Le Hoang Triet Thong

Generative AI Engineer — LLM Applications on AWS

Ho Chi Minh City | (+84) 947 685 335 | lhtthong.forwork@outlook.com
github.com/AlecVOV | linkedin.com/in/lehoangtrietthong | chilonthon.com

SUMMARY

Generative AI engineer who builds LLM applications on AWS from prototype to production. Shipped guardrailed Amazon Bedrock agents, RAG pipelines over Pinecone and pgvector, and LLM evaluation workflows on event-driven serverless stacks, designed for security and near-zero idle cost. AWS Certified AI Practitioner; professional English (IELTS 7.0).

TECHNICAL SKILLS

Generative AI & LLMs: Amazon Bedrock (Agents, Action Groups, Guardrails, Titan Embeddings), Claude models, RAG, prompt engineering and chaining, LLM output evaluation, prompt-injection defence; vector databases (Pinecone, pgvector)
AWS Cloud: Lambda, API Gateway, S3, CloudFront, Cognito, ECR, DynamoDB, SageMaker Pipelines, EC2; event-driven serverless design
Backend & Serving: Python, FastAPI, Mangum, Node.js / Express, TypeScript, REST APIs, JWT auth, Row-Level Security
ML & Inference: scikit-learn, XGBoost, CatBoost, LightGBM, DistilBERT, ONNX Runtime (INT8 quantisation), pandas, NumPy
DevOps & Frontend: Docker, Docker Compose, Git/GitHub, CI/CD, Vitest, Cypress; 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 users conversational task management through a four-operation action group.
- Built a retrieval-augmented recommender that embeds curated media with Amazon Titan into pgvector and matches content to the user's post-session emotional state.
- Ran a quantised DistilBERT classifier (ONNX INT8) inside Lambda with zero per-inference API cost, and hardened the agent against prompt injection with Bedrock Guardrails, ES256 JWT verification and Row-Level Security.

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

- Built a RAG evaluation service that replaced manual grading of English-to-Vietnamese AWS technical translations for FCAJ members with an automated internal workflow.
- Designed multi-stage Bedrock prompt chains over a Pinecone vector store that classify errors by type (grammar, terminology, tone, accuracy) and severity, returning a score and suggested correction.
- Brought cost to about $15 per 1,000 assessments on an event-driven serverless backend behind API Gateway; ran QA/QC on LLM outputs to raise precision, with Cognito-secured admin access and Vitest/Cypress tests.

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

- Built an MLOps pipeline over 1.5M+ rows of 5G telemetry where SageMaker Pipelines retrains models on every data upload; CatBoost led a 4-model benchmark at average $R^2 = 0.919$.
- Served predictions through FastAPI on Lambda (Mangum, ECR container image) behind CloudFront with Origin Access Control, at under $0.50/month.

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

- Built a hybrid detector combining static rules with a six-model weighted-vote ensemble trained on 47,858 samples, served via a FastAPI endpoint and self-hosted with Docker Compose behind Cloudflare Tunnel.

EXPERIENCE

AWS Study Group — AI/ML Ambassador *Jan 2025 – Present*

- Produce study material and briefings on generative AI on AWS (prompt engineering frameworks, prompt optimisation); completed a multi-service AWS capstone and contribute to community workshops.

EDUCATION

Swinburne University of Technology — Bachelor of Computer Science (AI Major) *Sep 2022 – Oct 2026*

- Pioneer Scholarship (20%). Thesis: generative AI system that batch-assesses translation quality in student reports.

AI Vietnam — AIO 2024 Specialised Programme (machine learning and deep learning) *Jan 2024 – May 2025*

CERTIFICATIONS

- AWS Certified AI Practitioner (Jul 2026) • AWS Certified Cloud Practitioner (Jun 2026)
- AWS Certified Data Engineer – Associate (in progress, exam Dec 2026) • IELTS Academic 7.0 (Aug 2026)

LEADERSHIP

- **ITea Lab**, Head of Media (Jan 2025 – Jan 2026) • **SMACK Vovinam Club**, Vice President & Head of HR (Mar 2023 – Aug 2024)