Anh Dung CONG

0373-716-958 | dungca1512@gmail.com | Portfolio | Blog | Linkedin | GitHub

Summary

AI/MLOps engineer, 3+ years shipping production AI end-to-end: Terraform/Ansible provisioning, Kubernetes (GKE + bare-metal kubeadm), GitOps CI/CD, GPU model serving. Delivered speech (ASR, pronunciation/tone scoring) and embedding services on GCP and DigitalOcean for users in VN, JP, KR and CN. Cost-conscious by default – GPU right-sizing, CPU-first inference, self-hosted alternatives to paid APIs.

Education

VNU University of Science (HUS) — Hanoi, Vietnam
Computer and Information Science — *Mar 2019 – Jun 2025*
Good academic achievement

Experience

AI/ML Systems Architect — Jan 2025 – Present
eUp Group — *Hanoi, Vietnam*

AI Infrastructure & MLOps ownership

Containerized and operated the production ML stack (Docker, Compose, Swarm) on GKE and DigitalOcean, with releases gated by Jenkins signed tags plus GitLab CI (pytest, DinD image hygiene).

Ran production ML services for VN/JP/KR – faster-whisper/CTranslate2 ASR, TTS, wav2vec2 scoring, Qwen3 embeddings – including a self-hosted OpenAI-compatible `/v1/embeddings` on RTX 4080 that replaced the paid API.

Cost engineering: benchmark-driven GPU sizing – measured **p95 1.86 s at 20 concurrent users** with only ∼8% GPU utilization, so a commodity CUDA GPU ($150–370/mo) replaced an H100 plan (∼$2,475/mo); ruled CPU-only out with data (∼1 req/s throughput wall from GIL and memory bandwidth) and ran RCA on a 3x ASR latency gap.

Multi-market Speech Scoring Platform (JLPT, TOPIK, HSKK, English)

Four FastAPI + Gunicorn services on one architecture: multi-engine STT with fallback (Kotoba-Whisper, faster-whisper/CT2, SenseVoice ONNX, ReazonSpeech), Praat/`parselmouth` analysis under a concurrency semaphore, parallel scoring pipelines.

Language-specific scoring: Needleman–Wunsch alignment (char/mora/bigram) and pitch accent (JA, SudachiPy/MeCab/pykakasi), Mandarin tone classification (ZH), wav2vec2 CTC GOP with espeak-ng G2P (EN) – scores deterministic, GPT-4o-mini only for word-level feedback.

AI Content & Translation Systems

Hey Translate: parallel multi-LLM translation (Claude, Gemini, GPT, Qwen) with strict row-alignment validation (no silent merge/drop) and an LLM-as-judge stage picking the best rendition.

Ultimate Lesson: Japanese lesson generation on Gemini 2.5 Flash with a fine-tuned Flash Lite extractor on Vertex AI, quality-check/auto-retry, Redis cache and multi-key rotation with 429 cooldown; plus a subtitle pipeline (yt-dlp → Whisper → GPT-4o-mini translation, JLPT vocab) as an async FastAPI job API.

AI Engineer — Nov 2024 – Jan 2025
AMELA Technology — *Hanoi, Vietnam*
R&D: Japanese handwriting OCR, RAG for chatbots, AI-guided automatic cosmetic eyeliner.

AI Engineer Intern — Mar 2023 – Nov 2024
FPT Smart Cloud — *Hanoi, Vietnam*

FPT AI Enhance (Home Credit, FE Credit, MB Bank, FPT Long Chau, FPT Shop): built and deployed the pipeline; big-data processing, model training/evaluation for automatic call-center scoring (NLP), automation via n8n.

Guardrails for LLMs: developed and tested guardrails improving LLM safety and controllability.

Virtual assistant chatbots: development and deployment for the State Securities Commission of Vietnam, Home Credit and FE Credit.

Projects

AI Gateway 2025

Personal Project *Java 21, Spring Boot WebFlux, FastAPI, Terraform, GKE*

Java (Spring Boot WebFlux) gateway + Python (FastAPI) worker for unified AI request management: multi-provider routing (OpenAI, Gemini, Claude, local LLM) with fallback, retry, circuit breaker and rate limiting (Bucket4j).

Worker serves local LLM inference, embeddings and a RAG pipeline (document indexing + search), with Redis cache and Prometheus metrics; Docker Compose locally, Kustomize on GKE.

Replaced a manual `gcloud` runbook with Terraform (API enablement, GKE Autopilot, Artifact Registry, static IP outside the cluster lifecycle): rebuilds with `apply`, near-$0 with `destroy`.

Whisper Japanese ASR Finetune 2025

Personal Project *PyTorch, HuggingFace, LoRA, Kaggle, GitHub Actions*

LoRA fine-tune of Whisper Tiny for Japanese ASR on ReazonSpeech, exported to faster-whisper (INT8) for cheap inference.

