Aditya Rajeev

U.S. Citizen | 669-237-8489 | umailadi@gmail.com | github.com/adi25r | cs.toronto.edu/∼arajeev

Education

University of Toronto Sept. 2023 – May 2027

***BS, Computer Science**, focus in Computer Systems and AI* *GPA: 3.99/4.0*

- **Relevant Coursework**: DSA, Deep Learning, OS Design, Compilers, Databases, Computer Architecture, Linear Algebra
- **Awards**: Department of CS Award ($12,000), University of Toronto Excellence award ($7,500), Dean's List
- **Teaching Assistant**, Systems Programming (Winter 2026)
- **Intending**: Graduate school (Ph.D., Computer Systems/AI) beginning Fall 2027

Technical Skills

Languages/Frameworks: C/C++, Python, Golang, CUDA, Java, Javascript, Bash, HTML/CSS, SQL
Full-Stack: React, Django, PostgreSQL, Redis, Firestore, REST APIs, WebSockets, Docker, Git, GitHub, OpenAPI
Cloud & DevOps: AWS (S3, SageMaker), GCP, Kubernetes, GitHub Actions, Linux
ML & Data: PyTorch, NumPy, Pandas, OpenCV, Scikit-learn, TensorFlow

Work Experience

Research Assistant (supervisor: Prof. Nandita Vijaykumar) May 2026 – Present

embARC Lab | *University of Toronto* *Toronto, ON*

- Writing high-performance GPU kernels for **finite field arithmetic**, accelerating multi-scalar multiplication for inference
- Optimizing workload configuration and execution to improve kernel throughput by **34x** and overall workloads by up to 6x
- Building profiling infrastructure and theoretical models for **hardware-independent optimization** across architectures

Research Assistant (supervisor: Prof. Gururaj Shaileshwar) January 2026 – May 2026

SITH Lab | *University of Toronto* *Toronto, ON*

- Developed custom CUDA/HIP kernels to parallelize attack execution, achieving a **1000x speedup** over baseline
- Reverse-engineered memory and bank layouts via timing analysis, forcing a **23,500x increase** in Rowhammer bit-flips
- Reduced exploit runtime from 21 hours to under 2 minutes; paper **GPUThor** accepted to ACM CCS 2026

Research Assistant (supervisor: Prof. Brokoslaw Laschowski) August 2025 – May 2026

Machine Intelligence Lab | *University of Toronto* *Toronto, ON*

- Engineered Transformer layers in PyTorch with customized self-attention, scaling across a **50M-parameter** architecture
- Built optimized data loaders and multi-threaded pipelines to preprocess **15,000+ Wikipedia articles**

Software Engineering Intern June 2025 – December 2025

Mezzi *San Jose, CA*

- Architected Go backend APIs for financial data aggregation, scaling throughput to serve **10,000+ user portfolios**
- Engineered an enterprise memory caching layer with Redis and Firestore, cutting P99 database access latency by **85%**
- Cut LLM inference costs by routing **70%+** of inbound queries through deterministic logic

Software Engineering Intern June 2025 – August 2025

Tigera (Part-time) *San Jose, CA*

- Deployed microservices executing distributed network policy validation for production clusters
- Shipped a RAG backend infrastructure on GCP, reducing ticketing turnaround and escalation times by **30%**
- Awarded **company-wide award** for engineering core AI integrations across validation and ticketing systems

Projects

Mini-Dynamo (Fault-Tolerant Key-Value Store) | *C, pthreads, Sockets*

- Designed a primary-secondary replication protocol with heartbeat-based failure detection and coordinator-driven recovery
- Resolved a **circular-wait deadlock** caused by synchronous cross-server forwarding by redesigning PUT handling

LatentSeas (Link) | *Java, Python, FastAPI, Sparse Autoencoders*

- Built a navigable **3D world** representing the latent space of LLMs, mapping **24,000+ features** with **Sparse Autoencoders**
- Built a **FastAPI** backend exposing **REST endpoints** to run inference and project features onto a KDE-constructed terrain

Train in a Snap (Hack the North) (Link) | *Snap Spectacles, VAPI, Flask, Heroku*

- **Semifinalist (Top 12% out of 256 teams)** — Won Top 10 in the Snap Spectacles AR track at a national hackathon
- Integrated Snap3D coordinates with Vapi Voice AI over WebSockets to build a low-latency AR training simulator

Publications

C. S. Lin, J. Qu, A. **Rajeev**, and G. Saileshwar, *"GPUThor: Amplifying Rowhammer Attacks via Non-Uniform Patterns to Exploit ECC-Protected GPUs,"* Proc. ACM CCS, 2026.

S. Zbaranska, A. **Rajeev**, S. A. Josselyn, and B. Laschowski, *"A Brain-Inspired Framework for Memory Prioritization in Neural Networks Based on Valence,"* NeurIPS, 2026 (under review).