Muhammad Ibrahim Khalid

mikhalid@connect.ust.hk | Hong Kong
github.com/1brah1 | 1brah1.github.io | linkedin.com/in/muhammad-ibrahim-khalid-475621382

PROFILE

Final-year HKUST electronic engineering student working on efficient ML systems: running large models under tight memory and latency budgets (edge devices, wireless GPU offload, KV-cache placement).

EDUCATION

The Hong Kong University of Science and Technology (HKUST) | Expected 2027

BEng Electronic Engineering. Thesis: *Hot/Cold Weight Swapping and KV-Cache Placement for LLM Inference on Wireless GPU Computing Infrastructure*, advised by Prof. Zili Meng

RESEARCH

Final-Year Thesis, HKUST | Advisor: Prof. Zili Meng | Sep 2026 – Present

- WiCi (arXiv:2608.24204) runs a phone's CUDA calls on a nearby server GPU over Wi-Fi. I am designing what stays in an H800's 80 GB when the model does not fit: hot layers and KV cache in VRAM, cold layers prefetched from DRAM/NVMe.
- Evaluating against fully resident, no-policy and FlexGen baselines on TTFT, inter-token latency and decode tokens/s. HK$22,500 compute funding.

Gray Scrubs AI (grayscrubsai.com) | Research Intern | Sep 2026 – Present

- Research on generating rare, safety-critical events.

Undergraduate Research (UROP1100, UROP2100), HKUST | Feb 2026 – Present

- Added Omi Glass support (BLE, microphone input) to github.com/hkust-spark/xg-glass-sdk, merged into main; surveyed GPU remoting and offloading (rCUDA, FlexGen, PIPO) to scope the thesis.

INDUSTRY EXPERIENCE

LWK + Partners (lwkp.com) | Workflow Automation Intern | Jun 2026 – Aug 2026

- Cut Hong Kong PNAP APP-152 building-separation checks from 10–30+ min to 30–90 s per run with a ~17k-LOC pyRevit extension and a two-stage plane search.
- Proof of concept: drove Autodesk Revit 2024 from Cursor over MCP (github.com/mcp-servers-for-revit).

SUSTAINALPHA (app.sustainalpha.pro) | Core Engineer | Feb 2026 – May 2026

- Built a Monte Carlo and mean-variance risk forecasting engine on EODHD data, plus an LLM news summariser, for an HKSTP-backed risk platform.

Gray Scrubs AI | Hardware Engineer Intern | Mar 2026 – May 2026

- Maintained the NVIDIA Jetson fleet and device access behind multi-country pilots and paper experiments.

PROJECTS

Real-Time VQA on Jetson Orin NX 16 GB | github.com/1brah1/VQA_Project-COMP4901D-

- Cut mean latency 747 → 346 ms (2.2×) by compressing SigLIP vision tokens 576 → 36, with accuracy up 25 → 50%; 80.75% validation accuracy on 1,621 held-out images.
- Wrote CUDA/PyBind11 token-pooling kernels: 6–10% faster per op but end-to-end flat (551.7 vs 550.4 ms), since generation is over 80% of latency. github.com/1brah1/jetson-vqa-cpp-extensions

SignalProof: Backtest Leak Auditor | github.com/1brah1/SignalProof

- Built walk-forward backtests with purge/embargo, a 46 bps cost model, Deflated Sharpe and PBO; injected look-ahead leaks fail at runtime (PBO 0.90 vs 0.50 clean). A local 7B LLM writes memos from metrics only, never prices.

Edge RAG Benchmark | github.com/1brah1/RAG_PROTOTYPE

- Compared local Ollama and NVIDIA NIM RAG on an RTX 4060 8 GB over 10-run trials: 14.94 tok/s (local q4) vs 33.66 tok/s (NIM, 0.771 s mean TTFT).

More projects: 1brah1.github.io

AWARDS & FUNDING

- HK$22,500 thesis compute funding, HKUST SuperPOD (2026)

SKILLS

ML systems: PyTorch, CUDA, PyBind11, LoRA, token compression, KV-cache and weight offloading (FlexGen, PIPO), RAG, Ollama, NVIDIA NIM
Software: FastAPI, React, TypeScript, SQLite, PostgreSQL, Docker, AWS EC2, GitHub Actions, pyRevit / Revit API
Languages: Python, C++, CUDA, C#, TypeScript, SQL