

MATHIEU AJAKA

AI Solutions Architect | Enterprise GenAI | Unstructured Data & LLM Systems | Financial AI

+966 58 095 0577 | mathieuajaka14@gmail.com | www.mathieuajaka.com

SUMMARY

AI Solutions Architect with a track record of designing and delivering production AI systems for enterprise clients across the GCC and Europe. Specialised in translating complex operational and data problems into scalable LLM and agentic platforms, with particular depth in unstructured document processing, bilingual RAG pipelines, and regulated-environment deployment. Built and shipped multi-agent platforms from 0 to production as sole technical owner across legal, regulatory, brand, and financial domains. Brings rare cross-functional fluency: technical architecture ownership combined with direct client engagement, solution scoping, and stakeholder communication. Currently seeking AI Solutions Architect, AI Engineer, GenAI Engineer, or Applied AI roles within Saudi Arabia (KSA), contributing to organizations building large-scale AI capabilities and digital transformation initiatives.

CORE SKILLS

- **Solution Architecture & Client Delivery:** end-to-end AI solution design, client scoping, stakeholder communication, technical pre-sales support, solution documentation, cross-functional delivery
- **LLM & GenAI Systems:** RAG pipelines, OpenAI SDK, Azure OpenAI Service, Azure AI Studio, Azure ML, vLLM, LangChain, prompt engineering, QLoRA / LoRA fine-tuning (PEFT, TRL), open-weight + API routing, adapter gating, conversational AI
- **Unstructured Data & Document AI:** document parsing, domain-aware chunking, clause-level extraction, structured workflow design, hybrid retrieval, multi-format ingestion, financial document processing
- **AI Agents & Agentic Systems:** LangGraph, tool-use agents, multi-agent orchestration, HITL checkpoints, agent evaluation and guardrails, OpenAI SDK, Anthropic SDK, state machines
- **Arabic NLP & Multilingual AI:** CAMEL Tools, morphological tokenisation, Arabic-aware chunking, BGE-M3, multilingual-e5-large, bilingual eval set construction, multilingual RAG
- **Evaluation & Observability:** RAGAS, LLM-as-judge with calibration, Langfuse, Prometheus, Grafana, human-in-the-loop review, golden-set release gates
- **MLOps & Infrastructure:** Docker, Kubernetes, GitHub Actions CI/CD, ECR, MLflow, W&B, AWS EC2 g5/g6, on-prem GPU serving, private cloud deployment, air-gapped environments
- **Data Engineering:** Kafka, Airflow, PostgreSQL, Parquet on S3, Elasticsearch, Pandas, Polars
- **ML Core:** Python, FastAPI, PyTorch, Hugging Face Transformers, XGBoost, LightGBM, scikit-learn

EXPERIENCE

AI Solutions Architect | WAJEEH Agency | Riyadh, Saudi Arabia

Jan 2025 – Present

Sole technical hire and de facto technical lead for an AI agency serving Saudi enterprise clients across CX, brand, and regulatory domains. Responsible for full-cycle solution design: scoping client problems, architecting AI systems, leading technical delivery, and owning production operations. Mentor junior engineers on production delivery standards.

- Designed and architected AI solutions for ~12 Saudi enterprise accounts: translated client operational and CX problems into LangGraph-orchestrated agentic workflows with state machines, HITL checkpoints, and tool-use routing across Arabic and English pipelines. Reduced campaign execution from ~3 days to <6 hours. Zero failed client handoffs across all accounts.
- Diagnosed and resolved Arabic unstructured document retrieval failure: default BM25 collapsed on clitic-heavy queries (ف, ل, و, prefixes) in regulatory and brand corpora, producing silent hallucinations. Redesigned with CAMEL Tools morphology-aware chunker, BGE-M3 dense retrieval, and RRF fusion. Recall@10: 0.61 to 0.89.
- Built model-routing layer for cost-efficient LLM delivery: local Llama 3.1 8B / Mistral 7B for templated content; Claude Sonnet / GPT-4o for regulated and executive-facing outputs. Reduced per-token cost by ~68% vs. API-only baseline.
- Established agent evaluation discipline where none existed: RAGAS and LLM-as-judge with calibration caught 2 retrieval-drift episodes before client delivery. QLoRA adapter release gate (4-bit NF4, rank 32, alpha=64) on 180-query held-out set rejected two adapters that passed voice but regressed factual grounding.
- Built CI/CD pipeline: GitHub Actions re-evaluates against golden set on every change; ECR stores versioned images; EC2 promotion blocked until RAGAS thresholds pass. No manual override to production.
- Deployed production LLM serving infrastructure: vLLM + AWQ 4-bit on AWS g5.xlarge (A10G). Benchmarked vs. TGI; vLLM held TTFT ~420ms median / p95 <950ms at concurrency 4-8.

