

Nguyen Dang Minh Quan

📞 0967106719 — ✉ ndmqan16@gmail.com — [🌐 Nguyen Dang Minh Quan](#) — [📷 Nguyen Dang Minh Quan](#)

Technical Skills

- **Languages:** Python, C++
- **Frameworks:** Keras, PyTorch
- **LLM & Vector:** LangChain, Qdrant, Transformers
- **Big Data:** Spark, Kafka, Hadoop
- **Devops:** Docker, Kubernetes, Prometheus, Grafana
- **Modelling:** Machine Learning, Deep Learning, Time Series Forecasting, RecSys, LLM
- **Database:** MSSQL, PostgreSQL, Neo4j
- **Others:** Data Analysis, Data Visualization

Education

University of Information Technology (VNU)

Information Technology

GPA: 3.3 / 4.00

Oct 2022 – Present

Ho Chi Minh City, Vietnam

Certifications

- IELTS – Overall Band 5.5
- TOEIC – 725

Work Experiences

TGL Solution

Dec 2025 – Present

AI Engineer Intern

Tech Stack: FastAPI, LangGraph, Google SDK, Qdrant, OpenSearch, SQL, Docker

- Built a FastAPI RESTful service to serve an Agentic RAG for complex invoice queries.
- Designed an Agentic RAG pipeline using LangGraph with Gemini 2.5 Flash Lite, combining semantic vector search and structured SQL queries.
- Integrated OpenSearch for automatic Japanese entity resolution, improving SQL query accuracy.
- Containerized a bilingual (English/Japanese) system using Docker for efficient deployment.

Achievement

Kaggle Competition - Konwinski Prize

Dec 2024 – Mar 2025

Gold Medalist (Top 6/617)

- A Kaggle challenge where participants built LLM-based systems that automatically fix real GitHub issues.
- Built a **Select-Patch-Verify-Choose** pipeline with open-weight LLMs via vLLM, scoring **0.008237** (based on the competition's custom evaluation metric).
- Through the competition, gained hands-on experience in LLM deployment, constrained problem-solving, and teamwork.

Projects

– **Coinbase Streaming Data Pipeline for Cryptocurrency Price Forecasting**

Tech stack: Python, Go, Kafka, Spark, Cassandra, Docker, PyTorch, Grafana

- Built a streaming pipeline using Coinbase API, Kafka, and Spark for high-frequency price data.
- Stored data in Cassandra and monitored system metrics via Grafana.
- Trained deep learning models LSTM achieving $R^2 = 0.9677$.
- Through the project, gained hands-on experience in big data streaming, deep learning, and system monitoring.

– **Deals Hunting**

Tech stack: Python, HuggingFace, OpenAI, Modal, Pushover, Transformer

- Developed system to detect online deals and notify users in real time.
- Fine-tuned **LLaMA 3.1 8B** for price prediction using Top-K decoding with weighted averaging, outperformed baselines (LightGBM, CatBoost, RF) with **RMSLE = 0.42** and **hit rate = 86.4%**.
- Deployed a scalable inference pipeline on **Modal (serverless)**.
- Through the project, gained hands-on experience in LLM fine-tuning, and AI system design.