Tran Ngoc Tan

AI Engineer

0382357800 tranngoctan910@gmail.com https://newttran.space/ Ho Chi Minh, Viet Nam

ABOUT ME

A motivated AI Engineer specializing in Agentic AI workflows, advanced RAG architectures, continuous LLM fine-tuning pipelines, and scalable backend systems. Combined with a solid background in Information Security, I am fully equipped to build secure, high-performance, and intelligent AI-powered applications.

SKILLS

AI & LLM ENGINEERING	LLMs, PyTorch, Transformers, RAG, Agentic AI, Vector Search, QLoRA, Unsloth, vLLM, Prompt Engineering, MLOps, Qdrant
LANGUAGES	Python, Java,JavaScript, TypeScript
FRAMEWORK & OTHERS	FastAPI, NestJS, SpringBoot, MySQL, PostgreSQL, Docker, Linux, Git, GitHub Actions

EDUCATION

Posts & Telecoms Institute of Technology HCM (PTIT HCM) **09/2021 - 12/2025**

- **Major - Information Security**

CGPA: 3.18

WORK EXPERIENCE

Daito Vietnam Co., Ltd. **05/2026 - Present**

AI R&D Engineer

- **Technology use:**
  - AI & Data: Agentic AI, RAG, Qdrant, PyTorch, Transformers, Unsloth, vLLM, Qwen2.5 7B, Gemini API, MLOps
  - Development: Python, FastAPI, Laravel, VueJS, MySQL
- **Content**: Researched and developed production-grade AI systems, including continuous LLM fine-tuning pipelines, hybrid local-cloud AI applications, and advanced RAG architectures. Built end-to-end solutions covering model optimization, agentic workflows, vector retrieval, MLOps automation, and full-stack AI platform development.
- **Projects**
  - **E-Commerce AI Chatbot & Advanced RAG System**
    - Description: Built a personalized AI shopping assistant powered by a Hybrid RAG architecture, capable of contextual product discovery, intelligent recommendations, and complex multi-turn conversations.
    - Responsibilities: Proposed and designed the end-to-end AI architecture from research to production deployment. Developed agent-based retrieval and reasoning workflows, implemented contextual memory and vector search capabilities, and built ETL pipelines and a real-time streaming platform for low-latency user interactions.
    - Technology use: Agentic AI, RAG, Vector Search, Qdrant, Gemini API, Python, FastAPI, Laravel, VueJS, MySQL
  - **Hybrid AI Email Intelligence & LLM Fine-Tuning Pipeline**
    - Description: Built a hybrid AI email classification platform and a continuous learning pipeline (Data Flywheel) for automated data collection, fine-tuning, and deployment of Qwen2.5 7B models.
    - Responsibilities: Architected an end-to-end continuous learning system (Data Flywheel) integrating Human-in-the-Loop (HITL) validation. Optimized inference of pre-quantized models in vLLM using dynamic LoRA adapters. Engineered structured prompting frameworks, implemented incremental QLoRA fine-tuning via Unsloth with replay datasets, and automated evaluation and zero-downtime adapter hot-swapping.
    - Technology use: Prompt Engineering, Gemini API, Transformers, PyTorch, Unsloth, vLLM, Qwen2.5 7B, MLOps, Python, FastAPI, MySQL

OTG Lab Co., Ltd **09/2025 - 05/2026**

Software Engineer

- **Technology use:**
  - Development: C#, ASP.NET Core Web API, Laravel, NestJS, MySQL, PostgreSQL (NeonDB), Hangfire (Background Jobs), Swagger, JWT Authentication, OpenAI API Integration, Speechmatics API, Figma.
  - DevOps/CICD: Bitbucket, Git, GitHub Actions, Git Webhooks, Docker, Jenkins
- **Content:** Developing web applications, working with the Singaporean operations team in English (mostly via messaging).....

- **Projects:**
  - **Educational Website**
    - Description: A classroom observation application that leverages Generative AI to automate the evaluation of teaching effectiveness. It processes session recordings to generate detailed reports on teacher-student interaction patterns, helping institutions scale their quality assurance efforts.
    - Responsibilities: Backend Developer (Architecting an AI-driven audio analysis platform, integrating advanced LLMs and media processing tools within a highly scalable .NET ecosystem.)
    - Technology use: C#, ASP.NET Core Web API, Entity Framework Core, Hangfire (Background Jobs & Schedulers), Multimedia & Data Processing, Supabase, MySQL, OpenAI API Integration, Speechmatics API, Swagger, JWT Auth, BitBucket, GitHub,...
  - **F&B Management System**
    - Description: A comprehensive F&B operation platform consisting of four integrated portals (Customer, Store, Driver, Admin) that streamlines the entire workflow from ordering to delivery. The system features real-time order synchronization, automated driver dispatching, and dynamic inventory management across multiple outlets.
    - Responsibilities: Back-end Developer (Ensuring scalability and maintainability. Integrated Google Maps API for real-time driver tracking and route optimization. Developed a robust real-time communication layer using Socket.IO for instant chat and live order updates, and implemented Firebase Cloud Messaging (FCM) for reliable cross-platform mobile push notifications.).
    - Technology use: NestJS, TypeScript, PostgreSQL, Google Maps API, Socket.IO, Firebase (FCM), Docker, Swagger...

TMA Solusions **06/2024 - 09/2024**

Front-end intern of social travel app at TMA Solutions

- Responsibilities: Font-end developer (Angular)
- Technology use: AngularJS, GitLab, Jira, Figma, Swagger,...
- Content: Working in scrum team as a Frontend developer. Develop UI and Integrate with API.

ACADEMIC PROJECTS

BUILDING AN EXPLAINABLE AI-ENHANCED SURICATA IDS/IPS FOR DDOS ATTACK DETECTION AND PREVENTION **09/2025 - 12/2025**

Graduation Project

- Link Report: https://drive.google.com/file/d/1dgYmOZTnv5wNsOObBYe4frJL6KXCNKXp/view?usp=sharing
- Designed and deployed a high-performance IDS/IPS that combines Suricata with XAI for accurate, transparent, and automated DDoS mitigation.
  - Developed an automated pipeline: Suricata Alerts → Traffic Capture (tcpdump) → Feature Extraction (CICFlowMeter) → AI Inference → Real-time Mitigation (iptables).
  - Optimized detection performance by training Siamese Neural Networks (SNN) and XGBoost on the CIC-DDoS2019 dataset, achieving high accuracy in identifying Layer 3/4 attacks.
  - Integrated SHAP (Explainable AI) to interpret model decisions, providing transparency for network administrators during threat analysis.
  - Orchestrated a distributed botnet environment using **Docker macvlan** to simulate real-world large-scale DDoS scenarios for stress testing.
  - Built a real-time monitoring dashboard with **Grafana** to visualize network traffic health and security alerts.
- Technologies: Python, PyTorch, XGBoost, Siamese Neural Network, SHAP (XAI), Suricata, Docker (macvlan), Grafana, CICFlowMeter, tcpdump, iptables, Bash Scripting, Linux (Ubuntu/Kali).