Full CI/CT/CD MLOps loop: GitHub Actions orchestration, Kaggle training, HuggingFace Hub hosting, quality gate for model promotion.

Published 3 Whisper-based Japanese ASR models on HuggingFace Hub (`dungca`); the LoRA variant reached 40+ downloads.

Homelab Kubernetes & GitOps Platform 2025 – 2026

Personal Project *kubeadm, ArgoCD, Prometheus, Grafana, Loki, MetalLB*

Bootstrapped a 3-node Kubernetes v1.31 cluster by hand with `kubeadm` (control-plane, etcd, kubelet, kube-proxy) instead of a managed distro: Flannel CNI, MetalLB (L2), ingress-nginx, local-path storage, self-hosted registry.

GitOps via ArgoCD App-of-Apps (multi-source, prune + selfHeal), so deploys are `git push` only – no SSH, no manual `kubectl apply`.

Observability with kube-prometheus-stack + Loki/Promtail, plus ∼2,400 lines of engineering notes, each lesson tied to a real cluster failure.

Raspberry Pi Homelab – Ansible IaC & Slack ChatOps 2026

Personal Project *Ansible, Docker, AdGuard Home, Cloudflare, Slack API*

Turned a hand-configured Raspberry Pi serving the whole LAN into IaC: one idempotent Ansible playbook rebuilds it from a bare OS (`fstab` mounts, Docker CE, journald caps, cloudflared, Tailscale, AdGuard Home, cleanup cron), host config snapshotted into git.

LAN-wide DNS on AdGuard Home (DoH upstreams, VN blocklists, safebrowsing/parental, named clients); measured the resolver and cut latency **199 ms** → **27 ms**.

Slack ChatOps on a Cloudflare Worker relay: `/homelab` slash command for live status, Block Kit alerts separating power loss from internet loss with outage duration and diagnosed cause, SSH-login notices, camera snapshots, weekly digest.

CI gates with ansible-lint, yamllint and gitleaks; GPG-encrypted config backups and a recovery runbook make a dead SD card a rebuild, not an investigation.

Technical Skills

Programming: Python, Java (17/21), Scala 3, TypeScript/Node.js, Bash, SQL

GCP: GKE (Autopilot + Standard, preemptible pools, Workload Identity), Compute Engine, Artifact Registry, Filestore, Cloud Load Balancing, IAM, Vertex AI – provisioned with Terraform

Other Cloud & Networking: AWS (EC2, EBS, S3, Lambda, Athena, CloudWatch, SNS, IAM), DigitalOcean, Oracle Cloud (OCI), Cloudflare (Workers, Pages, named Tunnel), vast.ai, Tailscale, AdGuard Home / DoH

Containers & Orchestration: Docker, Compose, Swarm, Kubernetes (GKE + bare-metal kubeadm), Helm, ingress-nginx, MetalLB, Flannel, Harbor / self-hosted registry

IaC & CI/CD: Terraform, Ansible (+ ansible-lint), GitHub Actions, GitLab CI (DinD, MR pipelines), Jenkins (tag-gated release), ArgoCD GitOps (App-of-Apps), Qodana, gitleaks

Observability: Prometheus, Grafana, Alertmanager, Loki, Promtail, kube-prometheus-stack

MLOps: HuggingFace Hub, LoRA/PEFT fine-tuning, CTranslate2 & ONNX INT8 quantization, CI/CT/CD training pipelines, model-promotion gates, GPU sizing & cost modeling

Speech & NLP: faster-whisper/CTranslate2, Kotoba-Whisper, SenseVoice ONNX, ReazonSpeech, wav2vec2 CTC forced alignment & GOP, Praat/parselmouth, espeak-ng G2P, SudachiPy/MeCab/pykakasi, librosa, ffmpeg

LLM & Agents: LangChain, LangGraph, Google ADK, MCP, A2A Protocol, RAG (BM25/FTS5 + vector), LLM-as-judge, grounding guardrails, multi-provider routing with fallback, Ollama, llama.cpp

Frameworks & Data: PyTorch, TensorFlow, Transformers, FastAPI, Spring Boot (WebFlux), Next.js; Kafka, Spark Streaming, Delta Lake, Elasticsearch, Redis, PostgreSQL, MySQL, MongoDB, SQLite, LanceDB

Tools & Languages: Git, Linux, n8n, Makefile, pytest, Ruff; English B1 (technical reading and communication)

Certification: AWS Certified Solutions Architect – Associate (SAA-C03) – in progress

References

Bui De Kien – CTO, eUp Group – 0941-238-989 – kienbd@eupgroup.net

Tran Trung Hieu – Senior AI Engineer, NLP team, FPT Smart Cloud – 0945-352-916 – hieutt125@fpt.com

Nguyen Duc Quang – Senior Developer, Team leader, AMELA Technology – quang.nguyen@amela.vn