

Ammar Hamdi Hammouch *AI Engineer*

ahlan@ahammouch.me | 055 498 4459 | Riyadh | [linkedin.com/in/ammam-hammouch](https://www.linkedin.com/in/ammam-hammouch) | github.com/AmmarsGeeks

PROFESSIONAL SUMMARY

Works on the infrastructure side of AI, not just the model calls. At Entropy, built the full dataset ingestion pipeline for LLM training data — quality gates, Label Studio annotation, BullMQ + Redis orchestration, upload to annotated export with no manual step — and connected an internal RAG chatbot to Teams and Slack via FastAPI. Built an LLM-evaluation data platform (Balsam) with three scoring tiers including an LLM-as-a-Judge, and a durable court-transcription bot coordinating STT and LLM calls over 4+ hour sessions. Currently working on Arabic-language LLM evaluation and benchmarking.

EDUCATION

Bachelor in Computer Science, Jazan University
GPA: 4.86 / 5 — First Class Honours

2021 – 2025

PROFESSIONAL EXPERIENCE

Software Engineer, Entropy

09/2025 – Present

LLM evaluation platform benchmarking Arabic-language model performance.

- Built a microservices ingestion pipeline for LLM training datasets: upload verification and corruption scanning, format validation for CSV, XML, and JSON, metadata quality checks, Label Studio annotation integration, and BullMQ + Redis job orchestration. Datasets move from upload to annotated export with no manual step between stages.
- Connected an internal RAG chatbot to Microsoft Teams and Slack via FastAPI — connection management, reconnection on drop, and query routing to an indexed internal knowledge base.
- Contributing to dataset management, annotation pipelines, and model scoring infrastructure for the Arabic-language LLM benchmarking platform.

Integration Specialist, Thmanyah

01/2025 – 05/2025

Media and podcast company. Internal systems across HR, TA, and Marketing.

- Built an internal HR platform covering payroll, notice periods, and 3-month probation evaluations — PDF report generation via Puppeteer, job queuing on BullMQ + Redis, and heavy schema migration work on Drizzle ORM.
- Designed a multi-source aggregation engine pulling from Tally, Google Forms, Airtable, and Excel into an internal CRM. Configurable flows similar to n8n, auto-distributed performance reports, and roughly 60% less manual reporting time.

React Native Developer, Motmaina Medical Center

04/2024 – 12/2024

Telehealth consulting platform for mental health.

- Built the React Native app end to end — state architecture managing health session state, patient–doctor handoffs, and booking flows across the full care cycle.
- Built the admin dashboard in Next.js with React Query (hydration, interceptors, SSR, query invalidation) for real-time visibility into sessions, providers, and clinical activity.
- Integrated real-time chat and a voice/video SDK; AWS S3 + Lambda for secure media upload and processing.

Technology Integration Intern, Reachware

08/2023 – 04/2024

iPaaS platform serving SMB and enterprise clients across the Middle East.

- Built and deployed RAG chatbots for internal knowledge retrieval, including LLM fine-tuning experiments on domain-specific corpora to improve response accuracy.
- Worked on ZATCA e-Invoice Phase 2 (FATOORA) compliance integration: XML schema validation, cryptographic invoice signing, and real-time clearance against the ZATCA API.
- Built data integration pipelines between Snowflake and ServeMe for financial reporting and analytics.

Ios Application Developer, COMT

08/2022 – 08/2023

F&B technology.

- Architected a full POS platform from scratch, comparable to Foodics and Rewaa: inventory, product catalog, and categories built into a transaction cycle that runs from entry through order, payment, and receipt in under 30 seconds.
- Built offline-first architecture with real-time sync, so the system keeps running through network interruptions and reconciles when the connection returns.

PROJECTS

Balsam: LLM Evaluation Pipeline

Batch LLM evaluation platform running an 11-task Airflow DAG: dataset validation, Spark ETL for preprocessing, human-in-the-loop annotation routing, and three scoring tiers — classical metrics, semantic similarity, and LLM-as-a-Judge — producing ranked leaderboards. The pipeline ingests datasets through format validation and metadata quality checks before they reach the evaluation stage, with schema contracts enforced in CI so downstream dashboards never see broken data.

Moyassier — Durable Meeting-Intelligence for Saudi Courts

Headless bot that joins a 4+ hour court session, records and transcribes it via STT, and runs structured LLM extraction to produce intelligence for the judge. Built on Restate durable execution, journaling each stage so a crash resumes from the last completed step instead of restarting an unrepeatable live recording. Chose Restate over Temporal for a single-binary in-region footprint under PDPL data residency, native Kafka alignment, and a concurrency model suited to dozens of mostly-idle concurrent sessions.

Mozn Platform — Mizan: LLM-Ops Observability

Real-time observability service for the LLM evaluation platform, built on Kafka and Faust with EWMA and z-score anomaly detection over streaming evaluation metrics. Sinks to TimescaleDB for time-series dashboards and Iceberg for long-term analytical storage, giving the platform operators visibility into scoring drift, pipeline latency, and annotation throughput.

SKILLS

Languages

Python, TypeScript, JavaScript, SQL

Infrastructure

Redis, BullMQ, Kafka, Docker, Webhooks

Databases

PostgreSQL, MongoDB, Firebase Realtime DB

AI and MLOps

FastAPI, RAG and vector databases, Label Studio, Airflow, LLM-as-a-Judge, model evaluation, embeddings, HuggingFace

Frameworks

NestJS, Next.js, React Native

Cloud

GCP (Cloud Run, Cloud Functions, Firebase), AWS (S3, Lambda)

CERTIFICATIONS

OCI Certified AI Foundations Associate — Issuer: Oracle | Date:

2025-10

AWS Cloud Practitioner CLF-C02 — Issuer: AWS

LANGUAGES

Arabic — Native, English — Professional