

NGUYEN VAN NGUYEN

AI Engineer — LLM Applications, RAG & Multi-Agent Systems

Binh Thanh, Ho Chi Minh City, Vietnam | 0396924462 | nguyenvannguyen192402@gmail.com

SUMMARY

AI Engineer with 2+ years of experience designing and shipping production LLM applications, including Retrieval-Augmented Generation (RAG) pipelines, multi-agent systems, and recommendation engines across the financial, social media, F&B, and manufacturing sectors. Comfortable owning the full stack from data ingestion and vector retrieval to API deployment and chat-channel integration (Zalo OA, Facebook Messenger). Strong foundation in PyTorch, LangChain, and LLM fine-tuning (LoRA/QLoRA).

TECHNICAL SKILLS

LLM & GenAI: LangChain, OpenAI API, Gemini API, Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Prompt Engineering, LLM Fine-tuning (LoRA, QLoRA, Unsloth AI)

Machine Learning / Deep Learning: PyTorch, TensorFlow, HuggingFace Transformers, LightGBM

NLP: Tokenization, Named Entity Recognition, Text Generation

Computer Vision: YOLOv8, CNN, Image Classification, Object Detection

Data & Infrastructure: Qdrant, ChromaDB, Pinecone, Redis, PostgreSQL, FastAPI, Docker, Git, GCP, AWS, CI/CD

Web & Automation: Next.js, TypeScript, n8n workflow automation, Zalo OA / Facebook Messenger webhooks, SePay payment integration

Languages: Python, TypeScript, Java, C/C++

WORK EXPERIENCE

AI Engineer | Le Tran Co., Ltd. (Le Tran Furniture)

09/2026 - Present

Nhi Xuan Industrial Zone, Hoc Mon, Ho Chi Minh City, Vietnam

- [Add 2-4 bullets: the AI systems you are building at Le Tran, the tech stack, and measurable results]

AI Engineer | HONGTHAI Technical

08/2025 - 07/2026

Ho Chi Minh City, Vietnam

- Led development of a Financial RAG ecosystem for parsing and querying high-density financial documents (annual reports, market news), using Docling for PDF-to-Markdown extraction and Qdrant as the vector store for context-rich retrieval.
- Architected a multi-agent financial analysis framework in which specialized agents, modeled on distinct investor personas (e.g., Warren Buffett, Stanley Druckenmiller, Peter Lynch), independently analyze the same document set and produce diverse, high-depth financial perspectives through iterative reasoning workflows.
- Designed and deployed a multi-stage recommendation engine (recall, ranking, re-ranking) combining LightGBM, collaborative filtering, and content-based filtering, with Redis-cached candidates served via FastAPI to personalize content for high-traffic users.
- Delivered an end-to-end AI chatbot automation project for enterprise client Hotbox (F&B technology): integrated multi-channel messaging (Zalo Official Account, Facebook Messenger) with automated payment processing (SePay), serving small and medium-sized enterprises in the Vietnamese market.
- Set up CI/CD pipelines and deployed AI services on Google Cloud Platform (GCP) and AWS; built n8n automation workflows powering both chatbot logic and automated AI video content generation, and contributed Next.js/TypeScript front-end components for internal tooling.

AI Engineer | TQT GLOBAL Trading Company Limited

05/2024 - 08/2025

Ho Chi Minh City, Vietnam

- Built a private, on-premise RAG assistant using a locally hosted Gemma 3 12B IT model served via vLLM, ChromaDB as the vector store, and multilingual sentence-transformer embeddings, automating internal knowledge retrieval while keeping all data on-premise.
- Developed an MCP-based FastAPI server to manage vector database operations and tool orchestration, and integrated SearXNG for real-time web search augmentation within the assistant.
- Contributed to an in-house Text-to-Speech (TTS) pipeline and shipped a Next.js (Vercel AI Chatbot) front end, launching a production-ready assistant for internal knowledge access and employee support.

PERSONAL PROJECT

Autonomous Financial Intelligence Agents | Independent Project

02/2026 - Present

- Building a 24/7 autonomous multi-agent system that monitors, analyzes, and synthesizes financial insights from global news and influential financial commentators, including sentiment/trend analysis agents and a cross-source fact-verification layer. Stack: Python, FastAPI, OpenAI API, Gemini API.

EDUCATION

B.Sc. in Computer Science

2019 - 2024

Can Tho University — GPA: 3.07/4.0

Certifications: VSTEP B1 — English Language Proficiency (2024)