Dinh Khai Nguyen

+84 37838454 nguyendinhkhaiqt2005@gmail.com ngwkhai.com github.com/ngwkhai Hanoi, Vietnam

SUMMARY

AI Application Engineer at VinSmart Future, working on AI data and its quality control, with research experience in natural language processing and LLM fine-tuning, and AI products built and deployed end to end. Over the next 3–5 years, my objective is to take on technical leadership or project management roles and to make long-term contributions to the growth of the company.

EDUCATION

VNU University of Engineering and Technology Aug 2023 - Present
Bachelor of Artificial Intelligence GPA: **3.61/4**

EXPERIENCE

VinSmart Future
AI Application Engineer Nov 2026 - Present
Take on in-depth AI data work and own its **quality control**, ensuring every dataset meets the required standard; support the management of a **12-person team**.
Junior AI Application Engineer Sep 2026 - Nov 2026
Carried out **data quality control**, reviewing data against the required standard before use; supported the management of a **six-person team**.
AI Intern May 2026 - Aug 2026
Built the **data collection and quality-control pipeline** for a travel dataset, from crawling and ingestion to the verification pass.

Viettel Networks Sep 2025 - Jan 2026
AI Intern
Researched generating **structured project layouts** from natural-language specifications, addressing the structural blindness of LLMs on hierarchical output.
Built the training set and evaluation metric, and fine-tuned compact open models with **LoRA** and a loss aligned to that metric.

PROJECTS

ResearchMap AI — *Solo* (researchmap.ngwkhai.com) 2026
Built the full product solo: a FastAPI backend on PostgreSQL with pgvector, a React frontend, Google OAuth and TOTP sign-in, quotas and an admin panel.
Designed the analysis pipeline: an LLM plans OpenAlex queries, papers are clustered with KMeans, and gaps come from topics that share no papers.
Built an **agentic RAG** chat (7 tools, at most 4 rounds) that cites only retrieved papers: **36/36** cited papers verified real, retrieval precision@5 **0.867**.
Technologies: Python, FastAPI, PostgreSQL, pgvector, scikit-learn, OpenAI API, React, TypeScript, Cloud Run.

TaxMate AI — *Team lead (team of 4)* (taxmate.ngwkhai.com) 2026
Led a four-person team building four cores (Legal RAG, Risk, Graph, Optimization); set scope and releases, and integrated them into one product.
Rebuilt legal retrieval with **BM25** over Vietnamese syllable pairs, citing only retrieved chunks: recall@5 rose from 32 to **52 of 56** questions.
Built the Risk core, multi-tenant auth on Supabase, and dev and production deployments on Cloud Run through GitHub Actions.
Technologies: Python, FastAPI, PostgreSQL, Supabase Auth, OpenAI API, React, TypeScript, Cloud Run.

Neural Machine Translation — *Leader (team of 2)* (github.com/ngwkhai/Neural-Machine-Translation) Jun 2025 - Jul 2025
Built the **first large-scale Vietnamese diacritic restoration dataset** (12.5M pairs) and implemented RNN, GRU, LSTM and Transformer **Seq2Seq** from scratch.
Achieved **BLEU 81.31** and **ChrF++ 88.57** with Transformer, significantly outperforming RNN and GRU baselines.
Technologies: Python, PyTorch, NumPy, Pandas, BLEU, ChrF++.

Optimization for AI Inference Engines on GPUs 2025
Optimized ResNet50 and DistilBERT GPU inference (FP16, INT8, torch.compile, ONNX); CUDA profiling found GEMM and ReLU bottlenecks.
FP16 raised DistilBERT throughput from 305 to 1202 samples/s at batch size 8 (**≈4×**); INT8 (TensorRT) gave **2–5×** with careful calibration.
Technologies: PyTorch, TensorRT, ONNX Runtime, TorchScript, CUDA profiling.

SKILLS

Programming	Python, C/C++, Java, SQL, TypeScript
Machine Learning	PyTorch, Scikit-learn, TensorFlow, Keras, XGBoost, Transformer, Seq2Seq, YOLO, LoRA fine-tuning
LLM & Retrieval	OpenAI API, RAG, agentic tool calling, BM25, pgvector, hybrid search
Systems & Inference	FastAPI, PostgreSQL, React, Docker, Cloud Run, GitHub Actions, Hadoop, PySpark, TensorRT, ONNX Runtime
Languages	Vietnamese: Native, English: Professional

HONORS & AWARDS

Third Prize, Student Scientific Research Award — Artificial Intelligence, University of Engineering and Technology	2026
Third Prize, Student Track — Humanitarian Logistics Hackathon	2026
Academic Encouragement Scholarship, University of Engineering and Technology	2023-2024
Second Prize, Thai Nguyen Provincial Excellent Student Contest in Physics	2023
Third Prize, Thai Nguyen Provincial Excellent Student Contest in Informatics	2021