Austin Jiang

+1 (604) 754-1808 | a68jiang@uwaterloo.ca | linkedin.com/in/austinboyujiang | github.com/austinboyujiang

Experience

DeepInfra Palo Alto, CA
Software Engineer Intern | **Python, PyTorch, Triton** *Incoming Jan 2027 – Apr 2027*

- Will work on inference acceleration for open-source large language models.

Huawei Noah's Ark Lab Toronto, ON
Research Intern | **Python, PyTorch** *May 2026 – Present*

- Designed layer-specific attention halting for SOTA feed-forward 3D reconstruction, achieving up to **2.45× speedup**; **first-author paper submitted to CVPR 2027**.
- Improved pose AUC by **3.0%** and reduced point/depth errors by **11.4%/2.5%** on average across 3 public datasets.
- Developed VLA-based embodied AI pipelines with agentic planning as a member of the **1st-place team at the CoRL 2026 BEHAVIOR-1K Challenge**, achieving **0.7–0.8 task Q-scores** across 2 tasks.

Multicore Lab, University of Waterloo Waterloo, ON
Undergraduate Research Assistant (Supervised by Dr. Trevor Brown) | **C++, CUDA** *Sep 2025 – Jul 2026*

- Developed a concurrent versioned pointer adapter within the benchmark, boosting throughput by ∼**2x**.
- Integrated a GPU B-Tree batching update mechanism, accelerating benchmark prefilling speed by **370%+**.
- Implemented high-throughput B+ Tree and ART Tree variants, supporting **10M-key** queries under **100+ threads**.

Wolfram Research Remote
Research Intern (Part-time) | **Mathematica** *Sep 2025 – May 2026*

- Engineered a multi-threaded halo-exchange framework for **800K+** keys across **8 cores**, cutting runtime by **56%**.
- Implemented sparse updates and dense scans via dynamic load-balancing, boosting throughput by **27%**.
- Authored a first-author paper on parallel algorithms, presented at the **Wolfram Technology Conference 2025 and 2026**.

Projects

HonkPack, HackMIT 2026 | **Python**: Architected an end-to-end per-object 3D reconstruction pipeline using SAM3, Depth Anything 3, and ShapeR from phone video, coupled with a custom 3D bin-packing algorithm.

Rewrite, TreeHacks 2026 | **Python**: Built a fully interactive real-time 3D VR game leveraging Hunyuan WorldPlay 1.5, deploying tensor parallelism on Modal across 4x H200 GPUs to boost frame rates from 1 to 12 FPS.

ADaptiv, CMU NexHacks 2026 | **Python**: Built a video object replacement pipeline in ComfyUI using Gemini, SAM3, and Wan2.1, optimizing execution on an RTX 4090 to achieve a 3x speedup .

Education

University of Waterloo Waterloo, ON
Bachelor of Computer Science (Honours) | *Major Average: 97.0%* *Expected Graduation: Aug 2028*

- **Scholarships**: René Descartes National Scholarship ($25,000), President's Scholarship of Distinction ($2,500)
- **Coursework**: Data Structures, Algorithms, Development Tools, Compilers, Object-Oriented Programming

Skills

Languages: C++, Python, C, Unity (C#), Java, Bash, TypeScript
Tools/Libraries: PyTorch, Linux, Git, CUDA, CMake, NumPy, SQL, Docker, Nginx, Flask, FastAPI
Concepts: Transformers, Large Language Models, Concurrency, Compilers, Distributed Systems, Databases

Achievements

Canadian Computing Olympiad 2026: Authored Day 1 Problem 3 (Contestant average score: 0.7/25).
Discover Citadel Program: Invited as 1 of ∼100 students (< 1% acceptance rate).
Jane Street First-Year Trading and Technology Program: Invited as 1 of ∼100 students (< 1% acceptance rate).
Generation Google Scholarship: Google's flagship global undergrad scholarship for impact in technology (1 of 55).
Canadian Computing Olympiad 2024 & 2025: Silver medalist x2 (5th and 6th out of 10,000+ participants).
USA Computing Olympiad 2024: Achieved the platinum division, ranked top 100 out of 15,564 participants.