

Quang Phu Nguyen

quangphunguyen1804@gmail.com | [linkedin.com/in/quangphunguyen](https://www.linkedin.com/in/quangphunguyen) | github.com/pdz1804

Vietnamese – Native | English – Professional Working Proficiency (IELTS 6.5)

Professional Summary

AI Engineer with 2 years of experience shipping production **LLM, RAG and Agentic AI** systems end to end — agent runtimes, tool-use orchestration, permission and governance layers, and multi-tenant administration with auditing, usage and cost reporting.

Full-stack across **FastAPI** backends, **React/TypeScript** portals and cloud deployment on **AWS, Azure and Google Cloud**; comfortable operating services through modern observability and disciplined practice — default-deny permissions, forward-only migrations, staged releases and layered testing through agent-behavior evaluations. Graduated **Excellent, GPA 3.8/4.0**.

Education

Ho Chi Minh City University of Technology (HCMUT)

Ho Chi Minh City, Vietnam

Bachelor of Computer Science (Major in Applied Artificial Intelligence)

Sep 2022 – 2026

- **Graduated** with **Excellent** classification — **Cumulative GPA: 3.8/4.0**
- **Academic Incentive Scholarship** (4/8 semesters), **OISP Scholarship** (3/8 semesters), **Consolidation Prize** — Bach Khoa Innovation Contest (Jun 2023).

Le Hong Phong High School For The Gifted

Ho Chi Minh City, Vietnam

Mathematics Honors Class

2019 - 2022

WORK EXPERIENCE

Associate AI Engineer

Nov 2025 – Present

Supervisor: Dr. Nguyen Duy Khuong

FPT Software AI Center, Ho Chi Minh City – F-Town 3

- Full-stack engineer on an **Agentic ERP Platform** (team of 15) — an enterprise AI workspace where autonomous agents act inside a company's business systems; my scope covers the **agent runtime, permissions and governance, multi-tenant administration, integrations** and customer-facing interfaces.
- Built the runtime **permission layer** deciding which tools an agent may use under a **default-deny** model, plus per-customer administrator controls with a full **audit trail** and **usage/cost reporting**; authored the permission-model specification.
- Shipped the first release of **user-authored Skills** and team sharing of agents, skills and artifacts; cut a slow shared-content endpoint from **tens of seconds to a single query** and consolidated user/group management onto one source of truth.
- Built a **third-party integration** from scratch and improved streaming chat (conversation search, composer usable while replies stream); operate services via **OpenTelemetry, Prometheus, Sentry and Arize Phoenix**, and review peers on architecture boundaries, permission correctness and migration safety.
- Architected a **healthcare agentic chatbot** and its **Management Portal** on **AWS Bedrock AgentCore** (team of 10), leading POC-to-production at **>90% end-to-end accuracy**; team received the **“Best Team” award** within FPT Americas (ST25).
- Presented 3 internal **AI4ALL** knowledge-sharing sessions and mentored an intern alongside senior engineers.

AI Engineer Intern

Jun 2025 – Oct 2025

Supervisor: Dr. Nguyen Duy Khuong

FPT Software AI Center, Ho Chi Minh City – F-Town 3

Remote Supervisor: Prof. Kazuyuki Motohashi

The University of Tokyo, Japan

- Built a **Blog System with AI search & recommendation** on Azure AI Search, Cosmos DB (NoSQL), Redis and Azure OpenAI, using **hybrid retrieval** (BM25 + vector + semantic + freshness scoring).
- Designed **M3ARAG**, a locally deployable, GPU-accelerated **Multi-Agent RAG** system answering questions over PDFs, HTML, Office documents and text.
- Developed a **Dual Attention Model** extracting technical keywords from company websites and a Transformation Matrix aligning Company–Patent data for the end-to-end Innovation Discovery pipeline.

Research Assistant

Jun 2024 – Dec 2024

Supervisor: Mr. Bui Cong Tuan (PhD Candidate)

Ho Chi Minh City University of Technology

- Designed an unsupervised **Knowledge Graph** construction framework requiring minimal domain-expert input, and evaluated SOTA **LLMs** for entity/intent extraction on raw Vietnamese documents.

