UDOCHUKWU ECHEFU

AI ENGINEER

Lagos, Nigeria · +234 901 796 1739 · udochukwuechefu10@gmail.com · udochukwu.cv · GitHub

PROFESSIONAL SUMMARY

AI engineer building evidence-led, human-in-the-loop applications with Python, FastAPI, LLMs, retrieval, and modern web interfaces. Shipped live systems for support triage, contract intelligence, onboarding review, and auditable decision workflows. I pay close attention to LLM context: what the model receives, what it can retrieve, and what it is allowed to return. I use evaluation harnesses to check whether outputs stay grounded, how the system behaves in failure cases, and whether responses remain consistent. I focus on making model output explainable, testable, and useful in real operational workflows.

TECHNICAL SKILLS

AI Engineering: LLM application development, RAG, hybrid search (BM25 + vector), structured outputs, prompt design, embeddings, evaluation, deterministic guardrails, human-in-the-loop workflows

Backend & Data: Python, FastAPI, SQLAlchemy, Alembic, PostgreSQL, pgvector, Redis, SQLite, REST APIs, background workers, load testing (k6), audit logs

Frontend & Cloud: React, Next.js, TypeScript, Streamlit, Vite, Tailwind CSS, Cloudflare Workers/R2, Railway

PROFESSIONAL EXPERIENCE

CLINIMU Healthcare | AI Engineer — Mar 2026 - Present

- Designing and integrating AI-assisted features into the CLINIMU healthcare system, including a retrieval-augmented generation workflow connected to existing product flows.
- Shaping the LLM context, retrieval and output boundaries so responses are grounded in approved clinical and operational sources, with supporting evidence available for review.
- I handle data preparation, prompt and test-case writing, interface wiring and documentation while keeping implementation aligned with product requirements.

The Gift House / Zoba Events | Brand Designer — May 2025 - December 2025

- Developed cohesive brand visuals and marketing assets for gifting and event experiences, maintaining a consistent visual language across customer touchpoints.

Upthrust | Junior Product Designer — Jun 2023 - Jan 2025

- Created interface concepts and reusable visual patterns for a digital product, supporting the team with wireframes, layout decisions, and polished UI assets.

Health Emergency Initiative (HEI) | Volunteer Graphic Designer — Aug-Oct 2022

- Produced social-media and campaign graphics for the media team while maintaining visual consistency across public-facing communications.

SELECTED AI ENGINEERING PROJECTS

Runna Gateway | AI Engineer — Sep-Oct 2026

Write-up · Live demo · Source

- Built an OpenAI-compatible, multi-tenant model gateway in FastAPI: streaming chat and embeddings across six providers, per-team model policies, and per-request usage and cost receipts in PostgreSQL.
- Enforced rate, concurrency and budget limits across replicas with atomic Redis Lua scripts. A load test exposed boundary over-admission; I replaced the limiter with GCRA, which held a 600 requests/minute limit with a burst of 30 to a 629-request maximum across three runs.
- Measured gateway overhead with k6 against a fixed-latency test provider: p99 of 7.2 ms at 100 requests/s on one replica and 5.6 ms at 200 requests/s on two. A ten-minute soak reconciled 30,001 responses with 30,001 receipts.
- Stack: Python, FastAPI, PostgreSQL, Redis, Next.js, TypeScript.

Hybrid Retrieval | AI Engineer Sep-Oct 2026

Source

- Built a FastAPI retrieval service on PostgreSQL that fuses ParadeDB BM25 and pgvector results with Reciprocal Rank Fusion, with versioned ingestion of TXT, Markdown and PDF sources and page-level provenance.
- Evaluated on BEIR SciFact (5,183 documents, 10,980 chunks): hybrid Recall@10 of 84.44% against 79.07% BM25-only and 81.97% vector-only, and nDCG@10 of 70.75% against 67.73% and 65.89%.
- Stack: Python, FastAPI, PostgreSQL, pgvector, ParadeDB.

WHT Recovery Control | AI Engineer Aug 2026

Live workspace · Source

- Built an AI-assisted control workspace for Nigerian withholding-tax receivable recovery, using DeepSeek to extract receipt fields through a strict schema with confidence scores, exact source quotes, and page provenance.
- Implemented deterministic reconciliation of document evidence to ledger cases, with human review for recognition, disputes, and closure; persisted workflow records, Cloudflare D1/R2 storage, and audit events.
- Stack: TypeScript, DeepSeek, Cloudflare Workers, D1 and R2.

Lenslayer | AI Engineer Aug 2026

Live workspace · Source

- Built an evidence-led contract operations and onboarding workspace combining a Next.js product interface with FastAPI services, persistent workflow records, private document storage, and human-owned decisions.
- Implemented portfolio-wide grounded Q&A, source-linked findings, identity-evidence reconciliation, assignments, approvals, deadlines, audit history, and reporting; AI findings never become material actions without reviewer confirmation.
- Stack: Python, FastAPI, Groq-hosted LLMs, LangChain, ChromaDB, Hugging Face embeddings, Next.js, React, TypeScript, SQLAlchemy, PostgreSQL, Cloudflare Workers and R2.

Kora Support Triage | AI Engineer Jul-Aug 2026

Live demo · Source & evidence

- Built a React and FastAPI support-operations system that uses Groq structured outputs for intent, urgency, sentiment, entity extraction, and response drafting across English and Nigerian Pidgin cases.
- Added deterministic policy guardrails, tenant-scoped knowledge citations, human approval, SQLite customer memory, durable jobs, manual fallback, proof mode, and persisted model/human audit trails.
- Verified with 132 backend tests, a 100-case synthetic policy dataset, and an 11-case live Groq smoke run that passed 11/11; deployed on Railway.

ContractGuard | AI Engineer Jun-Jul 2026

Live app · Source

- Built a Streamlit document-intelligence workspace for PDF, DOCX, TXT, and scanned-document review, with extraction diagnostics, evidence-linked risks, obligations, negotiation questions, and grounded follow-up answers.
- Implemented retrieval with ChromaDB and all-MiniLM-L6-v2 embeddings, Groq through LangChain, citation-backed outputs, conversational memory, report exports, and privacy-aware retention controls.

EDUCATION

B.Sc. Computer Science | Federal University of Technology, Owerri 2019-2024

ADDITIONAL INFORMATION

Language: English