FORTUNATUS OLUWADARA ADEGOKE

AI Engineer

+234 903 8901 471 fortunatusoluwadara@gmail.com linkedin.com/in/fortunatus-adegoke github.com/coding-fortunatus
Abeokuta, Nigeria

SUMMARY

AI Engineer with hands-on production experience building computer vision pipelines, RAG-based LLM agents, and audio intelligence systems. Deployed a real-time AI proctoring service detecting **6 violation types** via WebRTC stream analysis, and architected a RAG lesson-note assistant reducing repeated LLM inference by **~60%** through metadata-filtered vector caching. Proficient across the full AI stack — YOLO, MediaPipe, Silero VAD, LlamaIndex, Qdrant, OpenAI API, prompt engineering, and event-driven orchestration with Inngest. Backend-strong: production APIs in Python (FastAPI, Django), PHP, and Node.js provide the engineering foundation for scalable, observable AI services.

AI & TECHNICAL SKILLS

Computer Vision	YOLOv8 (object detection — person, phone), MediaPipe (face detection, Face Mesh, gaze estimation), OpenCV
Audio / Speech	Silero VAD (voice activity detection), TorchAudio, Librosa, adaptive RMS noise calibration
LLM & RAG	OpenAI API (GPT-4.1, GPT-4 Turbo, text-embedding-3-large), LlamaIndex, prompt engineering, structured JSON output, context-window management
Vector Databases	Qdrant Cloud (payload indexing, COSINE similarity, metadata-filtered search, collection management)
AI Orchestration	Inngest (durable functions, event-driven workflows, step-based inference, retry/observability)
ML Runtimes	PyTorch 2.x (CPU inference), torch.hub, Uvicorn/ASGI, singleton model loading
AI Frameworks	FastAPI, Django, LlamaIndex, SentenceSplitter chunking
Real-Time AI	WebRTC (signaling + server-side stream analysis), Socket.IO, Python asyncio, Semaphore concurrency control
Infra	Docker (CPU inference containers), Railway, Qdrant Cloud, PM2, Systemd

EXPERIENCE

AI Engineer

2025 – Present | Hybrid

Bluespectra Limited, Abeokuta

- Built production TRCN AI Proctoring Service — real-time WebRTC stream analysis detecting **6 violation types** using YOLOv8n, MediaPipe Face Detection/Face Mesh, and Silero VAD; containerized with Docker for CPU-only deployment at **~2–4GB memory footprint**.
- Implemented adaptive per-candidate RMS audio baseline calibration (first 10 chunks, 30s silence reset) — reducing false positives in varied exam environments compared to fixed-threshold approaches.
- Designed behavioral sequence detection state machine (counters per session, threshold-based flagging) and thread-safe in-memory session store with **10-minute TTL** and automated cleanup; singleton model loading eliminates **2–5s cold-start penalty**.
- Architected TRCN AI Lesson Note Agent — RAG pipeline with LlamaIndex SentenceSplitter (**700-token chunks, 100-token overlap**), Qdrant payload-indexed vector store, and GPT-4.1 generation; Inngest durable orchestration decouples request receipt from LLM inference for reliable async delivery.
- Reduced repeated LLM inference by **~60%** through metadata-filtered vector caching on 4 educational dimensions (subject, class, term, week); Semaphore(4) limits concurrent agent invocations to protect API rate limits.

Software / Backend Engineer (AI Integration)

2023 – 2025 | Hybrid / Contract

Zealarax Technologies / Bluespectra Limited

- Integrated OpenAI GPT-4 Turbo into the OAU Amaranthus Virtual Assistant — multilingual Q&A system (5 languages) with DB-backed answer cache reducing API response latency by **~60–70%** and lowering repeat inference calls by

~50–65% over deployment lifetime.

- Engineered keyword-based domain guard (**~190 amaranth-specific terms across 5 languages**) as pre-filter before GPT calls — protecting API quota from off-topic queries; used GPT JSON mode with structured multilingual output (10-key response object).
- Implemented admin-controlled batch GPT processing with rate-limit-aware sequential execution and feature toggle for AI on/off control.

AI PROJECTS

TRCN AI Proctor

FastAPI · YOLOv8 · MediaPipe · Silero VAD · PyTorch · Docker · WebRTC

- Multi-model pipeline: face detection, 468-landmark gaze heuristic, YOLO person/phone detection, VAD speech segmentation, adaptive noise baseline — all per-frame/per-chunk, CPU-only.
- Gaze direction via mean X-coordinate of 468 Face Mesh landmarks — simple, effective, no heavy gaze estimation model needed.

TRCN AI Lesson Note Agent (WIP)

FastAPI · LlamaIndex · Qdrant · OpenAI GPT-4.1 · Inngest · Socket.IO

- Event-driven: Socket.IO → Inngest → agent → Qdrant search → GPT inference → Socket.IO push; supports PDF/DOCX/HTML/URL ingestion with boilerplate filtering.

OAU Amaranthus Virtual Assistant

Node.js · Express.js · OpenAI GPT-4 Turbo · MySQL · Sequelize

- DB-first cache strategy: exact-match lookup before GPT call; column-per-language schema (EN/YO/IG/HA/PCM); batch processing fills all null language columns in one GPT call.
- **~50–65% reduction in repeat API inference** over deployment lifetime through aggressive caching.

EMS-IBR Seat Allocation Engine (Co-owned)

Django · Celery · Python

- Multi-pass constraint satisfaction: checkerboard/diagonal pattern pre-placement + **500-attempt random search** with 8-directional adjacency validation; graceful unplaced fallback over forced constraint violation.

BACKEND FOUNDATION

Production API experience across PHP (Laravel, OpenSwoole), Python (FastAPI, Django), and JavaScript (Node.js/Express): fintech ledgers, hospital management systems, real-time WebSocket servers, supply chain APIs, and government property platforms — providing the engineering backbone for scalable, observable AI service deployment.

EDUCATION & CERTIFICATIONS

HND, Computer Science — Federal Polytechnic Ilaro 2021–2024

Tunga DSA (2022) · Introduction to Backend Engineering, META/Coursera (2024) · Software Engineering 9-in-1, META/Coursera (In Progress)