TUAN ANH NGUYEN

Senior AI Engineer — LLMs, Computer Vision & Production AI Systems

CONTACT

Hanoi, Vietnam · Remote
tuananhhust05@gmail.com
+84 339 170 155
tuananh.it.com · Portfolio

PROFILES

LinkedIn	linkedin.com/in/tuấn-anh-nguyễn-939b97317
GitHub	github.com/tuananhhust05
Hugging Face	huggingface.co/Tuan050901
Upwork	upwork.com/freelancers/~01d5a83ab3a8f17b9f

PROFESSIONAL SUMMARY

Senior AI & Full-Stack Engineer with **5+ years of experience** building production software, AI-powered applications, scalable backend systems, and computer vision solutions. Designs and delivers **LLM applications, RAG systems, AI agents, computer vision, voice AI, real-time applications, and scalable APIs**, from model fine-tuning and GPU infrastructure to backend, frontend, deployment, monitoring, and production optimization.

Works with **international clients and distributed teams**, including founders, CTOs, and product teams. Completed **26+ Upwork projects and 1,100+ hours** with a **100% Job Success Score and 4.9/5 average client rating**.

AI/ML Engineering · LLM Apps & AI Agents · RAG & Vector Search · Computer Vision · Voice & Conversational AI · Backend & API Architecture · Microservices & Distributed Systems · GPU Infrastructure & AI Deployment · DevOps & Cloud · Real-Time Communication · Performance Optimization · Monitoring & Observability

PROFESSIONAL EXPERIENCE

HTI Group · Artificial Intelligence Engineer

Apr 2026 – Present

Hanoi, Vietnam · Remote · Part-time

- Designed and developed scalable **Camera AI & Video Management Systems (VMS)** for real-time video analytics, scaling from **200 to thousands of cameras** operating 24/7 at 25 FPS with **>90% accuracy**.
- Deployed and optimized **production AI workloads across multiple NVIDIA GPU types (RTX 4090, RTX 5090, RTX PRO 6000)** using CUDA, NVIDIA drivers, Docker, and Kubernetes.
- Optimized **YOLO-based real-time object detection pipelines** using **NVIDIA TensorRT, FP16 precision, and GPU acceleration** to improve inference speed, reduce latency, and enhance GPU resource utilization for scalable computer vision applications.
- Trained, fine-tuned, and deployed computer vision and multimodal models using YOLO, YOLO Pose, VLMs, LLMs, and MediaPipe, including:
  - **Fight Detection VLM:** fine-tuned Qwen3-VL-4B with 4-bit QLoRA (0.74% of weights trained). Cut errors from 38 to 4 on 322 held-out images; **accuracy 88.2% → 98.8%** across 5 seeds.
  - Built a perceptual-hash audit to catch train/test leakage and fixed an image-resolution bug, cutting visual tokens from ~1,450 to ≤390 per image; **~110 ms p95 latency** (offline benchmark, single RTX 4090).
  - **Speech Moderation LLM:** fine-tuned Qwen2.5-1.5B with QLoRA (3.4 GB VRAM) to output structured JSON verdicts, severity levels, and flagged terms. Beat zero-shot, lexicon, and TF-IDF baselines: **2× term-level F1** vs. zero-shot; invalid JSON 28 → 1.
  - Found and fixed a dataset-parsing bug affecting ~4,000 training records; raised **throughput 67×** (4.4 → 294 req/s, 111 ms p50) via LoRA merging and vLLM (offline benchmark, RTX 4090).
- Built and maintained backend services and RESTful APIs in Python/FastAPI for production AI applications.
- Developed frontend and desktop applications with React.js and Electron, integrating AI features into user-facing products.
- Owned the full AI lifecycle: training, fine-tuning, inference optimization, backend, frontend integration, and infrastructure deployment.

Tech: Python, FastAPI, React.js, Electron, YOLO, VLM, LLM, MediaPipe, CUDA, TensorRT, NVIDIA GPU, Docker, Kubernetes
Product: Applications for national defense and security (Government of Vietnam)

Stippl · Senior AI Engineer

Oct 2025 – Oct 2026

Netherlands · Remote · Freelance / Part-time

- Designed and developed an **AI-powered travel chatbot** for a platform with **100k–200k users**, understanding travel intent and recommending POIs, hotels, and travel experiences in context.

