Trinh Quang Vinh

AI Engineer

Ha Noi City, Vietnam | trinhquangvinhna@gmail.com | +84 398 072 726 | LinkedIn Profile |
Github

Professional Summary

AI Engineer with 1 years of experience in NLP and Generative AI, specializing in chatbots, RAG systems, and AI agents. Skilled in delivering end-to-end AI solutions from research to production deployment. Eager to apply AI expertise in a collaborative environment to build innovative solutions that transform business operations.

Education

University of Science, VNU, Bachelor's Degree in Electronic Engineering and Informatics - Artificial intelligence — Sept 2021 – May 2025

- GPA: 3.06/4.0
- **Relevant Coursework:** Machine Learning, Deep Learning, Data Structures & Algorithms, Statistics, Linear Algebra, Computer Vision, Natural Language Processing
- **Thesis:** "Development of a Real-Time Motorcycle Feature Extraction System Using Vision-Language Models" (Grade: A+)

Professional Experience

AI Engineer, Authentic Education Hub – Ha Noi — Jan 2025 – Present

- Developed end-to-end AI workflows including RAG systems and intelligent agents for production deployment
- Created and maintained internal libraries and APIs to streamline integration of AI capabilities across multiple projects
- Increased response accuracy and consistency of LLM-powered systems through advanced prompt engineering strategies
- Built automated data collection pipelines (web crawlers, scrapers, annotation tools) to generate high-quality training datasets and knowledge bases
- Packaged and deployed machine learning and services to production using Docker and CI/CD pipelines

AI Engineering Internship, Authentic Education Hub – Ha Noi — Nov 2024 – Jan 2025

- Implemented vector databases and embedding models for document retrieval systems
- Researched and experimented with vision-language models for image captioning and analysis
- Fine-tuned LLMs using PyTorch for domain-specific AI applications

Backend Developer, VMO Holdings – Ha Noi City — June 2024 – Sept 2024

- Developed REST APIs with Spring Boot and optimized database queries using JPA
- Implemented JWT authentication and achieved 85% test coverage with JUnit

Key Projects

The Edu AI Workshop | Educational AI Platform — Apr 2025 – Present

Technologies: OpenAI, Langchain, PaddleOCR, FastAPI, Docker, MinIO, YouTube Transcript API, Whisper ASR, FFmpeg

- **AI Workflow Architecture:** Designed comprehensive workflows and detailed technical specifications for educational AI tools supporting teachers, students, and parents.
- **Interactive Mind Map Generator:** Engineered end-to-end pipeline processing multi-format inputs through custom template prompting system, generating structured educational content with conditional image integration API and markdown export functionality.

- **Intelligent Assessment Generation System:** Built automated quiz creation tool supporting 1-100 questions across multiple difficulty levels with smart distribution algorithms, parallel processing, and multi-language support.
- **YouTube Content Extraction Pipeline:** Developed robust video-to-learning-material converter with fallback architecture (YouTube Transcript API → Whisper ASR), audio preprocessing using FFmpeg , concurrent multi-video processing, and intelligent content summarization with translation capabilities.
- **API Development:** Created scalable RESTful APIs using FastAPI for tool integration, cross-platform content management, and file I/O operations with MinIO storage, supporting diverse formats (PDF, DOCX, multimedia).
- **Advanced Content Processing:** Implemented advanced prompt engineering to transform raw data into structured educational content, improving content accuracy and usability.

AI Product Scanning System | KoreanBay Mobile App Jun 2025 – Jul 2025

Technologies: OpenAI, Langchain, FastAPI, RAG, Tavily Search, PostgreSQL, Docker, nanoOCR, paddleOCR

- **Korean OCR Processing:** Researched and tested multiple OCR solutions, then deployed a nanoOCR-based model for Korean text extraction from product packaging, achieving 90%+ accuracy on health supplement labels.
- **Logo Recognition System:** Finetuning YOLOv12m for Korean certification and insurance logo detection with bounding box localization, achieving F1-score >90% for multi-class logo classification.
- **Intelligent Translation:** Built an LLM-powered Korean–Vietnamese translation pipeline with integrated knowledge base search and context retrieval for accurate product information localization.
- **Smart Product Discovery:** Designed workflow and built RAG architecture with knowledge base search and Tavily web search fallback for comprehensive product matching.

Technical Skills

Programming Languages: Python (advanced), SQL, JavaScript (basic for integration)

AI/ML & NLP: LangChain, Hugging Face Transformers, OpenAI APIs, GraphRAG, PyTorch, TensorFlow, RAG Systems.

Data & Vector Databases: Elasticsearch, Pinecone, FAISS, PostgreSQL(pgvector), MongoDB

Cloud & MLOps: Minio, Docker, Git, CI/CD

Web & Deployment: FastAPI, Flask, Streamlit, RESTful APIs, React.js (for prototypes)

Certifications & Achievements

- **AI CAMPUS: Samsung Innovation Campus** (2024)
- **Deep Learning Specialization** - Coursera University (2024)

Languages

Vietnamese: Native **English:** 650 TOEIC