

LƯƠNG QUỐC DŨNG

AI Engineer | LLM Systems Builder

0965 801 265 • lqdung1375@gmail.com • Thanh Trì, Hà Nội

Portfolio: lqdung-portfolio.vercel.app • GitHub: github.com/quocdungTlu

OBJECTIVE

AI Engineer focused on shipping LLM systems past the demo stage — agents, RAG, evaluation, guardrails, telemetry and cost control.

TECHNICAL SKILLS

AI Engineering: LLMs (Claude, OpenAI) • RAG (LangChain, LangGraph, ChromaDB, hybrid dense+BM25) • AI Agents & Tool Calling • Prompt Engineering & Evaluation • Vision-LLM OCR

Backend: Python, FastAPI, REST APIs, SSE Streaming, SQLAlchemy, Alembic, Node.js, Pytest

Frontend: React, Next.js 14, TypeScript, TailwindCSS

Data: PostgreSQL (Supabase), MySQL, ChromaDB, Pandas, NumPy, Scikit-learn

Cloud & DevOps: Docker, Fly.io, Vercel, Render, GitHub Actions (CI), Git/GitHub

Languages: English — technical reading, documentation, and basic professional communication

EXPERIENCE

AI Engineer Intern — VinSmart Future, AI Models & Platform Division

06/2026 – 08/2026

- Worked across two internal AI platforms on agent architecture, plugin and tool safety, deployment governance, and LLM unit economics.
- Rebuilt a creative evaluation rubric after v1 passed 49/49 — a saturated scale is a measurement failure, not a success — into hard gates, anchored criteria, and forced ranking.
- Showed through controlled testing that a single user instruction (“remember X”) escalated unverified claims into long-term memory; proposed an authority-ranked write gate.
- Found two failure modes during implementation: a plugin able to issue its own approval receipt, and a lost update where the server returned HTTP 200 while overwriting a newer revision.

System Operations Collaborator — Viettel Customer Service

06/2025 – 05/2026

Operated V-Office, an enterprise document system for the Ministry of Justice and the Central Party Office; managed user permissions.

Front-end Developer Intern — OnTik

05/2024 – 12/2024

Built LoanCRM, an enterprise loan-management system; turned business requirements into UI and flows, and integrated REST APIs.

PROJECTS

TripNest AI — Production AI Travel Agent · Solo developer · Graduation project

05/2026 – 07/2026

11 agent tools · 255 automated tests · 20/20 RAG eval · 9/9 end-to-end booking flows · ~\$0.17 LLM cost / user / month

Live: tripnesttravel.online • Source: github.com/quocdungTlu/C2-App-031

FastAPI • Next.js 14 • Claude Haiku 4.5 / GPT-4.1 mini • LangChain • ChromaDB • Supabase PostgreSQL • Docker • Fly.io • Vercel

- An AI travel agent that takes a user from chat to a paid booking in one conversation: RAG-grounded recommendations, itinerary planning, VietQR payment and SSO — live with real users.
- Architected a multi-provider LLM layer behind a feature flag with a canonical message format — swapping providers is a one-line env change, enabling safe A/B testing and rollback.
- Cut cost and context bloat (prompt caching, summarization, tool-result pruning, TTL-evicted sessions) and tracked token/cost/latency, tool success and the booking funnel in an admin dashboard.
- Hardened the money path after red-teaming: payment confirmation moved out of the LLM to a server-verified sentinel, and PostgreSQL advisory locks make booking atomic under concurrent checkout.

TTHC Assist — AI Assistant for Public Administrative Procedures · Core contributor, 4-person 48h hackathon

07/2026

Demo: c3-tthc-assistant.onrender.com • Source: github.com/quocdungTlu/UBNDAI

FastAPI • LangGraph • Claude Haiku 4.5 • Chroma + BM25 hybrid RAG • Vision-LLM OCR • PostgreSQL • Docker

- Owned the vision-LLM OCR pipeline: structured output with bounding boxes, image-hash caching, and a confidence threshold routing low-certainty scans to human review.
- Designed a three-set evaluation (in-catalog / out-of-catalog / held-out) instead of one blended metric, measuring wrong-procedure commits separately: 0/60, at ~\$0.016 per case.
- Moved guardrails to the schema boundary: Pydantic validators reject LLM output carrying error-severity findings, and checklists generate only from the legal catalog.

agent-evaluator — LLM-as-Judge Evaluation Harness · Solo · from an internship task

08/2026

Source: github.com/quocdungTlu/agent-evaluator

Node.js • Claude API • SHA-256 corpus hashing • Wilson confidence intervals

- Across 9 runs a grader given the identical criteria vector reported different totals, so the normalizer recomputes every score deterministically and fails closed on a malformed vector.
- Attacked the grader itself (prompt injection, forged authority, oracle poisoning), kept security flags separate from the quality score, and reported Wilson intervals for small-n results.

Additional projects: DPO/ORPO alignment lab · multi-agent orchestration · data pipeline observability · React e-commerce storefront.

EDUCATION & LEADERSHIP

Thang Long University — B.Sc. in Information Technology

2021 – 2025

Applied AI Intensive Program (AI Thực Chiến, 05/2026 – 08/2026) • Vice President, Karatedo Club (2024 – Present)