Tri Ho Dinh

AI Research Engineer | Ho Chi Minh City, Vietnam
(+84) 901 364 158 | hodinhtri3010@gmail.com | hodinhtri.me | linkedin.com/in/trihodinh | github.com/HDT-0103

Education

University of Science, VNU-HCM — Ho Chi Minh City, Vietnam
B.Sc. in Information Technology (Information Security) ***GPA: 3.84 / 4.0*** *Sep. 2024 – Expected June 2028*
– **Availability: third-year student**; **part-time, 3–4 days/week** (from now); **full-time from Sep. 2027**.
– **Coursework:** Machine Learning, Linear Algebra, Probability & Statistics, Operating Systems, Database Systems.

Publications

L. U. Le*, **T. Ho Dinh***, C. Hoang, H. H. Ha. *"Meddies-PII: A Multilingual Framework for PII Extraction in Clinical De-identification."* arXiv:2609.12544, 2026. [arXiv] (*co-first author)

Research & Engineering Experience

Meddies — Ho Chi Minh City, Vietnam
AI Engineer (Part-time) – Research on Multilingual Clinical De-identification *May 2026 – Present*
– **Co-first author** of **Meddies-PII** (arXiv:2609.12544), a multilingual PII extraction framework for clinical de-identification spanning **17 languages** and **9 PII labels**.
– Design and maintain the **attribute-conditioned synthetic data pipeline** (12 medical document types, 11 formats, 12 scenarios, 9-LLM generator pool) that produced **1M** documents filtered by **13 deterministic validation gates**.
– The data trains a **350M-parameter** LFM2.5 BIOES token classifier with constrained Viterbi decoding, scoring **0.827 mean exact-match micro-F1** on 15 external benchmarks (vs. **0.658**) and **0.878** on a 5,100-document adversarial benchmark (vs. 0.540).
– Extending the pipeline to **20 languages** with **meddies-faker**, which samples personal attributes from country-specific **real-world frequency distributions** to better match each target population.

IaC Lab, Faculty of IT, HCMUS — Ho Chi Minh City, Vietnam
Undergraduate Researcher – Infrastructure-as-Code (IaC) Security *Feb. 2026 – Present*
– Curating a large-scale, open **IaC dataset** (Terraform, Kubernetes) from public sources and designing a classification taxonomy to serve as a reusable research benchmark.
– Evaluated static-analysis runners across **29 environments** and **1,346 Checkov rules**, studying false-positive reduction via AST context propagation; exported findings to **SARIF/CycloneDX** for CI/CD gates.

Projects

SmartATS – AI Retrieval Engine & LLM Telemetry | *FastAPI, pgvector, Azure* July 2026 – Aug. 2026
– Built an asynchronous CV parsing and extraction pipeline across **10 microservices**, with semantic search over **768-d multilingual embeddings** using **pgvector HNSW** plus boolean filters.
– Added LLM telemetry tracking **p95/p99 latency**, token usage and cost, with ABAC-based data isolation.

Insurance Fraud Detection – End-to-End ML Pipeline | *XGBoost, Docker* Feb. 2026 – Apr. 2026
– Engineered **145 behavioral features** and a leakage-safe SMOTE + XGBoost pipeline, raising recall from **18.4% to 65.0%** (**0.840 ROC-AUC**); served via Docker Compose.

Technical Skills

Languages: Python, C++, SQL, Bash
AI / ML: PyTorch, Hugging Face Transformers, Token Classification (NER, BIOES), LoRA/QLoRA, RAG, Embeddings, XGBoost
Data & Evaluation: Synthetic Data Generation, Faker, LLM-based Data Pipelines, Data Validation, Entity-level F1 Benchmarking
Systems & MLOps: FastAPI, PostgreSQL/pgvector, Docker, Kubernetes, Terraform, CI/CD, Azure, Linux
English: C1 (Oxford Placement Test)

Honors & Awards

RMIT Hackathon 2026 – **6th Globally**, **4th in Vietnam**, 3rd at Saigon South Campus (team of 4) 2026
Academic Merit Scholarship – Faculty of IT, HCMUS (two consecutive years) Jan. 2025 & Jan. 2026