THÁI HOÀI AN

Data Scientist Intern

hoaianthai345@gmail.com — LinkedIn — ahtportfolio.vercel.app

Summary

Data Science student with experience in multiple AI/ML and data analytics projects and research. Proficient in an end-to-end workflow: data collection, preprocessing, feature engineering, model training, evaluation, and insight communication. Seeking to apply models to deliver practical value for products and operations.

Education

University of Economics Ho Chi Minh City (UEH) — Ho Chi Minh, Vietnam
B.Sc. in Data Science — Oct 2023 – Mar 2027 (Expected)
Current GPA: 3.75/4.00

Experience

VNPT AI Hackathon — 3rd Prize — Team Project

- Designed an LLM-powered pipeline to convert meeting content into structured meeting minutes (summary, key decisions, action items).
- Built a RAG-based chatbot for document Q&A and contextual retrieval; instrumented basic auth/usage logging for iteration.
- Built and integrated a demo web app (UI + API integration) to showcase the LLM/RAG workflow.
- *Tech:* Python, LangChain, Vector DB (e.g., Qdrant/FAISS), Streamlit.

Faculty-level Scientific Research Competition — 1st Prize — Research
Topic: Vietnamese Fake News Detection (BiLSTM vs PhoBERT vs LLM)

- Benchmarked BiLSTM, PhoBERT (frozen/fine-tuned), and LLM prompting on ReINTEL; reported Accuracy/Macro-F1/AUC with error analysis.
- Achieved best result with PhoBERT fine-tuning: Accuracy 0.963, Macro-F1 0.929, AUC 0.980; analyzed deployment trade-offs (latency/VRAM).
- *Tech:* PyTorch, Transformers, scikit-learn.

Projects

- **Breast Cancer Ultrasound CAD (Segmentation + Classification)** — Multi-task vs sequential study on BUSI; Dice 0.7648 / IoU 0.6233; classification improved to Acc 0.853 (vs 0.620 sequential). *Tech:* PyTorch (U-Net + EfficientNet).
- **VN Bank Stock Analytics (Time-series + NLP/LLM-assisted insights)** — Built a multi-source pipeline for banking stocks; return regression MAE 0.0938 / RMSE 0.1189; risk forecasting Corr 0.9808. *Tech:* Python, XGBoost, time-series features.
- **Vietnamese Medical IE Pipeline (NER + Relation Extraction)** — Implemented IE workflow (labeling → training → evaluation) with semi-supervised hybrid RE; Acc 0.8125 / Macro-F1 0.6309. *Tech:* Python, Label Studio, (Pho)BERT embeddings.
- **Vietnam Weather Analytics & Forecasting (Time-series)** — Built forecasting + dashboard; binary Rain vs Not Rain Accuracy 0.836 (F1(Rain) 0.887); delivered Streamlit demo for visualization. *Tech:* Python, scikit-learn, Streamlit.

Honors and Awards

- **1st Prize** — Faculty-level Scientific Research Competition (Vietnamese Fake News Detection). *Oct 2025*
- **3rd Prize** — VNPT AI Hackathon (LLM Meeting Minutes + RAG Chatbot). *Dec 2025*
- **3rd Prize** — Faculty-level Scientific Research Competition (Breast Cancer Ultrasound CAD: Segmentation + Classification). *Oct 2025*
- **3rd Prize** — Logistics Hackathon (AI + IoT solution for logistics optimization). *Apr 2025*

Additional Information

- **Training:** Member of AI Viet Nam — AIO25 program. *Jun 2025 – Jun 2026*
- **Languages:** English: Intermediate.
- **Technical:** Python, SQL, Git; PyTorch, scikit-learn, Transformers; Computer Vision (OpenCV, image preprocessing); Streamlit; time-series modeling; NLP (NER/IE, RAG).
- **Research:** Experiment design, benchmarking, metrics (F1/AUC/IoU/Dice), ablation studies, error analysis, and technical writing.