

Anslem Akadu

Machine Learning Engineer | LLM Fine-Tuning, Inference & Forecasting

Nigeria (UTC+1) | anslemakadu1@gmail.com | +234 705 992 5770
anslem.dev | linkedin.com/in/anslem-akadu | github.com/Anslemakadu

SUMMARY

Machine learning engineer who fine-tunes, serves and deploys models: QLoRA fine-tuning of Gemma 3 1B and migration to Qwen, quantized GGUF inference behind FastAPI, and XGBoost/Prophet forecasting in health tech. Now builds LLM classification, extraction and RAG services on AWS with PII redaction, model selection tuned for cost, and automated tests.

TECHNICAL SKILLS

Machine Learning: PyTorch, Hugging Face Transformers, PEFT (LoRA/QLoRA), instruction tuning, SentenceTransformers, OpenAI embeddings, scikit-learn, XGBoost, Prophet

Forecasting & Data: time series forecasting, feature engineering, baseline/ML blending, model evaluation, pandas, NumPy, SQL, geospatial matching (haversine, Shapely)

LLM Inference: GGUF quantized inference (llama-cpp-python), OpenAI and Anthropic APIs, structured outputs (JSON Schema, Pydantic), tool calling, retrieval augmented generation (RAG), PII redaction

Deployment & MLOps: Python, FastAPI, Docker, AWS (EC2 GPU, Lambda, SQS, DynamoDB, EventBridge, CloudWatch), AWS SAM, Bash, Linux, Git, GitHub Actions, pytest

PROFESSIONAL EXPERIENCE

AI/ML Engineer | 20four7VA | Remote

Oct 2025 - Present

Remote staffing company. Own design, implementation, testing and AWS deployment of the ML and LLM services below.

- **LLM classification:** Built email priority triage (three tiers; rubric prompt plus deterministic validation) and Teams policy classification against 11 policies with strict JSON schema outputs, in event-driven Lambda pipelines fed by Graph webhooks.
- **RAG and extraction:** Built a RAG pipeline that retrieves similar past job descriptions (OpenAI text-embedding-3-small, cosine top 3) to guide new JD generation, after prototyping with local SentenceTransformers; plus multistage transcript extraction with a confidence gate.
- **Privacy and geospatial features:** Redacted 17 PII/secret types with prioritised regex rules before inference; matched GDACS and WMO alerts to ~200 geocoded locations (haversine, tiered 50/100 km; Shapely).
- **Inference cost and reliability:** Tiered model selection (small model for high volume triage, larger for briefings), per user token metering, idempotent DynamoDB writes and dead letter queues; nearly 300 pytest tests.
- **LLM applications:** Built a streaming tool calling Teams assistant (18 tools) and Pydantic validated daily/weekly briefings over Graph mail, calendar and meeting transcripts.

ML/AI Engineer | LifeBank (Nigeria) Ltd | Health tech and medical logistics

May 2025 - Oct 2025

- Fine-tuned Gemma 3 1B with QLoRA (PEFT, Hugging Face) for hospital finance instruction alignment and domain adaptation.
- Migrated the assistant from Gemma to Qwen, cutting inference latency by 20%; automated AWS EC2 GPU training setup with Bash, reducing setup time by 40%.

Data Scientist, Financial Systems | LifeBank (Nigeria) Ltd

Jan 2025 - Apr 2025

- Forecast hospital profitability, liquidity, leverage and efficiency with XGBoost and Prophet, using hospital size and state as regressors for revenue and cash flow models.
- Designed a hybrid forecast (80% naive baseline, 20% ML) for sparse data and surfaced early warnings of negative cash flow.

Data Scientist, Healthcare Forecasting | LifeBank (Nigeria) Ltd

Oct 2023 - Dec 2024

- Built time series forecasting pipelines for blood, oxygen and medical stock demand, reducing stockout risk by 20%; developed predictive maintenance models and a scenario simulation engine for supply planning.

SELECTED PROJECTS

LLM FastAPI Boilerplate | Python, FastAPI, llama-cpp-python, Docker | github.com/Anslemakadu/llm-fastapi-boilerplate

- Served quantized GGUF models with llama-cpp-python and FastAPI (/generate, /chat, /health) with configurable sampling; Dockerized on Hugging Face Spaces after iterating across Phi, Mistral 7B, Qwen2.5 and TinyLlama for CPU only hosting.

ElevatRAI | Python, Flask, scikit-learn, Render | github.com/Anslemakadu/ElevatRAI | elevatrai.onrender.com

- Role matcher that parses PDF/DOCX resumes, maps skills to a JSON taxonomy (7 roles, 56 skills), ranks role fit by cosine similarity and builds a learning roadmap; deployed on Render.

EDUCATION

B.Sc. Computer Science | Institut Supérieur Saint Paul Tarse University, Cotonou, Benin Republic

2021