

Tai-Nguyen Phu, MSc Junior AI Engineer

📍 Ho Chi Minh, Vietnam [in taingph2502](#) [G taingph2502](#) [✉ tainguyenphu2502@gmail.com](#) [📞 +84 945 409 269](#)

Education

Master of Computer Science GPA: 3.5/4.0 University of Information Technology (UIT – VNUHCM)	12/2025 – Present Ho Chi Minh, Vietnam
Bachelor of Computer Science GPA: 3.5/4.0 University of Information Technology (UIT – VNUHCM)	09/2022 – 06/2025 Ho Chi Minh, Vietnam

Work Experience

Junior AI Engineer , <i>724Software</i>	03/2026 – 06/2026
- Engineered core components of an education-focused agentic RAG system using LangGraph, with dynamic routing across learning materials and structured databases; used Langfuse traces to diagnose retrieval, latency, and failure patterns. - Developed OCR and layout-understanding pipelines for 70,000+ educational document images, spanning VLM fine-tuning, layout preprocessing, and vLLM inference serving. - Designed a weighted, polygon-aware fusion algorithm for multi-model text detection, improving bounding-box F1 by 5-10% on complex document layouts.	
AI Engineer Intern , <i>CoverGo Insurtech</i>	09/2025 – 01/2026
- Improved internal RAG reliability by redesigning structured prompts and integrating citation-grounded Gemini workflows, reducing unsupported and hallucinated responses. - Built a DSPy-based prompt optimization pipeline evaluated on 1,000 synthetic product-data cases, reducing manual tuning for the engineering team.	
Research Assistant , <i>University of Information Technology</i>	09/2024 – 03/2025
- Evaluated LoRA and QLoRA across downstream NLP tasks, identifying trade-offs in convergence stability, GPU memory usage, and domain-adaptation performance. - Built reproducible fine-tuning pipelines with controlled hyperparameter experiments, enabling systematic selection of PEFT configurations for resource-constrained model adaptation.	

Personal Projects

Manga Scanlation Workspace	04/2026 – 07/2026
- Architected a local-first, human-in-the-loop AI platform that unifies region detection, OCR, contextual LLM translation, inpainting, typesetting, review, and export in one revisioned workflow. - Fine-tuned PaddleOCR-VL on 21,000 Manga109 images via SFT, raising Japanese Exact Match from 9.2% to 65.9% and lowering Character Error Rate from 55.4% to 9.1%. - Built context-aware LLM translation with speaker relationships, glossaries, and translation memory, preserving cross-panel narrative consistency.	
Adaptive Multi-Hop QA Research System	01/2026 – 04/2026
- Built a self-correcting agentic RAG pipeline in LangGraph, improving multi-hop QA F1 on 2WikiMultiHopQA from 25.2% to 70.3% via automated failure diagnosis and re-retrieval. - Designed a hybrid sparse-dense retrieval pipeline (Qdrant + Elasticsearch) with cross-encoder reranking and citation grounding to reduce unsupported answers. - Implemented cost-aware LLM routing to flag low-quality generations and trigger targeted remediation, balancing accuracy against latency and API cost.	

Skills

Programming	Python, C/C++, SQL, FastAPI, Git, Linux, Qdrant, Elasticsearch, PostgreSQL, OOP, SOLID
Machine Learning	PyTorch, Transformers, Scikit-learn, Pandas, Numpy, SciPy, XGBoost, GANs
NLP/Computer Vision	LangChain, LangGraph, PEFT, Unsloth, SFT, DSPy, ONNX, OpenCV, YOLO, SAM, Paddle, vLLM
MLOps	Docker, Langfuse, MLflow, Weights & Biases, GitHub Actions, PyTest

Certifications & Achievements

IELTS Academic	7.0 Overall
NVIDIA	Applications of AI for Anomaly Detection
Coursera	Deep Learning Specialization, Machine Learning Specialization, NLP Specialization