AI Engineer (Contract) | ARA Technologies MEA | Dubai, UAE

Mar – Dec 2024

Brought in to diagnose and rebuild the ML and data platform underpinning ARA's analytics product. Delivered production-grade ML systems and search infrastructure; provided technical direction to 2 internal engineers.

- Resolved Elasticsearch p95 regression (7 nodes, ~11M docs): profiled nested query fan-out, shard skew, missing custom routing via esrally replay. p95: 4.8s to 780ms, unblocking a BI product stalled on SLA for two quarters.
- Re-architected cron ingestion to Kafka: at-least-once delivery, idempotent upserts, offset-replay on failure. Data-gap incidents went from ~1/week to zero.
- Rebuilt client segmentation: XGBoost on stratified and time-based holdout (vs. prior random-split baseline), +14pp recall on high-value segment at fixed precision. MLflow introduced as model lifecycle backbone.
- Worked directly with enterprise stakeholders to integrate AI workflows into operational and client-facing business systems.

Freelance AI Solutions Engineer Independent Europe & GCC Three engagements on bilingual document retrieval, anomaly detection, and ML infrastructure. Sole technical owner on each; technical oversight of integrator teams at Takamol and Presight.	2023 – 2024
Takamol Holding — Arabic Legal Document Intelligence Riyadh, KSA - Designed and delivered a bilingual AI solution for unstructured workforce-compliance documents: Saudi Labor Law, ministry circulars, Qiwa/Musaned policy. Domain-aware chunker + multilingual-e5-large + BM25/RRF hybrid retrieval. Hit@5 = 0.91 on 140-query bilingual eval set. - Deployed on air-gapped Oracle Cloud with zero outbound data transfer: full data privacy and regulatory compliance for a government-adjacent entity. Entered active enterprise use within 30 days of handoff.	Jun – Sep 2024
Presight AI — TAQ Ingestion and Anomaly Detection Abu Dhabi, UAE ETL from heterogeneous multi-sector sources into unified analytical indexes with incremental-update support (~3.4M rows/day). Anomaly detection on operational telemetry: Isolation Forest (unsupervised) + XGBoost on labeled incidents with time-based holdout. MLflow for full experiment provenance.	Feb – May 2024
E-commerce Platform — Retrieval and Segmentation France Ingestion pipeline from PostgreSQL + Braze CRM webhooks into retrieval index. XGBoost segmentation with time-based holdout: +11pp F1 over logistic-regression baseline.	Oct – Dec 2023
Data and ML Analyst (Project-Based) BREI Paris, France Python/SQL telemetry pipelines. Isolation Forest anomaly detection on sensor data. First production ML work with threshold calibration and false-positive triage practices.	Mar – Jul 2023
Data Analyst (Project-Based) ALUSTEEL GROUP Jbeil, Lebanon Consolidated SQL Server and Excel operational data into unified analytics layer. Automated weekly reporting, ~80% reduction in manual effort. Baseline demand forecasts for production planning.	Jun – Sep 2022
EDUCATION	
Master's in Applied Artificial Intelligence Universite Saint-Joseph (USJ) Beirut, Lebanon Thesis: TCN vs. ANFIS under concept drift on financial time-series (BTC-USD and ETH-USD 2018–2024).	2024 – 2026
Bachelor's in AI and Business IA School Paris, France	2021 – 2024
CERTIFICATIONS	
- Google Cloud Professional ML Engineer — Expected June 2026 - Deep Learning Specialization (Andrew Ng) — Coursera, 2023	
SELECTED PROJECTS	
TradeBot — Autonomous Trading System Crypto Perpetual Futures (Bybit, BTC/USDT) Personal R&D, 2025 – Present	
- Runs the full trade loop end to end: a deep-learning chart-signal model and a market-regime classifier drive entries, while the system sizes positions and leverage dynamically and manages the complete order lifecycle with broker reconciliation, automatic recovery, and a runtime kill switch. - Trades selectively through an adaptive gate that acts only on high-confidence, regime-favorable setups, and reads live order-flow and order-book microstructure (buy/sell pressure, book imbalance, liquidity walls) to time entries beyond what price candles reveal.	
Saudi Labor Law RAG (open-source)	
Bilingual legal retrieval over Saudi Labor Law corpus: clause-level chunking preserving citation hierarchy, hybrid retrieval + BGE reranker. 120-query eval set across 8 legal domains: Hit@5 = 0.93, MRR@10 = 0.81. Containerized FastAPI service and public reference architecture for compliance-critical Arabic retrieval.	
LANGUAGES	
Arabic (native) French (native) English (full professional)	