

Hossam El-Kharbotly

AI Engineer

📞 +962 7 7521 6254 (Call/WhatsApp) 📩 +20 109 105 5947 (WhatsApp only)

✉️ hossam.kharbotly@gmail.com | linkedin.com/in/hossam87 | therock87.github.io

Cairo, Egypt (currently on assignment in Amman, Jordan) | Open to relocation & remote

Profile

AI Engineer with 3+ years taking Generative AI from prototype to production: multi-agent systems that take real actions, hybrid RAG over enterprise data, and the sovereign on-prem GPU stack that serves them to hundreds of concurrent users. Hands-on Arabic NLP.

Core skills: Multi-agent systems, on-premise LLM serving, RAG/GraphRAG, GPU concurrency, AI observability.

Work Experience

AI Engineer & Automation Specialist

Feb 2026 - Present

Dar Al-Handasah

Cairo, Egypt · on assignment in Amman, Jordan

- Embedded on-site with Dar's PMC unit in Amman: ran discovery with the PMC partner, the unit director, his directors, and the cost-engineering team to capture their use cases and business case, then demoed, tested, and onboarded the agents face to face, training ~10 cost engineers.
- Build production multi-agent systems on AIDA, Dar's firm-wide AI platform, serving **hundreds of users** across **3 departments** with **thousands of requests per month**.
- Built a human-in-the-loop layer that lets agents pause to ask the user for missing context, now a shared AIDA capability other teams build on.
- Designed the model-agnostic on-premise LLM stack, presented the architecture to the CIO, and operate it, selecting models through golden-set benchmarks against Azure OpenAI.
- Instrumented agents with Langfuse tracing and review code for AIDA contributors.
- Mentored ~30 trainees across **3 internship cohorts** in Dar's Egypt, Lebanon, and Jordan offices.

AI Engineer

Aug 2025 - Jan 2026

Edulga

- Designed and shipped an adaptive learning platform that turns unstructured content (documents, videos, web) into personalized learning roadmaps using knowledge graphs and LLM orchestration (**OpenAI API**), serving **12+ REST endpoints** at **<100 ms P95**.
- Built Neo4j-based knowledge-graph pipelines and LangChain integrations for roadmap generation and retrieval.
- Reduced end-to-end latency and **OpenAI API** spend via async parallelization, HTTP/2 pooling, prompt caching, and batched operations.
- Implemented multi-tenant data isolation across users and sources for secure, auditable operation, backed by **97+ automated tests** and containerized services on AWS.

Freelance AI Engineer

Mar 2025 - Aug 2025

Independent

- Delivered custom AI solutions to **3 business clients** in Egypt, owning each engagement end to end: discovery, requirements, live demos, deployment, and a paid-for product.

R&D Machine Learning Engineer

Mar 2024 - Mar 2025

Al Farid Scan

- Optimized MRI classification models for diagnostic accuracy, and engineered attention mechanisms into medical image segmentation pipelines to improve precision and interpretability.

Projects

 [See all projects](#)

BoQ Cost Agent (v1 → v2) | *Agent harness, LangGraph, Text-to-SQL, Embeddings, SQL Server*

 [Project Page](#)

- Architected an agentic cost-intelligence system on an **agent harness of 6 tools** (text-to-SQL over a **1.2M-row** database, grep-style search, embedding retrieval), cutting PMC cost engineers' lookups from **up to half a day** to **minutes**.

- Evolved the v1 LangGraph pipeline into autonomous tool planning, so one plain-language query returns a benchmark, cross-project comparison, and fair-price estimate instead of manual data pulls.
- Engineered a fallback ladder (**nearby country, adjacent year, converted unit**, USD normalization) that keeps estimates flowing without an exact match, citing every BoQ, project, and year for sign-off.
- Built the v1 foundation: typed LangGraph nodes, a self-repairing fix-agent, a 6-stage term-resolution cascade (fuzzy, LLM, Weaviate hybrid), and human-in-the-loop interrupts that stop SQL being built on an unclear question.

