Tran Trong Nhan

Email: trant.nhan2003@gmail.com | **Phone:** +84 797 070 237 | **LinkedIn:** linkedin.com/in/trantnhan2003
GitHub: github.com/nhantran0506 | **Kaggle:** kaggle.com/trngnhntrna | **Website:** trantrongnhan.com

Summary

Results-driven AI Engineer with experience building and deploying full-stack machine learning applications. Expertise in both NLP and Computer Vision, with a strong foundation in MLOps, cloud infrastructure, and creating high-performance, scalable AI services.

Technologies and Skills

Data Science Skills: Generative AI, Data Mining, Data Preprocessing, Data Augmentation, Machine Learning, Deep Learning, NLP (Natural Language Processing), Computer Vision

Frameworks & Libraries: PyTorch, TensorFlow, Scikit-learn, OpenCV, Tesseract, Langchain, Langgraph, LlamaIndex, Huggingface, Flask, Django, FastAPI, .NET, Next.js, AsyncIO.

Programming Languages: Python, C++, Golang, JavaScript, C#, Java, SQL

Databases & Vector Stores: PostgreSQL, MySQL, MongoDB, Weaviate, ChromaDB, Pinecone, Qdrant, Milvus

DevOps & MLOps: Docker, Git, CI/CD pipelines, MLflow, AWS, K8S

Development Methodologies: SCRUM, Design Patterns, Software Development Life Cycle, Business Analysis, Agile, MVC

Languages

- **English**: Professional (920/990 - **TOEIC**)

Education

Ho Chi Minh University of Technology and Education, BS in Software Engineer **2021 – 2025**

- GPA: 3.33 /4.0

Experience

Junior AI Developer April 2025 – Present
Grooo International

- **Spearhead** the design and development of an **Agentic AI** platform that delivers personalized stock recommendations and in-depth analysis for diverse client need.
- **Designed scalable data pipelines** handling over **7 million records** for downstream AI applications.
- **Apply** semi-supervised learning to build customer behavior and risk-profiling segmentation models, directly contributing to a **30%** revenue growth in the project.
- Developed OCR-based AI services for intelligent document extraction, analysis, and summarization, significantly improving processing efficiency.
- **Tech Stack:** Langgraph, Langchain, Pinecone, Kafka, Minio, Spark, Docker, Git, Google Cloud.

AI Engineer July 2024 – March 2025
Primas Group

- Developed an **AI Agent** for artist recommendation.
- Designed and implemented a real-time text-to-speech (TTS) system for voice chat.
- Research and developed a guardrail system for Retrieval-Augmented Generation (RAG)
- Fine-tuned distilled SLM for entity extraction and validation, **increasing** the project revenue by **10%.**
- **Technical Stack:** Flask, LlamaIndex, Milvus, ChromaDB, Ollama, LlamaCPP, Huggingface, k6, Docker, Git

AI Engineer Intern Aug 2023 – Dec 2023

PA Vietnam

- Developed a robust machine learning pipeline and predictive model for revenue time series forecasting, increasing the accuracy of the forecast by 85% and enabling data-driven strategic planning.
- Implemented an advanced anomaly detection system for real-time revenue monitoring, reducing critical issue response time by 90%.
- Designed and deployed an AI-powered FAQ chatbot for customer support on Telegram, reducing average customer **wait times by 80%** through intelligent automation.
- **Technical Stack:** Flask, LangChain, FAISS, Transformers, Docker, Git

Projects

AI-Powered E-commerce Platform

Engineered a full-stack e-commerce site with a suite of AI features to enhance user experience, including an LLM customer service agent, personalized recommendations, and AI-generated 3D product visualizations with virtual try-on.
Tech Stack: Next.js, Kafka, Django, LlamaIndex, Weaviate, Docker, AWS

Vietnam Stock Analysis and Prediction

Developed and deployed a full-stack web application that provides real-time analysis and novel price predictions for the Vietnamese stock market, featuring interactive data visualizations and user authentication.
Tech Stack: Next.js, Django, Docker, Git, AWS

Vietnamese Legal AI Assistant

Fine-tuned a specialized LLM for Vietnamese law, achieving an **87% ROUGE-1 score** on Q&A tasks. Architected a RAG system with semantic search to deliver accurate, citation-backed legal information via a microservice API.
Tech Stack: React, FastAPI, LangChain, Pinecone, Transformers, Docker

Autonomous Research Agent

Built an autonomous agent using Langgraph to automate end-to-end research and report generation, featuring a self-correction loop for iterative fact-checking and refinement to ensure high-accuracy outputs.
Tech Stack: Langgraph, Ollama, OpenAI, Taivy, Git

Video Moment Retrieval Engine

Developed a high-performance search engine to retrieve specific moments from a large video library (news, cooking shows) using text queries. Utilized a CLIP model for cross-modal understanding and the Qdrant vector store for millisecond-latency retrieval, implementing query optimization and multiprocessing for scalability.
Tech Stack: FastAPI, Qdrant, CLIP, Transformers, OpenCV, Docker, Git

Certificates

• Machine Learning Specialization	Coursera, Oct 2022
• Google Data Analytics	Coursera, July 2023
• Google Cloud Training Day: Core Infrastructure Fundamentals	Cloud Ace, Dec 2023
• SDWS Training	Axon Active, May 2024
• Natural Language Processing Specialization	Coursera, Apr 2024

Honors and awards

• Consolation Prize, Vietnamese Olympiad in Informatics (Open Source Category)	Vietnamese OLP, Nov 2023
• Prototyped a QR-code based digital wallet to enable payments for unbanked users.	Global Hack, June 2024