – May.'23 **CMU School of Computer Science** Teaching Assistant
Graded student work, wrote exams, and taught recitations for Principles of Imperative Computation (Summer '20), Introduction to Robotics (Spring '23), and Operating System Design and Implementation (Fall '23).
Received overwhelmingly positive student feedback. Read reviews here.

Dec.'21 – May.'23 **CMU Explorer's Club** Quartermaster
Maintained club's outdoor equipment and hosted weekly gear checkouts for members.

Dec.'20 – Jan.'22 **CMU Puzzlehunt** Staff & Puzzle Writer
Organized and wrote the biannual CMU Puzzlehunt for over 1500 participants.
My puzzles include: Mother Functions, The Pirate's Gambit, A Tartan's Responsibility.

Aug.'20 – May.'23 **CMU Recreational Running Club** Treasurer
Dec.'20 – May.'21 **CMU Housing Services** Resident Assistant

AWARDS

*** = team competition**

2024	**3rd Place**	CMU Algorithms With A Purpose AI Contest*
	$250 Recipient	CMU Robotics Club SHRG Grant
2023	**University Honors**	CMU
2022	**1st Place**	CMU Robot Arm Autonomous Jenga Contest*
	1st Place, $1000 Prize	CMU Mobile Robots Race
2020	**Category Prize**	CMU TartanHacks*
2019	**"Ring of Honor"**	CMU Intro Comp. Biology, Research Project
	10th Place	FAMAT Programming Contest*
2018	**$2000 Recipient**	Mu Alpha Theta Grant
	Alumni	Wolfram Summer School
2017	**2nd Place**	NSU Psychology Bowl*

LANGUAGES

fluent English
conversational Mandarin

PROGRAMMING

C, Rust, Python, Java, CUDA, MLIR/LLVM, SML, Mojo, Why3, MATLAB, Mathematica, Scala, Docker, Bash, Git, LaTeX

INTERESTS

puzzlehunts, 2d animation, biking, shogi, cooking, board games, pickleball

Made with ❤ using LaTeX