Nguyen Quoc Khanh

☎ (+84) 039-269-7777 nqkhanh2002.github.io nqkdeveloper@gmail.com linkedin.com/nqkhanh2002

Summary

Applied AI Engineer with hands-on experience building LLM-powered applications Machine Learning, MLOps, Data Engineering. Aspiring highly motivated MSc/PhD candidate in Computer Science.

Education

University of Information Technology - VNUHCM June 2020 - Nov 2024
*Bachelor of Science in **Information System** (GPA: 3.19 / 4.00)* *Ho Chi Minh, Viet Nam*

- **Relevant Awards:** Merit scholarships for outstanding academic performance in two semesters and a full-tuition scholarship for exceptional academic achievement in one semester.

Nguyen Du High School for the Gifted 2017 - 2020
Major Subject in Physics (GPA: 3.5 / 4.00) *Dak Lak, Viet Nam*

- **Relevant Awards:** Gold Medal in Physics Olympic 10-3 (2017), Silver Medal in Physics Olympic 10-3 (2018)

Experience

Silver Tiger Technology LLC - Dubai, UAE Aug 2025 – Present
AI Engineer *LangChain, LangGraph, Tensorflow Serving, LLM systems*

- Design and deploy RAG architectures, multi-agent systems, and LLM-powered applications for enterprise clients (Chatbot, Assistant, Phishing Detection).
- Self-hosted and optimized large-scale foundation models on dedicated 8×NVIDIA H200 GPU infrastructure, enabling high-throughput, low-latency inference for enterprise AI applications.
- Evaluate new technologies and tools to improve scalability, cost efficiency, and model lifecycle automation.

Speech Processing Department - VinBigdata - HCMC, Vietnam Dec 2024 – Jun 2025
AI Engineer *Data Engineering, MLOps, ASR, TTS*

- **Data Engineer** Developed and implemented **Data crawling, Labeling, Transcript Forced Alignment (TFA)** workflows within the SpeechWorld Data Framework. Improved Grapheme-to-Phoneme (G2P) models and acoustic models to enhance forced alignment accuracy and reliability.
- **Machine Learning Engineer/MLOps** Enhanced SpeechBase system with new features including model tuning, backend improvements, and online/active learning capabilities.

Vingroup AI Engineer Training Program - HCMC, Vietnam July 2024 – Dec 2024
AI Engineer *Mathematics, ML/DL/CV/NLP, AI Ethics, MLOps*

- **Participated in the Vingroup AI Engineer Training Program**, a comprehensive 10-month course focused on building high-quality AI engineers with practical problem-solving skills to meet the technological resource needs of the Vingroup ecosystem.
- **Completed extensive training in key AI and data science areas**, including foundational and advanced machine learning, computer vision, natural language processing, AI ethics, linear algebra, probability, and statistics.

THINKPROMPT CO., LTD - HCMC, Vietnam Jul 2023 – Sep 2025
Team Leader - Data Scientist *Agile, MLOps, AWS EC2, MinIO, Docker, FastAPI, TensorFlow, Scraping*

- **Led a team of 3+ AI Engineers and Data Scientists**, designed and deployed 20+ predictive ML models for horse racing, achieving 20% WIN rate and 50% PLACE rate, leading to 150% net ROI and 40% reduction in client analysis time.
- **Implemented comprehensive MLOps workflows** with automated model training, validation, and deployment pipelines, significantly improving development efficiency and enabling rapid iteration of predictive models.
- **Architected and deployed end-to-end ML infrastructure** using AWS EC2, Docker, FastAPI, and PostgreSQL, building scalable data collection systems through web scraping and optimized database management for real-time horse racing prediction workflows.

MAICO Group May 2022 – August 2022

AI Engineer Intern *Recommendation system on real estate system*

- **Developed collaborative filtering recommendation system** for real estate platform serving 50,000+ users, achieving 25% improvement in click-through rates and 15% increase in user engagement time.
- **Implemented A/B testing framework** for recommendation algorithms, analyzing user behavior data from 10,000+ daily interactions to optimize model performance and business metrics.

Vietnam Olympiad in Informatics (VNOI) Jul 2020 – Dec 2022

Algorithmic Problem Setter *Algorithms, Tester, Problem Setter*

- Designed and tested algorithmic problems for competitive programming competitions. Conducted beta testing, considered time complexity, and provided sample input/output.
- Collaborated with other designers, generated test data, and continuously improved the competition process.

Project

AI Solution Platform - SilverTiger | *LLM, RAG, NLP, GraphRAG, Neo4j, Python, FastAPI*

- **Developed modules of enterprise LLM-powered assistant** with RAG-based knowledge retrieval, integrating foundation models for real-time multi-language question answering across 100+ knowledge sources.
- **Designed ML evaluation and data generation framework** including golden test cases, RAG benchmarking, and systematic model quality assessment, ensuring consistent performance across multi-language inputs (Vietnamese, English, and others).

Oten Trust - AI Security Platform - SilverTiger | *LLM, XAI, NLP, Python, FastAPI*

- **Led AI development of LLM-based phishing detection system** with natural language reasoning, implementing explainable AI (XAI) for transparent trust scoring and cross-language scam detection across Vietnamese and English inputs.

Sign Language Translation - Graduation Thesis | *TypeScript, HumanGAN, LLM, Skeleton Viewer, Mobile/Web*

- Developed a real-time sign language translation solution with multi-language support, leveraging cutting-edge machine learning for accurate sign-to-text and text-to-sign conversions.
- Enhanced user interaction through an intuitive interface for both desktop and mobile, enabling speech-to-text, text normalization, and internationalization across 107 languages.
- Innovated with 3D avatar animations and Human GAN technology for realistic sign language visualization, complete with video output features for sharing and accessibility.

Ventus - Horse Racing Betting System - ThinkPrompt | *MLOps, AWS EC2, MinIO, FastAPI, TensorFlow, Scraping*

- **Spearheaded development of data-driven horse racing betting system** using Agile and MLOps frameworks, processing 1,000+ races monthly with 200+ variables per horse, achieving 20% win rate accuracy and 150% net profit return.
- **Orchestrated scalable deployment architecture** with FastAPI, Docker, PostgreSQL, and MinIO on AWS EC2, delivering 100+ daily predictions with sub-300ms response times for real-time betting decisions.
- **Developed ensemble ML models using TensorFlow** and built automated reporting system serving 5+ stakeholders with real-time performance tracking and betting recommendations.

Publication & Manualscript

Adapting WavLM for Vietnamese Speaker Diarization in Real-world Conversations 05/2025

Thang Tuan Duy, Do Tri Nhan, Nguyen Van Huy, Nguyen Quoc Khanh, Phan Trung Kien, Mac Dang Khoa ***MAPR 2025***

An Orchestrated Framework for Automated Speech Data Processing and Alignment 05/2025

Nguyen Quoc Khanh, Nguyen Van Huy, Phan Van Tuan, Do Tri Nhan, Mac Dang Khoa ***Manualscript***

The Ethics of Advanced Driver-Assistance System Based Computer Vision 11/2024

Nguyen Quoc Khanh, Nguyen Hoang Trung, Nguyen Trong Hoang, Duong Thi An, Dam Van Hien ***Manualscript***

Optimizing Horse Racing Predictions through Ensemble Learning and Automated Betting Systems 10/2024

Nguyen Quoc Khanh, Nguyen Phuoc Sang, Tran Thi My Quyen, Tran Vu Anh ***Manualscript***

Computer Vision for Safety Management: A Case Study in the Construction Industry 10/2024

Nguyen Quoc Khanh ***Manualscript***