BoQ Ingestion Agent | *Pi agent harness, Agent tools, VLM/OCR, SQL Server*

[Project Page](#)

- Front line of Dar's historic bill-of-quantities database: a Pi agent harness that absorbs any BoQ (Excel, digital or scanned PDF) regardless of row or worksheet count.
- Investigates each BoQ with its own tools (including screenshots), summarizes it, and proposes a column mapping that a cost engineer approves in a review dashboard before ingestion into the Cost Agent's SQL database.
- Surfaces reconciliation checks, stats, benchmarks, and charts so the cost engineer can confirm each BoQ is correct and complete; scaling toward **50+ historical projects** and **millions of line items**.

Enterprise Document Parsing Platform | *VLM OCR, Classical OCR, MinIO, FastAPI*

- Shipped Dar's **first document-digitization platform**, the parsing core behind all of its RAG systems, turning a **3,500-page document into searchable data in ~25 min** at **~100 files/day** and returning **Markdown and typed JSON** through APIs.
- Routes each region to VLM or classical OCR, absorbing PDFs of up to **3,500 pages** for **6 concurrent users** with scalable concurrency, self-healing services, and **Prometheus, Grafana, and Loki** monitoring.

Enterprise AI Infrastructure | *vLLM, TEI, LiteLLM, Prometheus, Grafana, Loki*

- Own Dar's firm-wide hybrid (Azure + on-premise) AI infrastructure across **6+ servers**, including **4 GPU servers** spanning A100 and RTX 4090 hardware.
- Deploy and operate open-weight LLMs (**Qwen-family** today, model-agnostic by design) plus embedding, reranking, and specialty models on **vLLM and TEI**, keeping all confidential data in-house.
- Tune concurrency, context windows, and GPU memory to co-locate multiple models per GPU, serving **hundreds of concurrent users** and running **months without OOM**, with auto-heal and fleet-wide monitoring.
- Integrated **LiteLLM as the enterprise AI gateway**: model routing, **personalized API keys**, and **per-user budgets**.

El-Mal El-Halal Arabic Podcast Semantic Search | *Whisper, SentenceTransformers, Arabic*

[Project Page](#) | [Dataset](#)

- Arabic semantic search over ~20 hours of podcast audio at **50 ms** median CPU latency; released the subtitles as a Hugging Face dataset.

Education

Nile University

Sep 2019 - Sep 2024

Bachelor of Engineering and Applied Science in Electronics and Computer Engineering

- GPA 3.03/4.00 (Very Good) | Graduation Project: Excellent (A) | **six** posters at the Undergraduate Research Forum (UGRF), **first place** for two.

Technical Skills

Programming Languages: Python, TypeScript, C++, SQL, Cypher

Web & APIs: FastAPI, React, Node.js, REST

AI & ML Frameworks: PyTorch, TensorFlow, Scikit-Learn, LangChain/LangGraph, deepagents, DSPy, HuggingFace Transformers

Agentic & NLP: Multi-agent orchestration, Human-in-the-loop pipelines, Text-to-SQL, RAG, GraphRAG, Fine-tuning LLMs (LoRA/PEFT), On-premise LLM serving (vLLM, TEI), VLM & classical OCR, Harness engineering, Graph engineering, Golden-set evaluation

Arabic AI & NLP: Arabic NLP, Arabic ASR/TTS, Arabic NER, Arabic semantic search, Dialectal Arabic

MLOps, Infra & Observability: Docker, CI/CD, AWS (EC2, S3), Azure, Azure DevOps, Azure AI Foundry, GPU fleet management, Langfuse, MLflow, LiteLLM (model routing), Prometheus, Grafana, Loki, MinIO

Data & Vector Stores: MongoDB, Neo4j, SQL Server, PostgreSQL, Weaviate, Qdrant, FAISS

Languages

Arabic (native), English (fluent, IELTS Academic Band 7, C1)