

Malcolm Shi

[linkedin.com/in/malcolm-shi](https://www.linkedin.com/in/malcolm-shi) | github.com/malcolms123

TECHNICAL SKILLS

Languages: Python, C/C++, Rust, Go, Java, JavaScript/TypeScript, SQL, Kotlin, HTML/CSS
Frameworks & Libraries: PyTorch, TensorFlow, vLLM, CUDA, Spring, React, Angular, Node.js, Express.js
Developer Tools: Git, Docker, GitHub Actions, NVIDIA Nsight, PyTorch Profiler, Selenium, Sikuli, Android Studio
Platforms & Technologies: Modal, AWS (EC2, S3), Prometheus, OpenGL, Vulkan, DirectX, MongoDB, MySQL

EXPERIENCE

Huawei Technologies Co., Ltd.

Jan 2025 – Apr 2025

Software Engineer Intern | Software-Hardware Optimization Lab

Markham, ON

- Engineered **4+** neural rendering models using **3DGS** and **NeRF** in **C++** and **Python** with **CUDA**, **OpenGL**, **PyTorch**, and **TensorFlow**, reducing GPU graphics pipeline latency across Huawei's embedded systems by **56%**
- Implemented **2+** **Vulkan** compute shaders in **Android APKs** on embedded **Linux** using **Android Studio** and **Kotlin**, improving firmware-level performance benchmarks on mobile operating systems by **33%**
- Leveraged **DirectX 12 Work Graphs** in **C++** and **HLSL** to build an **occlusion culling** system profiled using **NVIDIA Nsight Tools**, optimizing memory usage to accelerate task execution by **24%**

Nest Wealth

May 2023 – Aug 2023

Software Engineer Intern | Pro 1.0 Platform

Toronto, ON

- Maintained **3+** **RESTful API** endpoints on **EC2** instances with **Docker**, integrating **DocuSign** form logic and least-connection **load balancing** to query **MySQL** databases and increase system throughput by **38%**
- Built an interactive donut chart in **TypeScript** using **Angular** components to visualize asset allocation, applying a **weighted decision matrix** to process client questionnaire data stored in **MongoDB**
- Conducted security reviews on **10+** third-party applications, performing **fault injection** and **penetration tests** to identify vulnerabilities, assess security controls, and support **SOC 2 compliance**

OpenText

Sep 2022 – Dec 2022

Software Engineer Intern | Content Server Core

Waterloo, ON

- Executed monthly **end-to-end**, **integration**, and **functional tests** on **8+** core enterprise modules by enhancing **Selenium** automation scripts that interact with complex **HTML** structures, achieving **98%** test coverage
- Developed **JUnit test report** generators that automatically interpret stack traces and summarize error logs into **XML** documents following test suite execution, accelerating debugging processes by **42%**
- Streamlined **regression testing** workflows by writing **Sikuli** scripts that revert **VMware** virtual machines and restore **SQL Server** database snapshots, enabling fully automated test cycles

PROJECTS

LLM Inference Optimization Lab 🧠 | GPU Inference Performance Benchmarking

Oct 2026

- Architected GPU-accelerated **model serving** infrastructure using **vLLM** and **Modal**, deploying **Qwen3-4B** on **NVIDIA A10** and **L4** GPUs with **BF16** precision, chunked prefill, and authenticated streaming APIs
- Constructed an asynchronous benchmarking framework in **Python** using **asyncio** and **httpx**, executing **1,000+** measured requests to evaluate concurrency, scheduling token budgets, and **TTFT**-throughput tradeoffs
- Investigated **KV-cache** prefix reuse through **A/B experiments** across **500+** requests, analyzing **Prometheus** telemetry to observe up to **77%** lower median TTFT and **89%** less prefill computation
- Profiled **CUDA kernels** using **PyTorch Profiler** to correlate cache reuse with changes in GPU execution patterns, measuring **36%** lower aggregate kernel duration in a controlled diagnostic

EDUCATION

University of Waterloo

Jun 2026

BASc in Computer Engineering (Co-op), With Distinction

President's Scholarship

CERTIFICATIONS

Scrum Alliance

Nov 2022

Scrum Foundations — Agile methodology