

Tran Huu Thuc

LEAD AI ENGINEER · GENAI & SPEECH SPECIALIST

Ho Chi Minh City, Vietnam

[\(+84\)832935181](tel:+84832935181) | thuc.th.qt@gmail.com | cv.tahutu.tech | github.com/thuc.th.qt | linkedin.com/in/thuc.th.qt

Profile

Lead AI Engineer with a strong product mindset, bridging cutting-edge AI research and enterprise production scale. Proven track record architecting real-time multi-agent conversational systems, multilingual speech language models, and high-throughput vision platforms. Extensive experience across the full AI lifecycle: from model training and fine-tuning to sub-second inference optimization, hardware acceleration, and resilient MLOps architecture.

Experience

Cake Digital Bank

Ho Chi Minh City, Vietnam

Lead AI Engineer

Jun 2026 - Present

Senior AI Engineer

May 2024 - Jun 2026

VNG Digital Business – AI Lab

Ho Chi Minh City, Vietnam

AI Engineer

Apr 2023 - Apr 2024

VNG TrueID

Ho Chi Minh City, Vietnam

AI Engineer

Aug 2021 - Apr 2023

VNG Cloud

Vietnam

Associate Software Engineer (During university)

Nov 2020 - Aug 2021

Projects

High-Volume Real-Time Multi-Agent Voice Systems

Dec 2024 - Present

Role: Lead AI engineer

Architected and deployed real-time multi-agent systems for enterprise call centers, serving 10,000+ daily calls with <2s end-to-end latency.

Engineered state-of-the-art multilingual speech language model (SLM), achieving 3rd place globally at Interspeech 2026.

Built modular agent core logic orchestrating TTS, STT, VAD, and LLM backends with streaming low-latency infrastructure and comprehensive telemetry.

Fraud Detection in eKYC & Liveness Verification

Apr 2024 - July 2025

Role: AI engineer (main PIC)

Main point of contact driving FIDO Alliance certification for the eKYC flow, achieving iBeta Level 1 (2024) and Level 2 (2025) presentation-attack detection (PAD) compliance with a passive liveness solution.

Achieved 1st place in the VLSP 2025 speech quality assessment competition.

Enterprise Generative AI & Diffusion Platform

Feb 2023 - Mar 2024

Role: Technical lead / AI engineer

Led end-to-end development of large-scale text-to-image systems based on Stable Diffusion for commercial and creative production.

Reduced overall inference latency by over 40% via targeted hardware optimization, dynamic batching, and asynchronous queuing.

Deployed high-throughput serving on NVIDIA H100 GPUs using Triton Inference Server, monitored with Prometheus and Grafana.

TrueID Biometrics & Facial Recognition System

Aug 2021 - Apr 2023

Role: Researcher, Computer Vision engineer

Engineered the 'TrueID' facial recognition and ID verification system, achieving a top-56 global ranking on the prestigious NIST FRVT 1:1 benchmark on its inaugural submission.

Researched and trained face embedding models serving millions of daily identities in production face authentication systems.

Identified and implemented optimized loss functions outperforming state-of-the-art methods across five public benchmark datasets.

Owned AI-core computer vision components for eKYC: ID spoofing detection, card integrity verification, and occlusion handling.

LLM Meeting Minutes & Intelligence Platform

July 2023 - Jan 2024

Role: AI engineer

Engineered production LLM pipelines leveraging GPT-4 and Claude to extract structured action items, summarize transcripts, and identify key discussion agendas.

Implemented resilient monitoring and diagnostics using Elasticsearch and Kibana for system reliability.

Awards

Jun 2026 **3rd Place**, Interspeech 2026 (2nd MLC-SLM) — Multilingual Conversational Speech Diarization & Recognition

Oct 2025 **1st Place**, VLSP 2025 — Track 1: Speech Quality Assessment

paper

2022 **Top-56 Global**, NIST FRVT 1:1 Benchmark — Facial Recognition inaugural evaluation

Education

Skills

Generative AI & Voice	Multi-Agent Systems, Speech Language Models (SLM), TTS, STT, VAD, LLM Orchestration, Prompt Engineering
Machine Learning & CV	PyTorch, TensorFlow, ONNX, TensorRT, Stable Diffusion, Face Recognition, Anti-spoofing (iBeta Level 1 & 2)
MLOps & Serving	Triton Inference Server, NVIDIA GPU Stack (H100), Docker, Prometheus, Grafana, Kibana, GCP, AWS
Engineering & Systems	Python, C++, gRPC, Microservices, Redis, MySQL, RabbitMQ
Certificates	TOEIC 715, Neural Networks and Deep Learning