HOANG BA MINH QUANG

AI/Machine Learning Engineer

Address: Ho Chi Minh, Vietnam, 71300
Phone: 0896208479
Email: hbminhquang.19@gmail.com
Linkedin: linkedin.com/in/quangai/
Hackerank: hackerrank.com/profile/hbminhquang_19
Github: github.com/minquan-alt

SUMMARY

AI Engineer Intern with hands-on experience building end-to-end AI systems for real-world and industrial applications. Strong background in ML/DL, NLP with proven experience deploying models via APIs and integrating AI into business systems (ERP, web platforms) under practical constraints.

EDUCATION

University of Information Technology (UIT)
Bachelor of Information System
2023 – 2027

SKILLS

- AI & Machine Learning: Machine Learning, Deep Learning, NLP, Computer Vision, Time Series Modeling.
- Generative AI: LLMs, Prompt Engineering, Fine-tuning.
- AI Engineering: End-to-end ML pipelines, model deployment, RESTful API integration, RAG systems.
- Programming: Python (PyTorch, TensorFlow, NumPy, Pandas, Scikit-Learn), Java, JavaScript.
- Backend & Data: FastAPI/Flask, REST APIs, SQL/MySQL, ETL pipelines.
- Tools & Systems: Git, Linux, Vector Databases (Qdrant).

CERTIFICATION

- Coursera - Deep Learning Specialization (DeepLearning.AI)
- Coursera - Advanced Relational Database and SQL
- Coursera - Get Started with Python

PROJECT

SignLearn - AI-Powered Sign Language Learning Platform **Oct 2025 - Dec 2025**
Sudo Code 2025 - WTM HCMC | Team of 4 | Role: Member | Top 3
Outcome: An AI-powered Vietnamese sign language learning platform featuring real-time gesture recognition, personalized learning paths, and an interactive AI tutor chatbot.

My work:

- Developed AI components for real-time gesture recognition and an NLP-based AI tutor chatbot.
- Built backend services to serve AI models and integrate them into a web-based learning platform.
- Applied RAG techniques with vector databases to enhance content retrieval and learning feedback.
- Integrated the AI components into a web-based system and optimized inference latency for real-time usage.

Source code: Github/sudo2025

Demand Forecasting and Supply Optimization for ERP **Oct 2025 - Dec 2025**

Team of 5 | Role: Leader

Outcome: Delivered an AI-powered demand forecasting solution integrated into an ERP system, enabling more accurate inventory planning, reducing stock-related risks, and supporting improved operational efficiency and revenue performance.

My work:

- Built an end-to-end machine learning system for demand forecasting using real-world ERP data.
- Developed forecasting and demand recovery models to support inventory planning and operational decision-making.
- Deployed models via FastAPI and integrated them into a web-based ERP application.
- Led my team in implementing an AI-powered business solution.

Source code: Github/ERP-Optimization

Advanced LLM Reasoning **Oct 2025 - Dec 2025**

Team of 1 | Role: Leader

Outcome: Built and deployed a cost-efficient LLM-based reasoning component by optimizing small language models for reliable performance under limited computational resources, suitable for real-world NLP applications.

My work:

- Optimized small language models (SLMs) to achieve reliable reasoning performance under limited computational resources.
- Designed and deployed an end-to-end NLP pipeline from data preprocessing to model serving.
- Demonstrated cost-efficient LLM deployment suitable for practical and industrial NLP applications.

Source code: Github/AdvancedLLMReasoning