

THÁI HOÀI AN

AI Engineer Intern/Fresher — LLM/RAG, NLP/CV, MLOps, FastAPI
0868490191 — hoaianthai345@gmail.com — LinkedIn — GitHub — Portfolio

Skills

- **AI/ML & Deep Learning:** Python, PyTorch, scikit-learn, XGBoost, Hugging Face Transformers, PhoBERT/BERT, CNN/RNN/Transformer concepts, model training, fine-tuning, evaluation, error analysis, ablation study.
- **LLM / RAG / Agents:** LangChain, LangGraph, RAG workflows, embeddings, semantic search, PostgreSQL/pgvector, prompt engineering, citation-aware QA, metadata/ACL filtering, tool-calling workflow design.
- **NLP / CV / Speech:** Vietnamese text classification, NER/IE, relation extraction, WER/CER benchmarking, computer vision, object detection, segmentation, YOLO, Knowledge Distillation.
- **AI Engineering & Data Stack:** FastAPI, WebSocket, REST APIs, Gradio, Streamlit, Docker, MLflow, MinIO, ONNX, TensorRT, PostgreSQL, ClickHouse, Kafka, Spark Structured Streaming, Git, Linux basics.

Experience

VNPT AI Hackathon 2025 — Minute / MeetMate

3rd Prize, Track 1

Scope: 2 code contributors on GitHub — Role: Lead Developer / Data Engineer / AI Engineer

- Built the data and retrieval layer for a BFSI AI meeting co-host: PostgreSQL/pgvector schema, OCR/SmartReader ingestion, document metadata, ACL filtering, and permission-aware RAG for cited answers.
- Integrated LangGraph/LangChain, FastAPI, and WebSocket flows across pre-meeting, in-meeting, and post-meeting stages, turning live transcript and enterprise documents into recap, ADR, minutes, and action items.
- Result: became the main code contributor in the repository (188/278 commits, about 68%) and helped the team win 3rd Prize; learned to balance latency, RAG quality, traceability, and audit-ready outputs.
- Tech Stack: Python, LangGraph/LangChain, FastAPI, WebSocket, PostgreSQL/pgvector, Electron.

Faculty-level Scientific Research Competition — Vietnamese Fake News Detection

1st Prize

Role: Lead Researcher / Developer — Ownership: benchmark design, model training, error analysis

- Designed a comparative NLP benchmark on 9,713 ReINTEL Vietnamese social-media posts with severe class imbalance, evaluating BiLSTM, frozen/fine-tuned PhoBERT, and LLM prompting configurations.
- Fine-tuned PhoBERT reached Accuracy 96.30%, Macro-F1 92.90%, and AUC 98.00%, outperforming BiLSTM accuracy 81.89% and the best LLM setup 73.25% while remaining more practical for deployment.
- Result: won 1st Prize at faculty level and presented the work at NCTD 2025; established a metrics-first evaluation frame using Macro-F1, minority-class recall, error analysis, and efficiency trade-offs.
- Tech Stack: Python, PyTorch, Transformers, PhoBERT, scikit-learn, LLM prompting.

Selected AI Engineering Projects

- **HPC MLOps Object Detection with YOLO Knowledge Distillation** — Built a deployment-oriented traffic object detection pipeline with YOLO teacher/student training, KD, FastAPI/Gradio serving, Docker Compose, MLflow/MinIO, and monitoring-ready infrastructure. Student KD reached mAP50 0.725 and mAP50-95 0.490 while reducing parameters from 58.82M to 2.51M and size from 112.85 MB to 5.14 MB; optimized inference from PyTorch 9.4 ms/image to ONNX 5.1 ms/image and TensorRT FP16 1.84 ms/image on Tesla T4.
- **Vietnamese Streaming ASR Benchmark** — Built a reproducible benchmark for Vietnamese ASR across offline and pseudo-streaming settings: JSONL manifests, Vietnamese normalization, WER/CER, real-time factor, latency proxy, bootstrap confidence intervals, and result aggregation. Produced Pareto frontier, delta WER heatmap, significance analysis, robustness views, and cloud matrix scripts for report regeneration.
- **Breast Cancer Ultrasound CAD: Sequential vs Multi-task Deep Learning** — Developed and evaluated segmentation-classification pipelines on the 780-image BUSI ultrasound dataset using U-Net/EfficientNet variants. Multi-task learning improved classification accuracy from 62.05% to 85.26% (+23.2 percentage points) while maintaining segmentation Dice around 76%; the project received BIT Genesis Research Award 2025 and was presented at NCTD 2025.
- **KKBox Real-time Customer Churn BI Dashboard** — Built a near real-time decision-support stack using Kafka log replay, Spark Structured Streaming, ClickHouse, FastAPI, and React/Vite. Smoke-tested the artifact pipeline across 906,157-row descriptive snapshot, 887,123-row validation set, and 907,471-row scored test/simulation outputs, connecting data engineering outputs to churn-risk and retention scenario views.
- **VN Bank Stock Analytics with XGBoost + LLM Reasoning** — Led a 5-member data mining project integrating market OHLCV, financial reports, macro indicators, and news sentiment for 14 Vietnamese banking stocks. Built 31 engineered features, four XGBoost models, and an LLM reasoning layer; achieved MAE 0.094/RMSE 0.119 for return forecasting and 0.98 correlation for risk forecasting.

Education

University of Economics Ho Chi Minh City (UEH)	Ho Chi Minh City, Vietnam
B.Sc. in Data Science	Oct 2023 – Mar 2027 (Expected)
Current GPA: 3.78/4.00	

Honors and Awards

• 1st Prize — Faculty-level Scientific Research Competition, Vietnamese Fake News Detection.	Oct 2025
• 3rd Prize — VNPT AI Hackathon 2025, AI meeting-minutes and RAG chatbot workflow.	Dec 2025
• BIT Genesis Research Award 2025 — Breast Cancer Ultrasound CAD with deep learning.	Oct 2025
• 3rd Prize — Humanitarian Logistics Hackathon, AI + IoT logistics optimization concept.	Apr 2025
• Encouragement Prize — National Excellent Student Competition in Chemistry.	2023

Additional Information

• Training: Member, AI Viet Nam — AIO25 Program; completed 11 verified credentials covering Python, SQL, Git, ML/DL, computer vision, NLP, and data analysis.	Jun 2025 – Jun 2026
• Languages: English — Intermediate working proficiency.	