

IDARA SAMUEL OSU

Machine Learning Engineer

LLMs/Speech/ML Algorithms · Inference Engineering · MLOps · Agentic Pipelines

idaraosusamuel@gmail.com · linkedin.com/in/idaraosu · github.com/aidyai · Nigeria · Remote · 3+ Years

ML Engineer with 3+ years building and shipping LLM-powered products in production. Built the first bilingual English-Ibibio AI platform for a language with 10M+ speakers and zero prior digital resources — corpus, training, and deployment from scratch. Cut p99 latency from 1,800ms to 320ms via TRT-LLM; 40% GPU cost reduction. Fine-tuned and adapted NLLB, Whisper, and Orpheus TTS on self-collected data: TTS MOS 3.14/5, ASR WER 15.33% / CER 5.73%, NMT +28 BLEU

CORE SKILLS

LLM & AI Fine-tuning · Instruction Tuning · NMT · TTS · ASR · Structured Output · JSON-constrained Decoding Agentic Systems LangGraph · LangChain · Compiler-Composer Pipeline · Schema-constrained Routing · Tool Calling · Multi-agent Orchestration

Inference TensorRT-LLM · CTranslate2 · vLLM · KV Caching · In-flight Batching · FP8 / INT8 / W4A16 Quantisation Data & **Infra** AWS SageMaker Pipelines · Terraform · ECR · IAM/OIDC · MongoDB · Supabase · Large-scale Data Collection & Annotation

MLOps Modal Labs · Docker · GitHub Actions CI/CD · AWS SageMaker · S3

EXPERIENCE

Head of Machine Learning · USEM AI ·

Jan 2024 – Jun 2026 · Nigeria

Sole ML lead for a bilingual English-Ibibio AI platform. Built the first digital corpus for Ibibio: 80 hrs of bilingual speech + 205,000 parallel sentence pairs. Every layer from scratch — data collection through production inference.

- Pioneered the first bilingual AI dataset for Ibibio: supervised collection and curation of 80 hours of English-Ibibio bilingual speech data and 205,000 parallel English-Ibibio sentence pairs — the first digital linguistic corpus for this language.
- Fine-tuned production bilingual TTS and ASR models on NVIDIA H100 GPUs using the bilingual speech dataset. Whisper adapted via LoRA fine-tuning; Orpheus TTS via Parameter-Efficient Fine-Tuning (LoRA). Achieved MOS 3.14/5 for TTS naturalness and ASR WER 15.33% / CER 5.73%.
- Adapted Meta's NLLB for English-Ibibio NMT via transfer from Igbo, a closely related Niger-Congo language, using 205,000 parallel sentence pairs; achieved +28 BLEU over the zero-shot baseline.
- Converted TTS/ASR to TensorRT-LLM and NMT to CTranslate2 with KV caching, in-flight micro-batching, and quantisation. p99 latency: 1,800ms to 320ms (82%), GPU cost down 40%.
- Architected a LangGraph Compiler-Composer-NMT pipeline: schema-constrained intent routing across 4 ops, NMT-optimised composition, full Ibibio output. 250+ active users across translation, transcription, and audio generation APIs. - Built supporting production systems: agentic Q&A; generator, lifecycle email cron, Supabase with retry/dead-letter logic, and Orpheus TTS in FP8 on H100 with streaming. Deployed on Modal Labs serverless with full CI/CD

ML Engineer (Contract) · Demz Analytics / Storri AI ·

Jul 2024 - Nov 2024

- Built the LLM orchestration pipeline converting raw news articles into structured content — headlines, digests, discussion prompts, image/video generation prompts — powering Storri AI's core product.
- Added output validation and JSON-repair logic across ~30 chained GPT functions, enabling unattended production operation without manual editorial correction.
- Designed the MongoDB data layer storing and serving all AI-generated content for fast retrieval across the video-generation workflow.

PRODUCT FREELANCE

EchoScript · Automated SageMaker + TRT-LLM Transcription Platform

Jul 2026

- 3-step SageMaker Pipeline (Download, Transcribe, GenerateDocs) with dependency-chained S3 I/O and containerised ScriptProcessors. All infrastructure — S3, ECR, IAM, CloudWatch, GitHub OIDC — provisioned as code with Terraform.
- TRT-LLM Whisper engine with paged KV cache; threaded 20ms batcher collects mel tensors; drain thread resolves futures by request_id. 6-stage GitHub Actions CI/CD: ruff, pytest, Terraform apply (OIDC-federated), Docker push (3 ECR images), model upload, pipeline upsert.

EDUCATION

Sep 2017 – Jul 2023

BSc. Urban & Regional Planning · University of Uyo Nigeria · CGPA: 4.35 / 5.0

- Spatial analysis, housing policy, environmental planning, infrastructure development, and Nigerian land law. Projects: site analysis, master plan preparation, and GIS-based mapping of urban growth patterns across Nigerian cities