Mohamed Rayen Fekih

AI/ML Engineer | Generative AI · RAG · Agentic Systems

rayen.fekih6@gmail.com | +216 52 591 307 | LinkedIn | GitHub | Portfolio | Open to Relocation

PROFILE

AI/ML Engineer with 5+ years building production Generative AI and ML systems: Arabic RAG over **165K+** documents, multi-agent assistants serving **1,000+** institutes, and recommendation and forecasting pipelines. Claude Certified Architect and Google Cloud Certified Professional ML Engineer, experienced in agentic workflows, LLM evaluation and cloud-native deployment on GCP.

TECHNICAL SKILLS

GenAI & Agents: Agentic Systems, Tool Calling, RAG, Semantic Search, LLM Evaluation, LangGraph, LangChain, Google ADK, Agent Engine, Gemini, Hugging Face, Arabic NLP
Vector Search: BigQuery Vector Search, Qdrant, BM25, RRF
Machine Learning & CV: PyTorch, Scikit-learn, Recommendation Systems, Forecasting, YOLO, RF-DETR, OpenCV, Roboflow
Languages & Data: Python, SQL, PySpark, BigQuery
Cloud & MLOps: GCP, Vertex AI, Cloud Run, FastAPI, Docker, CI/CD, MLflow
Spoken Languages: Arabic (native), English (full professional), French (professional)

CERTIFICATIONS

Google Cloud Certified: Professional Machine Learning Engineer | Google Cloud | 2026 | Credential
Claude Certified Architect, Foundations (CCAR-F) | Anthropic | 2026 | Credential

EXPERIENCE

AI Engineer, iHorizons - Qatar **Jul 2025 - Present**

Arabic Semantic Search and RAG Platform | FatwaTok

- Owned the core implementation of hybrid semantic search across **165K+** Arabic documents for Qatar's Ministry of Endowments and Islamic Affairs, improving retrieval beyond keyword matching.
- Designed the retrieval architecture with Gemini embeddings, BigQuery Vector Search, BM25 and RRF.
- Reduced end-to-end latency from **12s to 4s** by optimizing retrieval, vector search and BigQuery execution.
- Operated **3** production FastAPI services for semantic search, summarization and Arabic TTS.

AI Content Processing Pipeline

- Automated the daily content lifecycle so production data and search indexes stay current without manual work.
- Built Python/GCP pipelines on Cloud Run, Cloud Scheduler and BigQuery for embeddings, Arabic summarization, classification, Arabic TTS and search-index refresh.

Multi-Agent Khateeb Assistant

- Led the technical development of a multi-agent assistant that helps sermon writers generate grounded content from Quran, Hadith and trusted Islamic books, supporting adoption across **1,000+** institutes in Qatar.
- Co-designed the architecture and implemented a Google ADK / Agent Engine harness for planning, retrieval, generation, evaluation, moderation and revision.
- Supported the migration from an n8n proof of concept to production, with strict content constraints.
- Evaluated retrieval relevance, answer accuracy, latency and cost to guide model selection.

Computer Vision Oil Rig Safety Monitoring

- Developed CCTV monitoring for PPE violations and restricted-zone entry; fine-tuned and benchmarked YOLO and RF-DETR with PyTorch and Roboflow.
- Built OpenCV multi-object tracking and event detection, using MLflow for experiment tracking.

Machine Learning Engineer, Infor - New York, USA **May 2023 - Jun 2025**

GenAI Assistant

- Designed and built an agentic GenAI assistant with LangGraph, tool calling, contextual memory, Docker and Qdrant for RAG.
- Implemented retrieval workflows and optimized context management for scalable LLM applications.

Pricing Optimization and Recommendation Systems

- Led development of an ML pricing engine deployed for **3** customers, contributing to a **5% profit increase** for an enterprise customer; guided junior engineers.
- Reduced production inference latency by **80%** through Cython and NumPy optimization.
- Built recommendation models for upsell and cross-sell, and implemented CI/CD pipelines to automate builds and deployments.

Data Scientist, Cognira - Tunisia **Feb 2021 - May 2023**

Retail Demand Forecasting

- Built Python, PySpark and SQL time-series forecasting pipelines over **10M+** retail transactions within a 7-person team.
- Monitored production models for degradation and training-serving skew with automated alerts.

Process Quality Optimization

- Built regression and explainability models for a cheese manufacturer, contributing to a **15%** profit-margin improvement.

OPEN SOURCE

hadith-detector | PyPI | GitHub

- Built and published an offline Arabic Hadith detection library for source retrieval and LLM output verification: **14,940** records, trigram-anchored exact matching (**~1.1 ms**) and RapidFuzz fuzzy matching (**~4 ms**) on 60-word inputs.

EDUCATION

National Engineering School of Tunis (ENIT) **2016 - 2021**
Bachelor's Degree in Computer Science - Artificial Intelligence Specialization