

Yoshio Nomura

Ho Chi Minh City, Vietnam · yoshio.nomura@student.uts.edu.au

github.com/UniverseScripts · linkedin.com/in/yoshio-nomura-b3219438b · yoshio-nomura.vercel.app

Backend and AI-infrastructure engineering: routing, retrieval, and gating expensive compute behind cheap fast paths.

Education

University of Technology Sydney — Bachelor of Artificial Intelligence

Faculty of Engineering & IT · Expected 2028 · GPA 3.7 / 4.0

Experience

Pyrium — Co-founder · October 2026 – present

A three-person brand building private, cost-aware agentic AI; in development, not incorporated.

pyrium-private-ai.vercel.app

FlyRank AI — Backend AI Engineer, Internship (Remote) · July – September 2026

- Completed both the Backend AI Engineering and AI Fluency tracks; certificates issued 16 September 2026.

Projects

Ignis — maintenance intelligence for power plants

Team of 3 — backend AI engineering, benchmarking and system verification · October 2026 · AMD Developer

Hackathon, Act III (lablab.ai)

Python · SQLite · MinerU OCR · DeepSeek-V4-Flash (AMD-hosted) · React

- Work orders are imported with explicit column mapping, reviewed by a person, and issued as evidence-linked PDF reports; facts are computed deterministically and a hosted model drafts the explanation from a scoped evidence pack.
- Developed the work-order pipeline, the hosted-inference layer, document ingestion and the benchmark harness.
- Explanation prompt v3 produced a valid draft in **21 of 24** runs against **14 of 24** for v2, with median model time **10.3 s** against **21.3 s** — 12 synthetic scenarios, two runs each, preliminary.
- OCR mean character error **0.000** on text and table documents, **0.617** on diagrams (synthetic sets).

The Lantern — voice ordering agent for restaurants (tested in English and Spanish)

Team of 3 — dialogue logic and reliability · September 2026 · AssemblyAI Voice Agent Hackathon (lablab.ai)

Python · FastAPI · WebSockets · AssemblyAI streaming · Qwen3 4B via Ollama · Kokoro TTS

- A small local model returns one intent per turn; code resolves references, applies order changes and checks each against the live menu before anything reaches the kitchen.
- Team result: **10 of 10** six-turn order runs correct, and **5 of 5** full voice sessions with the correct basket (synthetic speech streamed in real time).
- Cut average time to first audio from **42.9 s to 5.8 s (-86%)** in the first version by replacing a fixed rate-limit delay with a sliding window — measured on that version, since replaced.

Pulsemind — critical-care telemetry, research prototype

Team — full-stack AI engineer · May 2026 – ongoing

Python · XGBoost · PyTorch · scikit-learn · React

- Team built an asynchronous ICU telemetry processor in which an XGBoost classifier scores incoming stream events and gates higher-cost LLM rationalisation calls, so they fire only on detected anomalies.
- About **43 ms** to score each reading, including its TreeSHAP explanation (demo service, 7 readings), against roughly 15 s for the LLM call it gates.
- Trained on the credentialed MIMIC-IV (PhysioNet) de-identified ICU dataset under its data use agreement; demonstrations run on a synthetic derivative built from it. No clinical deployment.

Develarper — two-tier LLM task router

Team of 4 — LLMOps · July 2026 · AMD Developer Hackathon, Act II

Python · Qwen 2.5 3B via Ollama · FastAPI · Fireworks

- A local Qwen 2.5 3B classifier scores each task's difficulty and dispatches hard tasks to a cloud API while easy tasks stay on the local model.
- On the team's routing evaluation, **711 of 800** tasks went to the correct tier (88.9%) against team-written labels.

- Separately evaluated on **200 tasks** (100 factual, 100 summarization): **93.00%** correct (186/200), scored by similarity and Jaccard overlap against expected answers — not human review and not an LLM judge — at **505.1 ms** mean latency running sequentially. Six of the 200 prompts contain unrendered template placeholders and are unanswerable as posed.
- The team rejected a dedicated BERT encoder classifier: holding a second resident model would not fit under the harness's memory ceiling, so the router reuses the generative model already loaded.

Roomie — student roommate and apartment matcher

Contributor, team of 4 — DevOps & Backend Engineering · April 2026 · GDGoC National Hackathon 2026 (Hanoi), team Hackaphobia

Python · FastAPI · Firestore · Vertex AI embeddings · React 19 · Vite · Google Cloud Run

- Onboarding survey and swipe interface matching on structured fields, with an optional free-text bio path using Vertex AI `text-embedding-004` embeddings and cosine similarity.
- **~50 real users** onboarded and swiped at demo day; **~12 ms** average request latency on the structured matching path, excluding embedding generation.
- Real-time chat over WebSockets, with an in-memory connection manager mapping each user to their active sockets so a user signed in on multiple tabs receives messages on all of them.
- Rejected an MCP-plus-n8n agentic pipeline as over-scoped for a four-person team with no operational runway.

Weatherise — weather-intelligence pipeline

Team — LLMOps & Backend AI Engineering · 9–11 June 2026 · Vietnam AI Open Hackathon (NVIDIA / OpenACC)

Python · Qdrant · NVIDIA Earth-2 surrogates · Nemotron Ultra · MCP

- Ingestion combining REST meteorological sources with NVIDIA Earth-2 surrogate models, feeding a multi-agent chain on Nemotron Ultra, with a Qdrant vector database supporting retrieval-augmented generation over the ingested data.
- Team selected as one of **10 from roughly 100 registrants** to compete.

Skills

Languages — Python, TypeScript, SQL, Java

Machine learning — XGBoost, PyTorch, scikit-learn

Backend — FastAPI, REST APIs, Firestore, async patterns, WebSockets, PostgreSQL, Redis

AI infrastructure — local LLM serving via Ollama, Vertex AI embeddings, Qdrant + RAG,

Nemotron Ultra multi-agent chains, NVIDIA Earth-2 surrogates, MCP

Frontend — Next.js, React, Tailwind, Vercel

Tooling — Git / GitHub, GitHub Actions, Docker

Credentials

- **Backend AI Engineering — Certificate of Completion** — FlyRank AI Internship Program — September 2026 (verify at internship.flyrank.ai/verify · ID FR-D11-62286-4A975)
- **AI Fluency — Certificate of Completion** — FlyRank AI Internship Program — September 2026 (verify at internship.flyrank.ai/verify · ID FR-D11-15C28-A543D)
- **Dean's List 2026** — University of Technology Sydney — July 2026
- **Certificate of Attendance** — Vietnam AI Open Hackathon (NVIDIA / OpenACC) — June 2026 (team selected as one of 10 from ~100 registrants to compete)
- **Top 30 Finalist Award** — GDGoC National Hackathon 2026, Hanoi, team Hackaphobia — May 2026
- **Next.js App Router Fundamentals** — Vercel — February 2026
- **Generative AI with Large Language Models** — DeepLearning.AI & AWS, via Coursera — January 2026
- **AWS Cloud Practitioner Essentials** — Amazon Web Services — foundational completion — December 2025