Trinh The Minh

(+84) 837-282-889 • Hanoi • trinhtheminh1104@gmail.co • github.com/trinhminh11 • linkedin

SUMMARY

Product-focused **Full-Stack AI Engineer** with proven expertise end-to-end lifecycle of AI applications, building production-grade **RAG**, **Knowledge Graph (KG-RAG)**, and **Multi-Agent** systems. Proven track record of replacing costly legacy/outsourced systems with optimized in-house solutions, achieving **50%+ cost reductions**. Comprehensive experience across the full stack—from designing **GCP** data pipelines to deploying scalable **MLOps** workflows on **Kubernetes**. Dedicated to creating high-impact AI products that drive user engagement and business efficiency.

TECHNICAL SKILLS

Programming Languages: Python, SQL, Java, C/C++

AI/ML & Data: LLMs, PyTorch, Langchain, LangGraph, Neo4j, Qdrant, Milvus, Cassandra, HuggingFace

MLOps & Cloud: Docker, Kubernetes, GitHub Actions, GitLab, GCP, AWS

Backend: FastAPI, REST, WebSocket, SSE, Postgres, Redis, Celery

Languages: English (IELTS: 6.5)

WORK EXPERIENCE

Founding AI Engineer (Core Product Team) **Full-time**
TeenUp (EdTech Startup) — https://teenup.vn *Sep 2025 - Dec 2025*

- Replaced the legacy outsourced infrastructure with a custom in-house solution through **GCP**, achieving an immediate **50%+ cost reduction** by optimizing compute and API usage.
- Architected and launched a **Knowledge Graph RAG** using **Neo4j** and **Vertex AI** to power a ”Parent Copilot,” driving a **30% DAU/MAU** ratio through high-relevance AI interactions.
- Developed a deterministic Multi-Agent system using **custom Python routing**, implementing strict guardrails to prevent hallucination in real-time user-facing chats.

AI Engineer & MLOps Engineer (via H&C Technology) **Remote**
Musashino Co., Ltd., Tokyo, Japan — https://www.musashino.co.jp *April 2025 - Nov 2025*

- Deployed a self-hosted **RAG** system serving **50+ child companies**, replacing Google’s Vertex AI Agent Builder to eliminate unit limits and reduce operational costs by 40%.
- Achieved ultra-low latency semantic search by optimizing the stack with **Redis caching**, **Celery workers**, **FastAPI**, and **Vertex AI embeddings**, handling high concurrency without degradation.
- Streamlined MLOps workflows by automating deployment and monitoring with Docker and health-check scripts, significantly reducing troubleshooting time.

AI Engineer & MLOps Engineer **Intern**
PIXTA Vietnam, Hanoi, VietNam — https://pixta.vn *Feb 2025 - April 2025*

- Engineered the ’Auto Review’ pipeline for Pixta Stock, processing **10,000+ daily images** with **99% precision** and **70% recall** by fine-tuning **CLIP (MLP head)** and **BLIP-v2** models.
- Implemented **NER models** to extract metadata and reduce false positives in multi-label classification, allowing the moderation team to focus on complex policy violations.
- Scaled containerized deployments on **AWS** using Docker to handle peak upload traffic in near real-time, ensuring zero bottlenecks in the content ingestion workflow.

ACADEMIC RESEARCH

Research Assistant **Member**
Optimization Lab - BK.AI International Center, Hanoi, VietNam — https://bkai.ai *Oct 2022 - Present*

- Co-authored journal articles on applying Deep Learning techniques to complex real-world optimization challenges.
- Conducted in-depth research exploring novel AI approaches, leveraging machine learning and data mining techniques to address complex optimization problems.

PROJECTS

Legal Documents Retrieval **Leader**
SOICT Hackathon 2024 Competition *Oct - Dec 2024*

- Developed a domain-specific retrieval engine using a hybrid **Bi-Encoder and Cross-Encoder** architecture, achieving **MRR@10 of 0.7352**.
- Ranking in the **Top 5** among all participants, showcasing expertise in specialized NLP tasks and retrieval optimization.

EDUCATION

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY (HUST) **EXPECTED GRADUATION: 2026**
Bachelor of Data Science and Artificial Intelligence

High School for Gifted Students, VNU University of Science **Sep 2019 - May 2022**
IT Specialized

PUBLICATIONS

"STEAM: Securing the connections in wireless sensor networks with varying target priority", Accepted at Computer Networks, Jan 2026

"DRONEs: Deep Reinforcement Optimization for Network k-Connectivity Restoration Enhancement in UAVs", SOICT, 2025

"SPARTA-GEMSTONE: A two-phase approach for efficient node placement in 3D WSNs under Q-Coverage and Q-Connectivity constraints", JNCA (Q1), July 2025

"Heuristic and Approximate Steiner Tree Algorithms for Ensuring Network Connectivity in Mobile Wireless Sensor Networks", JNCA (Q1), Jun 2025

"Enhanced Diffusion for Semi-Supervised Learning", NOLTA, Dec 2024

AWARD

HUST - SOICT HACKATHON 2023: First Prize

HUST - SOICT HACKATHON 2024: Second Prize