- Built and optimized AI workflows integrating **LLMs, travel APIs, search services, and contextual trip data**; integrated **Google Gemini** with **Serper** (web search) and **Nuitee** (hotel and travel data).
- Researched and developed facial recognition applications for real-world AI systems.
- Developed and maintained backend services with **FastAPI and NestJS** for scalable AI and travel workflows.
- Designed and ran **unit, integration, and stress tests** to validate chatbot reliability and behavior under high load.
- Implemented production monitoring and AI observability with **Langfuse** (LLM traces, latency, token usage, pipeline performance).
- Containerized and deployed production services with **Docker and Fly.io**; worked with product and engineering teams on production issues and UX optimization.
- Worked with **Fly.io** platform infrastructure, covering core **server, container, storage, and networking** services.

Tech: Python, FastAPI, TypeScript, NestJS, Gemini, Serper, Nuitee, Langfuse, Docker, Fly.io
Product: stippl.io

ForSkale · Chief Technology Officer (Part-time) — Jul 2025 – Aug 2026

Italy · Remote · Led a team of ~5 engineers

- Led architecture and development of AI-powered **sales and marketing products**, owning the stack across frontend, backend, AI systems, and DevOps.
- Built automated marketing systems with scalable **pipelines and workflow automation**.
- Developed **AI meeting agents** that join meetings, capture recordings and transcripts, and use **OpenAI, Claude, and Llama** to generate insights, analysis, and marketing strategies.
- Designed and implemented **RAG systems** over user-specific knowledge for contextual analysis and recommendations.
- Built **AI voice agents** with real-time voice via **Twilio and ElevenLabs**.
- Designed scalable AI services and production infrastructure using **microservices, containerization, and cloud-native technologies**.
- Worked with **AWS** infrastructure: **EC2** (servers), **ECS and ECR** (containers), **S3 and EBS** (storage), and **VPC, Route 53, and Application Load Balancer** (networking).

Tech: React.js, Node.js, Express.js, Python, FastAPI, MongoDB, Weaviate, Docker, Kubernetes, AWS, OpenAI, Claude, Llama, Whisper, LangChain, LangGraph, Twilio, ElevenLabs, Vexa AI
Products: forskale.com, forskale.it

Upwork · Freelance Full-Stack Developer / AI Engineer — Apr 2024 – Sep 2026

Remote · International clients

- Completed **26+ projects and 1,100+ hours** for international clients with a **100% Job Success Score and 4.9/5 average rating**.
- Delivered production-grade **LLM applications, RAG systems, computer vision, voice AI, automation platforms, and scalable backend infrastructure** across multiple industries.
- Worked directly with founders, CTOs, product and engineering teams to turn business requirements into technical solutions, owning projects from architecture through deployment and production support.
- Solved engineering problems in **system architecture, AI performance, GPU optimization, scalability, and production reliability**.
- Worked with **AWS** infrastructure for client projects: **EC2** (servers), **ECS and ECR** (containers), **S3 and EBS** (storage), and **VPC, Route 53, and Application Load Balancer** (networking).
- Worked with **PostgreSQL** relational databases for client backend services alongside MySQL and MongoDB.

Key client: Lighthouse Education (United States) — PHP Developer · Mar 2025 – Sep 2026

- Developed and deployed web applications using **PHP and CodeIgniter**.
- Built **AI-powered educational chatbots** for interactive learning, integrating **Google Gemini APIs**.
- Developed backend services in **PHP and Python FastAPI**, integrating LLM capabilities into production; data and knowledge infrastructure on **MySQL, MongoDB, and Weaviate**.

Tech: Python, FastAPI, Node.js, React.js, PHP, CodeIgniter, MySQL, PostgreSQL, MongoDB, Weaviate, Docker, Kubernetes, AWS, Gemini, LLMs, RAG, Computer Vision, Voice AI
Products: Mobile: Justvital (Google Play, App Store), Jobsy (Google Play, App Store) · Web: gobigblog.com, jobsy.dk, shopgt.com, easyforms.pro, okinihome.com

HTC Vietnam · Product Manager & AI Software Developer

Jul 2023 – Apr 2026

Vietnam · Remote

- Designed and developed scalable API systems with **Python FastAPI and Node.js**; designed resilient infrastructure on **Kubernetes and Docker**.
- Managed project workflows, assigned tasks, tracked KPIs, and guided engineering teams; defined product strategy and technical roadmaps with company leadership.
- Researched and implemented **database sharding** for large-scale data processing.
- Developed real-time communication systems with **MQTT, WebSocket, and WebRTC** for chat, VMS, and attendance applications.
- Led AI/ML R&D across **object detection, facial recognition, vectorization, OCR, audio processing, and LLM-based applications**; built vector search on **Elasticsearch and Pinecone**.
- Built and maintained web and mobile applications with **React and Flutter**; supported automated data crawling systems.

