

Moaz Muhammad

AI Engineer — LLM Applications, RAG & Document AI

moazmo27@gmail.com | +20 107 047 4845 | Cairo, Egypt | Open to remote
linkedin.com/in/moazmuhammad | github.com/moazmo | portfolio-moazmos-projects.vercel.app

Summary

AI engineer who builds LLM products that cite their sources or refuse to answer. Sole or lead engineer on live Arabic/English products (tender compliance, online academies), a clinical-documentation assistant and an open-source code-review agent, in Python and TypeScript on PostgreSQL, with tests and published evaluations, unflattering numbers included.

Experience

Yashfii | *AI & Software Engineer (part-time)* Cairo, Egypt / Hybrid | Jun 2026 - Present

- Sole engineer on AIDocFriend V2, the successor to the team-built V1 clinical assistant: a FastAPI/Python and Next.js/TypeScript monorepo that turns Egyptian-Arabic consultations into live differential diagnoses, editable SOAP notes, prescriptions and lab orders. The clinic owner approved V2 as the platform going forward.
- Built a speech-to-text adapter over four engines (browser speech, Deepgram real-time streaming, Mistral Voxtral, local Qwen-based Egyptian-Arabic ASR) with automatic fallback, feeding staged live analysis.
- After the LLM medication-safety pass, added a **deterministic filter that drops any finding for a drug not on the reconciled medication list**. Also built per-doctor isolation, a leased PostgreSQL job queue, versioned clinical edits and per-encounter AI cost budgets; 14 ADRs.

NTG Clarity | *AI Product Engineer Intern — Team Lead* Cairo, Egypt / Hybrid | Aug 2026 - Sep 2026

- Led a three-person team and wrote 196 of 197 commits for **BidProof** (live), which reads Arabic and English tenders and marks a requirement "satisfied" only when approved, cited, in-date evidence proves it.
- Raised requirement-extraction recall on 578 attorney-labelled CUAD clauses from **0.673 to 0.777**, and parser verbatim accuracy to **0.975 (Arabic) and 0.979 (English)**. An audit of **1,528 production requirements** found 1,309 quoted verbatim on the cited page and 7 not found.
- Pipeline: PDF/DOCX + OCR, extraction, pgvector + lexical retrieval, verification, evidence lock; Fastify, React, Supabase PostgreSQL, Cloudflare R2, QStash; 306 automated tests.

Product Manager Accelerator (PMA) | *AI Engineer Intern* Remote | Apr 2026 - Jun 2026

- Owned the AI-engineering workstream on ComplyAI, a Claude-based AI-regulation compliance checker: raised rule-data QA from **70.3% to 99.8%** (550/551 rules) across 11 rule sets and re-mapped 88 US rules mis-tagged with EU IDs.
- Validated the assessment engine end to end against live Claude: **13/13** legal-reasoning checks and **zero hallucinated rule IDs** across 67 generated gaps.

Lillup | *AI Engineer Intern (part-time)* Remote | Jan 2026 - Jun 2026

- Built parts of an on-device multi-agent coaching assistant: hybrid intent routing (LLM + rules), RAG with ChromaDB, BM25 and citations, output-safety gates, pytest suites.

Cellula Technologies | *Machine Learning Engineer Intern* Remote | Jul 2025 - Aug 2025

- Built taxi-fare regression and hotel-cancellation classification models (scikit-learn, XGBoost, Pandas).

Projects

Eilm *Founder & sole engineer — TypeScript, Next.js, Supabase/PostgreSQL, Playwright* Jul 2026 - Present

- Built and run a live, Arabic-first, multi-tenant learning platform alone (749 commits): **155 migrations**, deny-by-default row-level security, **about 840 automated tests plus 91 database test files**, and a release gate on lint, types, translation parity and build. Five real academies use it alongside three demo showrooms.
- Each academy is merchant of record on its own Stripe/Paymob account; the platform never holds funds. Lesson Q&A answers only from the transcript, with verbatim quotes, or refuses.

PR Sentinel *GitHub Marketplace Action — Python, LangGraph, GitHub Actions* Jun 2026

- Multi-agent code review (4 analysts x 3 self-consistency samples, evidence anchoring, verifier) in the user's CI with their own API key, ~\$0.01/review; 256 tests.
- On **60 real merged bug-fix PRs** it reached **28% recall and 71% precision**, against 91% on seeded fixtures. I published the real-PR result and shipped the accuracy levers that failed to beat baseline as off by default.

SecureAgentRAG *Python, LangGraph, FastAPI, Qdrant, Next.js* May 2026 - Jun 2026

- Privacy-first English/Arabic agentic RAG: a nine-node LangGraph pipeline, hybrid dense and sparse retrieval fused by reciprocal rank, cross-encoder reranking, role-based access enforced inside the vector filter, and enforced citations; about 700 tests and 41 ADRs.
- Fine-tuned BGE-Reranker-v2-M3 (**NDCG@10 0.774 to 0.790**); shipped opt-in, since a small gain didn't justify a ~2 GB dependency for every user.

Skills

AI / LLM: LLM application development, RAG (hybrid retrieval, reranking, citation verification), LLM evaluation, AI agents (LangGraph), document AI & OCR, Arabic NLP, prompt-injection hardening, speech-to-text; DeepSeek, Claude, OpenAI-compatible APIs, Ollama, Hugging Face

Engineering: Python, FastAPI, TypeScript, Fastify, Next.js, React, PostgreSQL, Supabase (RLS), pgvector, Qdrant, Drizzle, REST, WebSockets/SSE, Arabic/English i18n & RTL

Quality & delivery: pytest, Vitest, Playwright, pgTAP, GitHub Actions, Docker, Vercel, Hugging Face Spaces, Cloudflare, ADRs, Claude Code | **Languages:** Arabic (native), English (professional)

Education

Helwan University, Cairo | BSc Computer Science & Artificial Intelligence

Sep 2024 - Jul 2028 (expected)