Tai Nguyen Phu

+84 945 409 269 | tainguyenphu2502@gmail.com | linkedin.com/in/yuitc | github.com/YuITC

Aspiring AI Engineer specializing in LLM-powered systems, with a focus on Retrieval-Augmented Generation, agentic workflows, and model fine-tuning. Seeking to contribute to AI-driven product teams.

Education

University of Information Technology - VNUHCM — Ho Chi Minh, Vietnam
Master of Computer Science — *Dec 2025 – Present*

University of Information Technology - VNUHCM — Ho Chi Minh, Vietnam
Bachelor of Computer Science | *GPA: 3.5/4.0* — *Sep 2022 – Sep 2025*

Experience

AI Engineer Intern | ***CoverGo Insurtech*** — Oct 2025 – Jan 2026

- Collaborated on improving internal RAG pipeline quality through prompt engineering and Gemini integration
- Built a DSPy-powered prompt optimization pipeline, cutting operational costs and reducing manual tuning overhead

Undergraduate Research Assistant | ***University of Information Technology*** — Dec 2024 – Mar 2025

- Investigated algorithms and optimization techniques to advance LLM inference and RAG retrieval performance
- Applied fine-tuning methods including LoRA and QLoRA through hands-on Python experimentation

Projects

Metacognitive RAG System | *LangGraph, FastAPI, Qdrant, Elasticsearch, PostgreSQL, Docker* — Jan 2026 – Mar 2026

- Developed a self-correcting RAG system with LangGraph, boosting answer accuracy by 21 points on HotpotQA compared to naive RAG baseline
- Designed adaptive retrieval optimization leveraging Thompson Sampling for dynamic query routing
- Implemented hybrid search (Qdrant + Elasticsearch) with reranking for improved retrieval precision
- Reduced unnecessary LLM calls by 46% through fast-path monitoring and early-exit logic
- Deployed an async FastAPI service via Docker Compose, enabling production-scale question answering

Novel-Verse AI | *Next.js, React, TypeScript, DeepSeek, OpenAI* — Oct 2025 – Dec 2025

- Engineered a multi-provider LLM translation pipeline with prompt caching for cost efficiency
- Designed a RAG-based Q&A system powered by OpenAI's text-embedding-3-small embedding model
- Automated structured knowledge extraction from raw unstructured text at scale via structured LLM JSON outputs
- Engineered adaptive 3-tier scraping system with fallback strategies for reliable extraction from rate-limited sources
- Delivered full-stack Next.js application with real-time SSE streaming

Vietnamese Legal Document Retrieval | *PyTorch, FAISS, Sentence-Transformers, Fine-tuning* — Mar 2025 – May 2025

- Fine-tuned Sentence-BERT on 100K+ Vietnamese legal document pairs using contrastive learning
- Evaluated retrieval model via MTEB benchmark, achieving 79% Recall@10, greatly outperforming baseline model
- Built FAISS IVF index over 100K+ document embeddings, enabling sub-second semantic search across full corpus
- Containerized and deployed the full retrieval system using Docker

Skills

Programming	Python, C++, TypeScript, SQL, FastAPI, OOP, Git, Linux, Qdrant, PostgreSQL
AI/LLM	LangChain, LangGraph, LoRA, QLoRA, DSPy, RAG, AI Agents, Context Engineering
Machine Learning	PyTorch, Transformers, Scikit-learn, Pandas, NumPy, OpenCV, XGBoost, GANs
MLOps	Docker, AWS, MLflow, Weights & Biases, GitHub Actions, PyTest

Certifications & Achievements

IELTS Academic	7.0 Overall
NVIDIA	Applications of AI for Anomaly Detection
Coursera	Deep Learning & Machine Learning Specialization
Kaggle Competition	Home Credit - Credit Risk Model Stability