Richael

Auckland, New Zealand | huangd6666@gmail.com
dh8116.github.io | github.com/dh8116 | x.com/RicaV42

PROFILE

A student engineer who ships: GPU kernels benchmarked against PyTorch, a fine-tuned language model served in production, and a full-stack app with real users that I build and run alone.

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.

AI INFRASTRUCTURE

- Fine-tuned Qwen3-14B with LoRA on a 1,526-row corpus spanning five generation paths in six languages, for $1.30 of compute.
- Deployed an optimized cold-start vLLM architecture that led 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.

EXPERIENCE

Founder and sole engineer, Soulor — soulor.app
June 2026 – present

- Built and run an AI app for rehearsing hard conversations: Node.js/Express on Vercel, Supabase Postgres with chat encrypted at rest, Stripe billing, live in six languages.
- 15 users and Stripe billing live; every chat goes through my own multi-provider gateway (Groq, Gemini) that benches a provider at its per-model quota and alerts on outages; 200+ pull requests.
- Cut reply latency from 7.9s to 2.0s; logged every failed chat, voice and translation call, so a failed turn shows up in the database instead of disappearing.
- Every commit runs the full Playwright end-to-end suite before it lands. Built a listening voice call (browser speech recognition, text-to-speech that plays the first sentence first) in one evening after a tester reported the old one never answered.

VNportal — vnportal.vercel.app
May 2026 – present

- Music community platform; fine-tuned MusicGen-small for bar-locked stems and built the eval harness that scores clips as production.

RESEARCH & WRITING

- Leading a five-person team paper on RLVR verifier persistence (Aug 2026 – present): when a persistently wrong verifier harms GRPO training. Draft due by the end of 2026, targeting COLM/TMLR.
- Kernel and general writeups at dh8116.github.io/blog.

EDUCATION & AWARDS

Year 11, Auckland, New Zealand. Expected graduation 2028. IB Math AA HL, NCEA Level 2

- Score 10/15, AIME | Honor Roll (Top 5%), AMC 12 | Honorable Mentions, HiMCM and IMMC regional