Tech: Python, FastAPI, Node.js, React, Flutter, Kubernetes, Docker, MongoDB, Elasticsearch, Pinecone, MQTT, WebSocket, WebRTC

Products: Applications for national defense and security (Government of Vietnam)

Hưng Hà Payment JSC · Technical Manager & Full-Stack Developer

Aug 2021 – Jan 2025

Hanoi, Vietnam

- Managed and supervised engineering tasks for an app with **~1M registered users**, providing technical support and resolving complex implementation issues.
- Developed a real-time video calling application with **WebRTC, MQTT, and WebSockets**, and desktop applications with **Electron.js**.
- Led migration of legacy systems from **PHP/MySQL to Node.js/MongoDB** for scalability and maintainability.
- Built and deployed **Next.js** applications focused on performance and SEO: **~200 ms page load, Google PageSpeed >90**.
- Managed production infrastructure (**Nginx, PM2, Docker, firewall, domains**); applied distributed storage and system-level optimizations; supported AI-driven applications.

Tech: Node.js, MongoDB, Next.js, PHP, MySQL, Electron.js, WebRTC, MQTT, Docker, Nginx, PM2

Products: Web: timviec365.vn, viecday365.com, joblike365.com, hungha365.com · Mobile: Timviec365 Company app (Google Play, App Store)

PROJECTS

TravelViet · Founder & Lead Engineer

travelviet.org

Vietnam travel platform · Personal venture

- Built a Vietnam travel platform combining a social feed, an **AI trip planner**, real-time messaging, and portals for travelers, agencies, businesses, and guides.
- Designed the AI planner as a **FastAPI RAG service** (Gemini, Serper, Qdrant) with streamed answers that turn into editable day-by-day trips; accent-insensitive Vietnamese search over a catalog of 4,400+ places (Elasticsearch).
- Owns the full stack and operations: Node.js/Express API, Socket.IO, BullMQ, MongoDB, Redis, Next.js web, a Flutter mobile app in development, Docker Compose, Nginx, and Cloudflare edge caching.

Tech: Next.js, TypeScript, Node.js, Express, Socket.IO, BullMQ, Python, FastAPI, Gemini, Qdrant, Elasticsearch, MongoDB, Redis, Flutter, Docker, Nginx, Cloudflare

StudyVent · Maintainer & Operator

studyvent.org

Research platform for students · Taken over and operated

- Took over and operate **StudyVent**, a research platform where students search, browse, and share **term papers, essays, bachelor and master theses**, with **tens of thousands of users and documents**.
- Maintain the **PHP** backend and **Angular** frontend: topic search, paper uploads, subscriptions, and the knowledge base.

Tech: PHP, Angular

Red Panda & KumaTomo · AI Education Apps

WeMeet Limited (Hong Kong)

AI learning apps for students and children · iOS & Android

- **Red Panda App:** AI tutor that guides students step by step (homework help from photos, Math, English, Science, Kanji, exam prep). Built features including **book reading with read-aloud**, poetry reading, an **interactive whiteboard** (incl. Word documents), chat sessions with session summaries, and classroom challenges.

- **KumaTomo:** AI chat buddy for kids with voice chat, bilingual mode, whiteboard and drawing, a reading library, and a companion parent app (KumaGuardian).

Tech: Flutter, Firebase, Cloud Functions, TypeScript

Products: Red Panda (App Store, Google Play) · KumaTomo (App Store, Google Play)

TECHNICAL SKILLS

AI / ML	LLMs, RAG, AI Agents, Fine-tuning (QLoRA/LoRA), vLLM, Computer Vision, Object Detection, YOLO, YOLO Pose, VLMs, OCR, Facial Recognition, Audio Processing, Voice AI, Whisper, Prompt Engineering
LLM / AI Ecosystem	OpenAI, Claude, Llama, Qwen, Google Gemini, LangChain, LangGraph, Weaviate, Pinecone, Elasticsearch, Langfuse
Backend	Python, FastAPI, Node.js, NestJS, Express.js, PHP, CodeIgniter
Frontend	React.js, Next.js, Angular, Flutter, Electron.js
Databases	MongoDB, MySQL, PostgreSQL, Elasticsearch, Weaviate, Pinecone
Infra & DevOps	Docker, Kubernetes, AWS, Ubuntu, Nginx, PM2, Fly.io, Firebase, CUDA, NVIDIA GPU Infrastructure
Real-Time Systems	WebRTC, WebSocket, MQTT
Third-Party APIs	Serper, Nuitee, Twilio, ElevenLabs, Stripe