Otmane El Bourki

Generative AI Engineer

LLM, RAG, Multi-Agent & Cloud AI Systems

Paris, France

☎ +33 7 75 73 57 51
otmane.elbourki@gmail.com
linkedin.com/in/oelbourki
github.com/oelbourki

SKILLS

LLMs & Agentic Frameworks: Multi-Agent Systems (LangGraph, CrewAI, Autogen), RAG & GraphRAG (LangChain, Neo4j), Prompt Engineering (CoT, Few-Shot), DSPy
Model Optimization & Serving: Fine-tuning (LoRA/QLoRA), 4-bit Quantization (AWQ/GPTQ), vLLM, TorchServe, TensorRT, ONNX Runtime, GPU Optimization (CUDA)
Cloud & MLOps: GCP (Vertex AI, Cloud Run), AWS (SageMaker, EC2), Docker, Kubernetes, CI/CD (GitHub Actions), Prometheus, Grafana
Data & Vector Databases: Neo4j (Cypher), Redis, Qdrant, ChromaDB, Vector Embeddings & Similarity Search
Core Engineering: Python, PyTorch, Hugging Face, REST APIs, Microservices

PROFESSIONAL EXPERIENCE

Artificial Intelligence Engineer (GenAI & Multi-Agent Systems Lead) Feb 2025 - Present
Impactera *— Financial Consulting Firm* *Paris, France*

- Lead the end-to-end development of a specialized AI agent platform, translating C-level needs into scalable architectural roadmaps across model selection and infrastructure design.
- Architected a secure FastAPI/Redis ingestion pipeline for financial data (CCH Tagetik), ensuring GDPR compliance and reducing manual reporting overhead by 60%.
- Engineered an automated executive briefing system for financial trend analysis, deploying containerized microservices to cloud infrastructure to support enterprise decision-making

Artificial Intelligence Engineer (RAG & LLM Specialist) June 2023 - June 2024
Tetricks *— AI Solutions Startup for Hospitality* *Palaiseau, France*

- Deployed a production-grade RAG chatbot on GCP (Vertex AI, Cloud Run), cutting infrastructure costs by 47% with automated CI/CD retraining pipelines and containerized microservices
- Diagnosed and resolved critical latency bottlenecks by implementing 4-bit quantization (GPTQ), vLLM serving optimization, and LoRA fine-tuning, achieving a **3x increase in inference throughput**.
- Spearheaded the integration of Vision LLMs for automated room description generation, acting as the primary technical lead for a Proof-of-Concept that successfully secured Bpifrance innovation funding.

Machine Learning Research Engineer Feb 2022 - July 2022
Computer Science Department of Ibn Tofail *— Research Institution* *Kenitra, Morocco*

- Developed an emotion recognition system with CNN, achieving 90% accuracy and 20% improved robustness.

PROJECTS

Enterprise GraphRAG for Financial Document Intelligence | *Neo4j, Qdrant, LangChain, Llama-3*

- Engineered hybrid Vector Search + Knowledge Graph system for multi-hop reasoning over 50K+ documents, reducing hallucinations by **40%** through semantic relationship modeling and context-aware retrieval.

Multimodal Document Processing Pipeline | *GPT-4V, LlamaVision, FastAPI, Redis*

- Built multimodal pipeline with Vision LLMs for automated invoice/receipt processing with **96% accuracy**. Deployed via FastAPI processing 1K+ documents daily, reducing manual data entry by **75%**.

Intelligent LLM Gateway with Cost-Optimized Routing | *FastAPI, Redis, Qdrant, OpenRouter, Prometheus*

- Architected LLM gateway with dynamic routing to optimal models (GPT-4, Llama-3-70B) based on complexity and cost. Implemented semantic caching reducing API costs by **60%** and serving **42%** of queries in sub-50ms.

EDUCATION

IT Architecture Expert (RNCP7 — Master) — Data Architecture 2024–2025
École 42 Paris *Paris, France*

Master's in Artificial Intelligence (With Honors) 2022–2024
Ibn Tofail University *Kenitra, Morocco*

Bachelor's in Mathematics & Computer Science 2018–2022
Ibn Tofail University *Kenitra, Morocco*

CERTIFICATIONS:

GenAI Winter School, Ecole Centrale Casablanca (Mar 2024)
Practical Data Science, AWS (Apr 2022)
TensorFlow Developer, Deeplearning.ai (Oct 2020)

LANGUAGES:

French (Fluent)
English (Fluent)
Arabic (Fluent)

COMMUNITY:

Founded 1337AI.
Mentored in ThinkAI & HackAI hackathon.