

Richael

Auckland, New Zealand | huangd6666@gmail.com
[dh8116.github.io](https://github.com/dh8116) | github.com/dh8116 | x.com/RicaV42

PROFILE

A student building GPU kernels, fine-tuned language models, and full-stack applications with real users. Nine Triton kernels benchmarked against PyTorch over nine consecutive weeks, each published with its numbers; a fine-tuned Qwen3-14B serving in production behind vLLM, leading the provider chain and answering faster than the model it replaced.

EDUCATION

Year 11, Auckland, New Zealand. Expected graduation 2028

Relevant coursework: IB Math AA HL, NCEA level 2 (planning for AP calculus & cs exams next year)

TECHNICAL SKILLS

GPU & ML: Triton, PyTorch, Python, GPU benchmarking and profiling, LoRA fine-tuning, vLLM, Modal, eval harnesses

Full-stack: TypeScript, Node.js, Express, Next.js, React, Supabase/Postgres, Web Audio API, Stripe

Practice: Git, GitHub Actions, Playwright end-to-end testing, web security hardening (CSP, XSS)

GPU SYSTEMS & OPTIMIZATION

One Triton kernel per week, benchmarked against PyTorch.

- Cut fused cross-entropy to 15.90 ms against PyTorch's 24.03 ms at vocab 131,072 (1.51x), at 1.67x less peak memory, by collapsing five live copies of the logits tensor to three.
- Raised fused softmax from ~55 GB/s to ~230 GB/s, and established T4's ~240 GB/s practical ceiling by measurement.

DISTRIBUTED AI INFRASTRUCTURE & DISPATCH

- Fine-tuned Qwen3-14B with LoRA on a 1,526-row corpus spanning five generation paths in six languages; achieving full performance for \$1.30 of compute.
- Deployed optimized cold-start vLLM architecture, leading the production chain at 2.3–4.1s against the original 5.7s model.
- Built warm-gate routing: the gateway probes for a warm container and the probe doubles as the warm-up, so a 108 s cold start never reaches a user.
- Fine-tuned MusicGen-small for bar-locked stems, built the eval harness scoring clips as production.

PROJECTS

Soulor AI — soulor.app

- A regional-variant companion and multi-perspective simulation app with cross-session memories and full relationship stages. Also serving as a great tool for daily reminders and wellness enhancement.
- Cross-Mode Engine | Relationship Staging | Bias Adaptation | Stateful Sampling

VNportal — vnportal.vercel.app

- A vinyl and music community platform for creation, recreation, and discussion, promoting innovative music creation with AI-assistance.
- Controllable Audio Synthesis | Onset Charting | Source Cascade | Stem Synthesis

Personal website — dh8116.github.io

- A central hub sharing my skills, projects, current development stages, and blog writings.
- Next.js static export | Hreflang Sitemap | Postbuild CSP.

RESEARCH & WRITING

- Leading a five-person-team paper on RLVR verifier persistence: when a persistently wrong verifier harms GRPO training. Targeting COLM/TMLR.
- Publishing general and kernel writeups at dh8116.github.io/blog; co-authored several modelling papers for HiMCM and IMMC.

AWARDS & VOLUNTEERING

- Score 10/15, AIME (American Invitational Mathematics Examination) | Honor Roll (Top 5%), AMC 12 (The American Mathematics Competition 12) | High Distinction, AUSAMC (Australian Mathematics Competition)
- Honorable Mention, HiMCM (High School Mathematical Contest in Modeling)
- Honorable Mention, IMMC regional (International Mathematical Modeling Challenge Regional Round)
- Volunteer, Mustang Math Community — Technology and Curriculum Development