Ahamed Mohamed Wajibu

Mississauga, Ontario | a226moha@uwaterloo.ca | Github | Linkedin

EDUCATION

UNIVERSITY OF WATERLOO — Waterloo, ON

Honours Bachelor of Mathematics — Expected May 2028

SKILLS / TECHNOLOGIES

- **Languages:** C++, Rust, CUDA, Python, Go, C, C#, TypeScript, JavaScript
- **ML & Inference:** PyTorch, TensorFlow, cuBLAS, OpenCV, YOLO, LangChain, Ollama
- **Backend & Cloud:** ASP.NET Core, FastAPI, Flask, Node.js, React, SQL Server, PostgreSQL, Docker, Nginx, Linux, Git, Azure, AWS, Microsoft Fabric

EXPERIENCE

TD Bank — Toronto, ON

Software Engineer Intern — *January 2027 – April 2027*

- Design, develop, and deploy enhancements to enterprise applications, with a focus on debugging, incident resolution, and system reliability.
- Work with business partners to define requirements, perform unit and integration testing, and support code reviews and deployment planning.

RBC Borealis — Toronto, ON

*AI Researcher Fellow | **1 of 10** teams selected across Canada* — *October 2026 – Present*

- Developing an NFL impact-detection pipeline with YOLO, ByteTrack/DeepSORT, and PyTorch temporal models to track players and helmets, classify impacts, and estimate athlete exposure risk.

Mitsubishi — Mississauga, ON

AI Engineer Intern — *January 2026 – May 2026*

- Owned an internal RAG platform end to end over 2,000+ aircraft technical manuals, serving **40 engineers** (LangChain, Azure AI Foundry, PostgreSQL, Microsoft Fabric); ingestion pipelines extract structured fields, per-page summaries, and source metadata.
- Answered **98 of 100** questions correctly on a golden set written by senior engineers; both misses required image understanding.
- Self-hosted Gemma 4 and quantized open-source models via Ollama behind an Nginx reverse proxy on an RTX PRO 6000 (Blackwell) server, keeping sensitive documents on-prem.
- Added tool calling (document search, page retrieval, SQL keyword search, web search) with inline citations.

Mitsubishi — Mississauga, ON

Software Developer Intern — *September 2025 – December 2025*

- Replaced manual SQL Server scripts with ASP.NET Core/C# pages for QA call-up, aircraft line status, job updates, and Syspro service workflows used by hundreds of production workers.
- Refactored Entity Framework services, LINQ queries, and SQL Server access patterns, cutting request latency by **~30%** and reducing redundant database calls across production modules.

PROJECTS

Forge | *Rust, C++, CUDA, Go | GitHub*

- Built an LLM inference runtime **from scratch** for Qwen2.5-0.5B and SmolLM2-135M: transformer decode, KV cache, RoPE, RMSNorm, and attention kernels in C++/CUDA, a Rust runtime, and an OpenAI-compatible streaming server in Go.
- Implemented **continuous batching**, **a paged KV cache**, and **CUDA Graphs** for GPU serving; on an NVIDIA L4, aggregate decode throughput scales from 22.0 to 41.6 tok/s across batch sizes 1 to 8.
- Raised CPU decode from **0.44 to 23.0 tok/s (~52×)** through multithreaded kernels and weight-only INT8 quantization (group size 64), which also cut memory use from 2,404 to 1,020 MiB.

Block-GTQ WebGPU | *TypeScript, WGSL, WebGPU, PyTorch | GitHub*

- Ported Block-GTQ's RoPE-aware K-cache quantization (Liang et al.) to WebGPU/WGSL, with mixed-width bit allocation, bit-packed storage, and in-kernel key decompression; runs SmolLM2-135M and Qwen2.5-0.5B in the browser.
- Cut persistent K-cache memory **2.24×** at 4K context (45.35 → 20.24 MiB) and validated GPU attention against a Python reference across **120 strict comparisons** (max error 2.4e-6).