LE HOANG TRIET THONG

AI Engineer — LLM & ML Deployment on AWS

Tan Phu, Ho Chi Minh City • (+84) 947 685 335 • lhtthong.forwork@outlook.com
github.com/AlecVOV • linkedin.com/in/lehoangtrietthong • https://chilonthon.com/

SUMMARY

AI engineer focused on taking ML and generative AI models from experimentation to production on AWS. Built and deployed four end-to-end systems covering automated retraining with SageMaker Pipelines, FastAPI model serving on Lambda, guardrailed Amazon Bedrock agents with RAG over vector databases, and INT8-quantised transformer inference via ONNX. Designs for security and cost first, with serverless stacks running at under $1/month idle. Professional English (IELTS 7.0).

TECHNICAL SKILLS

Languages	Python, TypeScript, JavaScript, SQL
ML & DL	scikit-learn, XGBoost, CatBoost, LightGBM, Hugging Face (DistilBERT), ONNX Runtime (INT8 quantisation), BiLSTM / 1D CNN, pandas, NumPy
LLM & GenAI	Amazon Bedrock (Agents, Guardrails, Titan Embeddings), RAG, prompt engineering, LLM output QA, vector databases (Pinecone, pgvector)
Serving & Backend	FastAPI, Node.js / Express, REST APIs, API Gateway, AWS Lambda (Mangum, container images)
Cloud & MLOps	AWS SageMaker Pipelines, Lambda, ECR, S3, CloudFront, Cognito, EC2; event-driven serverless
DevOps & Quality	Docker, Docker Compose, Git/GitHub, CI/CD (Vercel), Vitest, Cypress, ESLint, Cloudflare Tunnel
Databases	PostgreSQL (Supabase, Row-Level Security), MySQL, DynamoDB, Pinecone
Frontend	Vue 3, Nuxt, React, Tailwind CSS

PROJECTS

Focus Mode — AI-Augmented Productivity Platform *Amazon Bedrock, Lambda, ONNX, Supabase, Nuxt 3*

- Deployed a guardrailed Amazon Bedrock agent (Claude Haiku 4.5) behind six Lambda microservices, giving conversational task management through a four-operation action group.
- Optimised inference by quantising a DistilBERT emotion classifier to ONNX INT8 and serving it inside Lambda, removing any dependency on external AI APIs and all per-inference cost.
- Built a retrieval-augmented recommender embedding curated media with Amazon Titan into pgvector, matching content to the user's post-session emotional state.
- Secured the AI layer for production: Bedrock Guardrails against prompt injection, in-Lambda ES256 JWT verification and Postgres Row-Level Security.

5G Network Quality & Predictive MLOps Platform *SageMaker Pipelines, FastAPI, Docker, Vue 3*

- Built an end-to-end MLOps pipeline over 1.5M+ rows of raw 5G telemetry (138 drive-test files), with AWS SageMaker Pipelines automatically retraining models on every data upload.
- Engineered the data pipeline: GPS validation, bounding-box filtering, outlier capping, 1-second resampling and 50+ features on a temporal train/test split.
- Benchmarked XGBoost, CatBoost, Bidirectional LSTM and 1D CNN for 12-hour throughput and latency forecasting; CatBoost best at average $R^2 = 0.919$. Selected GMM ($k = 3$, silhouette 0.419) over K-Means for coverage-quality segmentation.
- Served models via FastAPI on Lambda (Mangum, ECR container image) behind CloudFront with Origin Access Control, at under $0.50/month.

Automated Translation Assessment System (FCAJ) *Amazon Bedrock, Pinecone, Lambda, Cognito, Vue 3*

- Built a RAG evaluation service replacing manual grading of English-to-Vietnamese AWS technical translations for FCAJ members with an automated workflow.
- Designed multi-stage Bedrock prompt chains over a Pinecone vector store to classify errors by type (grammar, terminology, tone, accuracy) and severity, with a per-article score and suggested correction.
- Brought serving cost to about $15 per 1,000 assessments with an event-driven serverless architecture behind API Gateway; secured admin access with Cognito (email + OTP).
- Set up the quality gate (Vitest, Cypress end-to-end, ESLint) and ran QA on LLM outputs to raise detection precision.

Reconnaise.ai — ML Software Vulnerability Detector *FastAPI, Docker Compose, scikit-learn, React*

- Built a hybrid detector combining static rules with a weighted-vote ensemble of six models (Random Forest, XGBoost, LightGBM, Gradient Boosting, Logistic Regression, Naive Bayes), trained on 47,858 balanced samples from five public corpora including DiverseVul and CVEfixes.
- Served predictions through a FastAPI REST endpoint returning verdict, confidence, per-model breakdown and matched rule for SQL injection, XSS, command injection and path traversal.
- Containerised the full stack with Docker Compose and deployed it on a self-hosted server exposed securely via Cloudflare Tunnel.

Experience

AWS Study Group **Jan 2025 – Present**
AI/ML Ambassador *Ho Chi Minh City*

- Produce study material and briefings on generative AI practice on AWS, including prompt engineering frameworks and prompt optimisation, for the community.
- Completed a capstone integrating multiple AWS services, and contribute to technical workshops and community events.

Education

Swinburne University of Technology **Sep 2022 – Oct 2026**
Bachelor of Computer Science, Major in Artificial Intelligence *Ho Chi Minh City*

- Pioneer Scholarship (20%). Thesis: a generative AI system that batch-scans student report submissions to assess translation quality.
- Relevant coursework: Applied Machine Learning, Intelligent Systems, AI for Engineering, Cloud Computing, Data Science Principles.

AI Vietnam — AIO 2024 Specialised Programme **Jan 2024 – May 2025**
Mathematics for AI, machine learning, deep learning and applied projects *Online*

Certifications

- **AWS Certified AI Practitioner** — *July 2026* • **AWS Certified Cloud Practitioner** — *June 2026*
- **AWS Certified Data Engineer – Associate** — *in progress, exam December 2026*
- **IELTS Academic** — Overall Band 7.0, *August 2026*

Leadership

- **ITea Lab**, Head of Media Department (Jan 2025 – Jan 2026) — led the lab's knowledge-sharing content strategy across Facebook and LinkedIn.
- **SMACK Vovinam Club**, Vice President & Head of HR (Mar 2023 – Aug 2024) — co-led club operations and acted as liaison with university administration.