


ETIMBUK AFIA

AI Engineer

✉ afiaetimbuk100@gmail.com  github.com/etimbukafia  linkedin.com/in/etimbukafia

PROFILE

AI Engineer with 3+ years helping clients and businesses achieve goals and targets through AI, including increasing revenue, improving conversion workflows, reducing operational costs, and automating high-value processes with measurable ROI. Expert in building models, integrating LLMs, vector stores, APIs, tool-calling, and cloud deployments to deliver tailored AI solutions across multiple domains.

Skilled in agentic pipeline design, memory-based reasoning, and building high-reliability, production systems deployed on the cloud

PROFESSIONAL EXPERIENCE

AI Engineer, Seedgate Technology

2022 – Present

Architected multi-agent, multimodal, and high-scale automation systems across education, ASR, RAG, architectural intelligence, document intelligence, YouTube content automation, and marketing automation. Integrated LLMs with third-party APIs, CRMs, dashboards, and microservice ecosystems to deliver autonomous planning, structuring, reasoning, and execution capabilities.

Some Projects:

AI-Powered Long-Form Content Generation System (Language Learning Platform):

- Built an automated multi-step LLM workflows generating large-scale long-form content (book summaries, movie-based narratives, instructional materials).
- Designed prompt sequences for extraction, restructuring, synthesis, style enforcement, and metadata generation.
- Implemented quality checks for coherence, hallucination reduction, tone control, and narrative stability using structured tool-calling.
- Automated pipelines using Python, long-context APIs, and agentic workflow tools.

Architectural Drawing Intelligence (Multimodal Vision + RAG System):

- Built multimodal agents combining Vision LLMs, RAG pipelines, and vector search to interpret complex architectural drawings.
- Enhanced retrieval with hybrid chunking, re-ranking, multimodal embeddings, and structure-aware metadata.
- Delivered interactive Q&A reasoning for architectural plans using deterministic, schema-backed agent orchestration.

Real-Time Streaming ASR Optimization (Quran Recitation App):

- Re-architected a batch-trained ASR model for 200–300ms streaming inference with minimal hallucinations.
- Tuned NVIDIA Riva pipeline (beam width, VAD, frame size, timestamps, latency budgets) achieving sub-500ms end-to-end latency.
- Implemented right-context expansion, chunk-overlap, CTC smoothing, and chunk-level beam constraints.
- Built benchmark suite covering WER, hallucination rate, temporal continuity, and speaker robustness.

Enterprise Document Intelligence — Production RAG System Audit & Refactor:

- Audited a microservices-based Weaviate + LangChain RAG architecture used for large PDF/DOCX/XLSX pipelines.
- Redesigned chunking strategy, embedding flow, retrieval logic, hybrid search, reranking, OCR pipeline, and metadata handling.
- Improved search relevance, indexing quality, and prompt engineering for structured document analysis.
- Delivered evaluation frameworks, documentation, and monitoring dashboards.

LangChain Orchestration Layer for Institutions & Blockchain Verification:

- Built advanced LangChain pipelines with structured-output agents, tool routing, vector store integrations, and deterministic reasoning.
- Connected users, institutions, and blockchain verification systems through a unified orchestration layer.
- Ensured reliability through schema validation, agent state management, and optimized tool-calling flows.

YouTube Content Intelligence Agent (Research Analysis Scriptwriting):

- Built multi-agent system for topic research, trend analysis, competitor mapping, tone modeling, and script generation.
- Integrated YouTube APIs to ingest transcripts, analytics, metadata, and retention data.
- Implemented narrative structures optimized for CTR, AVD, watch time, and engagement patterns.

AuditAI — Autonomous Architecture, Code & Pipeline Auditor

- Designed multi-role agent system (Architect, Reviewer, Validator) for auditing codebases, ML pipelines, workflows, API integrations, and cloud architecture with CrewAI.
- Delivered structured reports, risk analyses, and optimization recommendations using memory-based reasoning.

EDUCATION

BEng, Computer Engineering, *Elizade University*

2018 – 2023

Final Year Project: Real-time Yorùbá sign language recognition and transcription using a custom CNN architecture (87% accuracy).

SKILLS

AI / ML / LLM Systems — LLMs | Multi-Agent Architectures | Tool Calling | Autonomous Planning | Reasoning Loops | LangChain | LangGraph | AutoGen | CrewAI | Google ADK | RAG | Multimodal LLMs | Embeddings | Vector Databases (Pinecone, Qdrant, Milvus, Weaviate) | Transformers | Conformer Models | CTC | Streaming ASR | Vision Models

Backend & API Engineering — Python | FastAPI | Node.js | TypeScript | PostgreSQL | MongoDB | Neo4j | Supabase | Firebase | REST/GraphQL APIs | API Orchestration | Microservices | ETL Pipelines | Redis

Infrastructure — AWS | Azure | GCP | Docker | Kubernetes

Workflow & Automation — CRM Integrations | Analytics APIs | Google Sheets Automation | Content Pipelines | Document Parsing | OCR | Data Enrichment | Workflow Orchestration | Agent State Management

Monitoring & Reliability — Prometheus | Grafana | Sentry | InfluxDB | ELK Stack | Structured Logging | Latency Optimization | Performance Dashboards

PROJECTS

AI Performance Optimizer Agent

Autonomous agent monitoring 12+ ML pipelines, handling drift detection, retraining, and anomaly recovery. Boosted uptime to 99.7% and cut recovery from 5 to 8 minutes.

AI CRE Intelligence Agent

Multi-agent system analyzing CRE data (docs, listings, comps, images) with RAG-powered insights for financial modeling, risk scoring, and deal analysis.

Cross-Platform Automation Layer

LLM-driven engine connecting CRMs, Sheets, dashboards, and APIs to automate routing, enrichment, reporting, scheduling, and content generation.