Aymen Rahmani

LLM & AI Agents Engineering Intern (PFE)
Ariana, Tunisia (open to relocation, Europe) | +216 93 402 936 | aymen.rahmani@esprit.tn | linkedin.com/in/aymen-rahmanii | github.com/Aymenrahmanii

Summary

Final-year Software & Data Science engineering student (ESPRIT) and evaluation-first LLM engineer: RAG, agents and MCP with measured results, CI eval gates and prompt-injection testing, deployed on Azure. Seeking a 4–6 month end-of-studies internship (PFE) from Feb 2027, internship agreement via ESPRIT.

Technical Skills

LLMs & agents: LangChain, LangGraph, hybrid RAG (BM25 + dense, reranking), GraphRAG, agentic RAG, multi-agent systems, tool calling, MCP, guardrails, Ollama, n8n
Eval & security: Recall@k, MRR, faithfulness, LLM-as-a-judge, CI eval gate, prompt-injection testing, PII masking, OpenTelemetry
ML & data: Python (proficient), SQL, PyTorch, OpenCV, Transformers, Scikit-learn, XGBoost, Pandas, Qdrant, PostgreSQL, Azure AI Search
LLMOps & cloud: FastAPI, Docker, Git, GitHub Actions (OIDC), Azure (OpenAI, AI Search, Container Apps), Terraform, MLflow, Kafka, ONNX
Languages: Arabic (native), English (C1), French (B2), German (B1–B2)

Projects

Regulatory RAG Assistant on Azure (GDPR, AI Act, NIS2, DORA) | 2026
Python, FastAPI, Azure AI Search, Azure OpenAI, Container Apps, Terraform, GitHub Actions, OpenTelemetry | Personal project
- Hybrid RAG (BM25 + dense, RRF), GraphRAG and agentic RAG behind one streaming API; the agent reaches **0.93 correctness** vs 0.81 for the baseline (60 questions).
- Versioned evaluation harness (40+ runs): **Recall@5 from 0.73 to 0.94**; a CI eval gate blocks any PR that lowers answer quality; deploys with approval and rollback.
- Document-level security trimming, PII masking, **0 of 240 prompt-injection attacks succeeded**; about $3.5 per 1,000 requests; 139 tests.

CrisisLink – Multimodal Flood Triage with LLM Agents | 2026
n8n, Groq (gpt-oss, Qwen VL, Whisper), Gemini (LLM + gemini-embedding-001), Qdrant | Team of 6 (Esprit 5DS final project)
- Telegram bot (photo, Derja/French/English voice, text, GPS) where AI agents estimate flood depth, extract distress, retrieve official protocols and grade incidents P1–P4 behind a deterministic safety check; orchestrated in n8n.
- **I own the Protocol RAG agent**: structure-aware chunking of the official protocols, Gemini embeddings and Qdrant search; **Hit@3 0.95, MRR 0.935** on the retrieval evaluation (24 chunks, 10 cited sources).

Autonomous Multi-Agent DevOps/MLOps Platform | 2026
LangGraph, Kafka, gRPC, PostgreSQL, Qdrant, SonarQube, Semgrep, Trivy, Gitleaks, MCP | Team of 6
- Built the platform's **orchestrator** and **context service**, which coordinate the agents that each teammate developed.
- Also built the **Security & Quality agent**: tests, coverage, lint, CVEs (Trivy), secrets (Gitleaks) and SonarQube merged into one ALLOW / WARN / BLOCK release verdict, wired into CI and exposed as an **MCP server**.

DriftScope MCP – Cloud-Deployed MCP Server for Data-Drift Auditing | 2026
Python, MCP SDK (FastMCP, SSE), Polars, SciPy, scikit-learn, Render | Personal project
- Built and deployed an MCP server that gives LLMs deterministic statistical tools (two-sample KS test, Population Stability Index with severity labels) to audit data drift in tabular ML pipelines, so the LLM handles triage and reporting while the maths stays exact.
- Exposes tools, a drift-policy resource and an audit prompt over SSE; usable from Claude Desktop through a remote MCP endpoint.

NCM Lab – Benchmarking a Coarse-to-Fine Text Generator | 2026
PyTorch, Hugging Face Transformers, FAISS, FastAPI, Prometheus, Docker, GitHub Actions | Personal project
- Built a toy Next-Concept Model (NCM) from scratch (multiscale VQ tokenizer, ~5.6M params) and a benchmark harness comparing the official NCM checkpoints with same-size autoregressive (AR) models on a 6 GB laptop GPU; reproduced the paper's XSum **ROUGE-L 27.8 vs 28.0**.
- Evaluated NCM as a RAG context compressor and inside a tool-calling agent, plus an 8-experiment limits suite (latency, batch size, faithfulness, robustness): compression dropped RAG accuracy **from 87.5% to 10%**; AR models were 2–10x faster at batch 32. FastAPI service with Prometheus metrics, Docker, 55 tests and CI.

Experience

Computer Vision Intern | Jun 2026 — Aug 2026
Supra Services & Technology · Internship | Tunis, Tunisia (hybrid)
- Engineered an Automated Optical Inspection (AOI) system for sub-millimetre PCB defect detection: hybrid dual-engine pipeline (RT-DETR-L supervised + PatchCore unsupervised) with EIoU + NWD losses, **+24.15% mAP50:95 (77.76%)** and +6.67% mAP75.
- Implemented SIFT + RANSAC homography for real-time template registration and added Eigen-CAM SVD heatmaps (explainable AI) as visual feedback for operators.
- Built and deployed multi-threaded REST/WebSocket APIs with FastAPI for **30 FPS** video streaming and SQLite logging (ONNX).
- *Tech: Python, PyTorch, RT-DETR, PatchCore, OpenCV, FastAPI, SIFT, SQLite, ONNX, WebSockets*

Machine Learning Intern | Jul 2025 — Aug 2025
Web Media International · Internship | Tunisia (hybrid)
- Developed and deployed an end-to-end ML classification service for email validation and workflow routing; built an XGBoost model: **90% accuracy, 91 F1-score** (feature engineering, hyperparameter tuning, cross-validation).
- Integrated Python ML microservices (FastAPI/Flask) with a Spring Boot backend and MySQL database for real-time inference.
- *Tech: Python, XGBoost, Scikit-learn, FastAPI, Flask, Spring Boot, MySQL, REST APIs*

Education

Engineering Degree (Master's level), Software & Data Science, *ESPRIT School of Engineering* | Sep 2024 – Jul 2027 (exp.)
Bachelor's Degree in Embedded Systems, *Higher Institute of Computer Science (ISI)* | Sep 2021 – Jul 2024

Certifications

NVIDIA DLI (6 certificates, 2025–26), incl. **LLM application development** and **LLM evaluation**; also deep learning, transformer NLP, prompt engineering. **AWS Academy Cloud Foundations** | **MASSAI AI Summer School** (Jun 2025, 30 h). Also (Nov 2026): Anthropic Academy – Introduction to MCP; 365 Data Science – Evaluating AI Agents, Intro to AI Agents and Agentic AI.