

SACHIN VED GUPTA

☎ 437-217-4947 ✉ SachinVedGupta@gmail.com in [linkedin.com/in/Sachin-Ved-Gupta](https://www.linkedin.com/in/Sachin-Ved-Gupta) 🐙 github.com/SachinVedGupta
📖 medium.com/@SachinVGupta9 👤 SachinVedGupta.com

Education

McMaster University

September 2023 – April 2028

Bachelor of Software Engineering Co-op

Hamilton, Canada

- GPA: 3.9/4.0, **Provost's Honour Roll Medal** (perfect 4.0 GPA in 1st year), **\$10,000** Scholarship (OpenText, Rogers)

Experience

Software Engineering Intern

May 2026 – August 2026

Microsoft

Vancouver, Canada

- **PowerPoint Web**, Performance Team: Cached copilot pane for immediate pop-up during boot so no async layout shift
- Ideated and built a swarm of agents that **act as real users** in varying environments (different: task complexity, RAM, WIFI, aspect-ratio) and auto report bugs, weird shifts, and craft improvements for a more polished PowerPoint UX
- Ran the tool concurrently for 7 days, **detecting 17 optimizations** resulting in 21% faster boot times and 13 prod fixes

Data Science Intern

January 2026 – April 2026

Zipline

San Francisco Bay Area

- Designed a parameterized **Site Operations Simulator** modeling order flow, autonomous drone navigation, pickup sequences, and mission completion times to forecast throughput and evaluate infrastructure vs. delivery time tradeoffs
- Validated simulation fidelity against real operational data by running K-S tests on distributional metrics
- Developed a **Codex-powered Simulation Skill** that translates Slack business queries (optimal fleet size for X demand?) to JSON configs, runs the simulator, and delivers actionable data-driven insights used by 3 Operations teams

Software Engineering Intern

September 2025 – December 2025

Shopify

Toronto, Canada

- Built an LLM-based Relevancy Judge that explains and filters unfitting products; fine-tuned it with Toloka labels and **distilled it into a 1B param model** for low-latency scoring of 10M+ recs, removing irrelevant items before the feed
- Shipped a debugging tool to detect “weird” Shop App recommendations and trace their origin in the **Recs ML pipeline, thus demystifying “why”** they came and enabling fast pain-point resolution (used by 80+ engineers)
- Created a Multi-Tier Caching system (local > remote > bigtable) for **11x faster filtering** of advertised products

Machine Learning & Open Source Lead

April 2025 – Present

Google Developer Student Clubs ↗

Hamilton, Canada

- Led a 10-member team building an open-source AI Browser Agent powered by user input, memory, and vision models
- Built a perception-planning-action pipeline combining LangGraph intent recognition with a **fine-tuned Grounding DINO model** trained on UI-to-meaning mappings, improving site understanding and autonomous action accuracy.

Machine Learning Lead & Co-Founder

January 2025 – Present

McMaster Aerial Robotics & Drone Club ↗

Hamilton, Canada

- **Fine-tuned YOLOv11** on a custom balloon/bucket dataset to better identify fire markers (+2.6% mAP, +4.4% recall)
- Architected a camera-to-NVIDIA Jetson Nano-to-flight controller pipeline and optimized it by deploying a TensorRT object detection model with **FP8 quantization**, achieving real-time, **low-latency inference with a 3.2× speedup**
- Led a 20-member team across autonomous drone development and Gazebo simulation testing to prepare for SUAS 2026

Impactful Projects

AWS - Llama LLM Inference Optimization Contest ↗ | AWS (Neuron Kernel Interface, Trainium EC2), Pytorch, Numpy

- Optimized performance by re-implementing fundamental transformer inferencing operations like Matrix Multiplication and Scaled Dot Product Attention to **utilize NKI** which is more efficient for the Trainium EC2 chips it's running on
- Utilized the benchmarking tool to monitor performance metrics (accuracy, NKI flops, throughput) - **now 1.13x faster**

LearnBridge - AI Agent for Students ↗ | Google Agent Development Kit, GCP, Python, Google Classroom & Calendar API

- Architected **ReAct-based multi-agent workflows** with memory/calender integration for 80% faster course info access

Skills

Programming: Python, C++, SQL, JavaScript, Linux, Docker, PostgreSQL, CypherQL (Neo4J), Google ADK, LangChain

Libraries: PyTorch, Tensorflow, Triton, CUDA, Flask, Pandas, Numpy, Matplotlib, Scikit-Learn, OpenCV, Hugging Face

Certifications: AWS Certified AI Practitioner, AWS Certified Cloud Practitioner, Nokia Bell Labs - Data Driven 5G ↗

Highlights: Won HackMIT (1000 competitors) ↗, Ignition Hacks Speaker ↗, AI articles ↗, Attended ASPLOS & ICSE