Nguyen Quoc Khanh

(+971) 52-572-4704 nqkhanh2002.github.io nqkdeveloper@gmail.com linkedin.com/nqkhanh2002

Summary

Applied AI Engineer with hands-on experience building LLM-powered applications. 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*** *Ho Chi Minh, Viet Nam*

Experience

Silver Tiger Technology LLC - Dubai, UAE Aug 2025 – Present
AI Engineer *LangChain, LangGraph, LangSmith, LLM Inference*

- Design and deploy RAG architectures, multi-agent systems, and LLM-powered applications for enterprise clients (Chatbot, Assistant, Phishing Detection).
- 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).
- 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.
- 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.
- 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.

THINKPROMPT CO., LTD - HCMC, Vietnam Jul 2023 – Sep 2025
Team Leader - AI Engineer *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**, building scalable data collection systems through web scraping and optimized database management for real-time prediction workflows.
- **Developed ensemble ML models** and built automated reporting system serving 5+ stakeholders with real-time performance tracking and 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***