Key Projects (delivered at work)

- Agentic ERP Platform** | *FastAPI, PostgreSQL, Temporal, Qdrant, GKE, React* Jun 2026 – Present
- Enterprise AI workspace where agents act inside a company's business systems: shareable user-authored agents and Skills, versioned Artifacts, scheduled Automations, and a plugin system for third-party software — with runtime permissions, audit trail and cost reporting.
- Healthcare Agentic Chatbot & Management Portal** | *AWS Bedrock AgentCore, Strands* Nov 2025 – May 2026
- Multi-agent chatbot (Product Recommendation, Health & Lifestyle, Product Education) at >90% end-to-end accuracy, plus a portal for knowledge/prompt/guardrail management, change history and real-time monitoring.
- Smart Product Recommendation System (SPR)** | *Amazon Bedrock, Testing Automation* Oct – Nov 2025
- Rule-based and LLM hybrid recommendation engine for a skin-health scanning device; regulatory compliance, latency and quality monitoring, and automated quality gates for production release.
- M3ARAG – Multi-Agent RAG System** | *LangGraph, LangChain, Docling, ColPali* | [Link](#) Aug – Oct 2025
- GPU-accelerated multi-agent RAG for local, cloud-free deployment, with an end-to-end document intelligence pipeline over PDFs, HTML, Office documents and text.
- Azure Blog System with Search & Recommendation** | *Azure AI Search, Cosmos DB* Aug – Oct 2025
- Full-stack blog platform with semantic search and a personalized recommendation engine; hybrid BM25 + vector + semantic retrieval with Redis caching for low-latency delivery.

Personal Projects (self-driven & academic)

- Fine-tuning Language Models for NLP Tasks** | *HF Transformers, PEFT/LoRA, PyTorch* | [Link](#) April 2025
- Fine-tuned and compared **T5-Base**, **BART-Base** and **Flan-T5-Small** across sentiment analysis, question answering and machine translation, weighing full fine-tuning against **LoRA** and adapter tuning (5-member team).
- Sentiment Analysis with Various Models** | *Scikit-learn, XGBoost, Keras, PyTorch* | [Link](#) April 2025
- Benchmarked Decision Trees, Naïve Bayes, SVM, XGBoost, Random Forest, MLP and Bi-LSTM over TF-IDF and word-embedding features, focusing on feature transformation and hyperparameter tuning (5-member team).
- Detect AI-generated Text** | *Scikit-learn, PyTorch, DistilBERT* | [Link](#) Dec 2024
- Compared classical models (SVM, Random Forest, XGBoost) against neural approaches (MLP, **DistilBERT**) to classify LLM-generated essays in the educational domain (3-member team).

TECHNICAL SKILLS

Languages: Python (advanced), JavaScript/TypeScript, C/C++, SQL, R

LLM & Agents: LangChain, LangGraph, AWS Strands, MCP, RAG Systems, Agent Evaluation/LLM-as-a-Judge, OpenAI API, Google Gemini, Docling, PEFT/LoRA

ML & Data: PyTorch, TensorFlow/Keras, Scikit-learn, XGBoost, Hugging Face Transformers, Sentence-Transformers, spaCy/NLTK/Gensim, NumPy, Pandas, Polars

Cloud & DevOps: AWS – Bedrock AgentCore/KnowledgeBase/Guardrails, Lambda, ECS/ECR, EC2, S3, OpenSearch Serverless; Azure – AI Search, Cosmos DB, OpenAI; Google Cloud – Cloud SQL, GKE, Artifact Registry; Docker, Kubernetes, Terraform, ArgoCD, CI/CD, OAuth 2.0/RBAC

Backend & Data Stores: FastAPI, Pydantic, SQLAlchemy/Alembic, PostgreSQL, MySQL, MongoDB, Redis, Temporal, Qdrant/ChromaDB

Observability, Quality & Web: OpenTelemetry, Prometheus, Sentry, Arize Phoenix, SonarQube, Backstage, Playwright/Selenium, ReactJS, TypeScript, Tailwind CSS, Linux/Nginx

CERTIFICATIONS & TRAINING

Anthropic: Model Context Protocol: Advanced Topics (May 2026), AI Fluency Framework & Foundations (May 2026), Claude with Vertex AI (May 2026), Introduction to Subagents & Claude 101 (Apr 2026)

Google: AI Specialization & AI Fundamentals (Feb 2026), Prompting Essentials (Jun 2026), Foundations of Data Science (May 2026), Gemini Certified University Student (Dec 2025)

DeepLearning.AI: AI Agents in LangGraph, Functions & Tools with LangChain, LangChain for LLM Apps, Build AI Apps with MCP Server | **Hugging Face:** AI Agents Fundamentals (Jun 2025)

DataCamp: AI Engineer for Data Scientists Associate (Sep 2025) | **Google Cloud:** TensorFlow, Responsible AI for Developers, Generative AI with Gemini, Vertex AI skill badges