

Le Tran Minh Dat

AI-NATIVE FULL-STACK DEVELOPER

Ho Chi Minh City, Vietnam +84 369 744 023 datltmse@gmail.com letranminhdat.com GitHub LinkedIn

Full-stack engineer with six production systems shipped in the past year, solo or leading small teams: a bank-wide LLM gateway, a VAS-compliant accounting ERP used daily by **20 accountants**, an on-prem RAG advisor for a 19-branch academy, a real-time patient-monitoring vision system, a voice ordering agent and an AI trading platform delivered in **20 days**. I own the full stack on every one: data model, backend, frontend, infrastructure and deployment.

Experience

LLM Gateway Engineer 2026 – Present

ENTERPRISE CLIENT (BANKING) · OUTSOURCED CONTRACT · 2-ENGINEER TEAM

OWNED Architecture and the large majority of the custom layer, including its test suite and technical documentation; one other engineer contributed alongside.

- Architected a bank-wide LLM gateway on LiteLLM as the single control point between every internal application and major LLM providers. Key decision: isolated all bank-specific code in a separate 95-file extension layer with a single unlock hook into the core, so upstream releases merge cleanly.
- Delivered enterprise access control: per-user, per-team and per-project budgets, SSO integration with the bank's identity provider, role-based permissions and an immutable audit log, alongside bidirectional guardrails that mask or block sensitive data such as phone numbers, national IDs, account numbers and cloud credentials.
- Engineered cost governance: rather than trusting the gateway's own estimate, per-call spend is reconciled against actual AWS invoices and tagged for department-level allocation; added automatic model failover, budget-aware downgrade to cheaper models, and OAuth-based access to developer AI subscriptions, a capability absent upstream.

AI Agent Consultant Engineer 2025 – Present

HƯỚNG NGHIỆP Á ÂU · FREELANCE · SOLO, END-TO-END

OWNED Everything: backend, frontend, prompt pipeline, data model, infrastructure and deployment. Client supplied CRM access and end-user feedback.

- Built a 24/7 RAG enrollment advisor for a **19-branch** academy (**100,000+** learner base), handling **202 leads** and ~1,800 messages to date, with an advisor cockpit for human-in-the-loop review. Hybrid retrieval fuses local ONNX embeddings (CPU, zero API cost) with Vietnamese full-text search; a CRM-sourced allow-list and topic pinning prevent hallucinated courses.
- Instrumented every LLM call (tokens, cost, latency, errors) and migrated providers on measured A/B evidence rather than guesswork.
- Deployed on-prem inside the client's VPN-only LAN under systemd, integrated with the internal CRM, and hardened it with auto-restart plus a 9-suite E2E test gate after a production incident.

Accounting Application Developer 2025 – Present

ENTERPRISE CLIENT (EDUCATION) · FREELANCE · 2-ENGINEER TEAM

OWNED Full backend, frontend and database design, delivered jointly with one other engineer.

- Built a VAS-compliant ERP (.NET 10, PostgreSQL, React) in daily production use by **20 accountants**; authored the SRS and **164 regulatory test cases** it is audited against (audit passed, currently in trial period).
- Shipped double-entry ledger logic and a full audit trail through Docker + AWS EC2 + GitHub Actions CI/CD, ahead of the **1 Jan 2026** regulatory deadline.

AI Team Lead, Computer Vision Jun 2025 – Jan 2026

WHAMMY TECH · CONTRACT · LED 4 ENGINEERS

OWNED All AI work (detection models, inference pipeline, LLM reporting) and technical documentation; advised the backend, frontend and mobile engineers.

- Led all AI development on IPBMS, a real-time fall/seizure detector on live RTSP streams (YOLO pose estimation plus LLM incident reports); designed a **5-frame temporal validation** strategy to suppress single-frame false alarms.
- Deployed and operated on-premise, live on **2 production cameras**.

AI Agent Engineer 2025 – Present

ĂNG Ê COFFEE · CONTRACT · SOLO, SELF-HOSTED

OWNED Everything, from agent logic to the self-hosted server it still runs on.

- Built a real-time voice ordering agent (LiveKit, TTS) in daily production use, with hand-built multi-LLM failover and a self-learning RAG loop.
- Rolled out in phases (text to voice to real-time voice); self-hosted and still operates the stack.

Freelance Software Engineer, Team Lead Oct 2025

NEURADAO · LED 3 ENGINEERS

OWNED AI trading logic, AI chatbot and trading-automation backend; one backend and one frontend engineer on the rest. Deployed on AWS with PostgreSQL; project closed on delivery.

- Led a **3-engineer team** from zero to a production AI trading platform in **20 days**, setting the architecture and delivery plan and owning AWS deployment through go-live.

Skills

AI WORKFLOW	AI-assisted development across architecture, code generation, refactoring, debugging and testing; prompt & context engineering
FULL-STACK	TypeScript, Python, C#, SQL · .NET 10, Node.js/NestJS, React + TanStack · PostgreSQL, pgvector, Redis · REST API design, relational data modeling, transactional integrity
LLM / AI	Multi-provider routing & fallback, hybrid RAG, guardrails, human-in-the-loop, YOLO/OpenCV, RTSP
CLOUD & OPS	AWS EC2, Docker, GitHub Actions CI/CD, systemd, on-prem/VPN deployment · observability (per-call cost/latency/error logging), E2E test gates, RBAC & audit logging
SPOKEN	Vietnamese (native), English (professional working proficiency)

Education

B.Eng. Software Engineering

FPT UNIVERSITY · HO CHI MINH CITY

GPA 3.47 / 4.0 · Graduated 2026

Awards

Top 4 — Viettel AI Race

HACKATHON · VIETTEL GROUP

Reference

Nguyễn Anh Khoa

CHIEF TECHNOLOGY OFFICER · WHAMMY TECH

+84 969 473 668 khoanalio1412@gmail.com LinkedIn