

MOHAMED AWED OGBA

Senior AI Engineer — Agentic AI · RAG · LLMOps

Riyadh, Saudi Arabia | +966 50 913 4503 | ogba.awed@gmail.com | [linkedin.com/in/mohamed-awad-079485115](https://www.linkedin.com/in/mohamed-awad-079485115)

PROFESSIONAL SUMMARY

Senior AI Engineer with 9+ years in software and enterprise delivery, now building production Generative AI for national-scale platforms. Core AI engineer at ELM Company on the **Nusuk Cognitive AI Assistant** for Saudi Arabia's Hajj & Umrah platform, designed for **20,000+ concurrent users**: agentic workflows, multilingual Arabic/English RAG, hybrid LLM routing and speech AI. Owns the LLMOps lifecycle on GCP and Kubernetes, from fine-tuning and low-latency serving to evaluation, guardrails and observability. Previously 6+ years leading engineering teams on ERP programs for government clients, including the Ministry of Defense and the Royal Court.

PROFESSIONAL EXPERIENCE

AI Engineer · ELM Company

Aug 2025 – Present · Riyadh, Saudi Arabia

Senior-level scope on the Nusuk Cognitive AI Assistant, a multilingual Arabic/English assistant for Saudi Arabia's Hajj & Umrah platform. The work improved digital service accessibility for pilgrims and significantly reduced operational support load.

- **Architected** agentic workflows for Nusuk: agents with function/tool calling and multi-step reasoning (ReAct, Plan-and-Execute), and multi-agent orchestration with shared memory and human-in-the-loop control, acting on downstream APIs across multiple enterprise products.
- **Built** production RAG on Milvus with multilingual Arabic/English embeddings, hybrid semantic search and re-ranking, grounding the agent layer in enterprise knowledge.
- **Delivered** hybrid LLM routing between OpenAI and Arabic-specialized models for high-concurrency public-sector workloads.
- **Fine-tuned** open LLMs (Qwen3, Llama) with LoRA/QLoRA/PEFT and served them on vLLM for low-latency, GPU-optimized inference.
- **Engineered** end-to-end speech AI (streaming ASR, TTS and full speech-to-speech with turn-taking and barge-in) for voice-based pilgrim assistance.
- **Designed** async FastAPI microservices with Redis caching, rate limiting, JWT and RBAC/ACL on GCP-hosted Kubernetes (Terraform/Pulumi, Helm); led AI security, including prompt-injection mitigation and tool-use guardrails aligned with government standards and the OWASP LLM Top 10.
- **Owned** the MLOps/LLMOps lifecycle: CI/CD with canary releases, model registries, LLM evaluation and tracing, Kubeflow orchestration and drift monitoring (Prometheus, Grafana, OpenTelemetry, Evidently AI).

Technical Team Lead — ERP & Enterprise Systems · ELM Company

Jun 2020 – Aug 2025 · Riyadh, Saudi Arabia

- **Led** the technical team delivering enterprise ERP systems for major government and confidential public-sector clients, owning the full SDLC from requirements to deployment, maintenance and stakeholder support.
- **Key engagements:** National Center for Government Resources (NCGR), Ministry of Defense, Royal Court, Riyadh Development Authority (CADC – PIF) and Sharqia Development Authority, plus additional classified digital-transformation programs.
- **Coordinated** multiple stakeholder groups to ensure delivered systems met business and quality requirements.

Technical Team Leader · SLNEE Company

Jun 2019 – Jun 2020 · Riyadh, Saudi Arabia

- **Led** the team maintaining ERP systems for the Royal Saudi Strategic Missiles Force and integrated military systems with external ministry platforms for cross-agency data exchange.
- **Deployed** Odoo at the Bankruptcy Commission (Ministry of Commerce); Finance Consultant and Senior Outsource Developer on ELM Company projects (Marin, ASAS).

Earlier Career — ERP & Software Engineering

Jan 2017 – Apr 2019 · Sudan

- **CTO & Founder, App-Script for Business Solutions** (Nov 2017 – Apr 2019, Khartoum) — founded and led a 9-member Agile/Scrum team of developers, testers and analysts as Product Owner; delivered an Odoo ERP for the University of Khartoum across academic and administrative processes.
- **Senior ERP Techno-Functional Consultant, Amarouse for IT Solutions** (Jan 2017 – Jun 2017) — built ERP systems for Matco Heavy Gear Equipment and Geely Motors, including a check-management module integrated with accounting workflows, and mobile and core-banking integrations.
- **Design Unit Manager, DAR ELMUSTAGBAL Organization** (Jan 2017 – Jun 2017) — developed accounting, HR and fingerprint-integrated attendance systems; implemented Odoo for MAS Health Care.

TECHNICAL SKILLS

Agentic AI: LangGraph, LangChain, LlamaIndex, AutoGen, CrewAI, Model Context Protocol (MCP), function/tool calling, multi-agent systems, ReAct, Plan-and-Execute

LLMs & RAG: OpenAI GPT-4o, Claude, Gemini, Llama, Qwen, Mistral; Milvus, Qdrant, pgvector, Elasticsearch; GraphRAG, hybrid search, re-ranking

Fine-Tuning & Serving: LoRA, QLoRA, PEFT, SFT, DPO, quantization (GPTQ/AWQ); vLLM, Triton, TensorRT-LLM, TGI, KServe

Speech AI: ASR, TTS, streaming Speech-to-Speech, Whisper, VAD, LiveKit, WebRTC

LLMOps & Evaluation: Kubeflow, MLflow, Airflow, W&B, CI/CD for ML, canary deployments; RAGAS, DeepEval, LangSmith, Langfuse, NeMo Guardrails

Cloud & Infrastructure: GCP (Vertex AI, GKE, Cloud Run), AWS (SageMaker, Bedrock), Azure OpenAI; Kubernetes, Docker, Helm, Terraform, Pulumi

Backend & Observability: Python, FastAPI, Redis, PostgreSQL, Kafka, REST/gRPC/WebSockets; Prometheus, Grafana, OpenTelemetry, Evidently AI

AI Security: OWASP LLM Top 10, prompt-injection mitigation, tool-use guardrails, JWT, RBAC/ACL, PII protection

EDUCATION

MSc, Big Data Technology & Artificial Intelligence — University of East London, United Kingdom (in progress)

BSc, Software Engineering — Sudan University of Science & Technology, Khartoum (2011 – 2015)

CERTIFICATIONS & LANGUAGES

Certifications: Generative AI & Workplace Transformation · Building Chatbots with OpenAI & LangChain · SAP Financials — Essential Training · Cisco Certified Network Associate (CCNA)

Languages: Arabic (native) · English (professional proficiency)