Otmane El Bourki

AI Engineer - Large Language Models, Multi-Agent Systems, MLOps (Machine Learning Operations)
otmane.elbourki@gmail.com | github.com/oelbourki | linkedin.com/in/oelbourki | Paris, France | +33775735751

Skills

LLM & Agentic Systems: RAG (Retrieval-Augmented Generation), Multi-Agent Systems, Prompt Engineering, LLM Evaluation, GPT Models, ChatGPT, Google Gemini
Model Optimization: LoRA, QLoRA, 4-bit Quantization (AWQ, GPTQ), vLLM, ONNX Runtime, CUDA
MLOps & Cloud: MLOps (Machine Learning Operations), GCP (Vertex AI, Cloud Run), AWS (SageMaker, EC2), Docker, Kubernetes, CI/CD, GitOps, Prometheus, Grafana
Data Systems: Neo4j, Qdrant, ChromaDB, Redis, Vector Similarity Search, SQL, Pandas, NumPy
Engineering: Python, PyTorch, TensorFlow, Hugging Face Transformers, FastAPI, REST APIs, Microservices

Experience

Self-Employed - AI Engineer — Sep 2025 – Present, Paris, France

- Architected production-ready multi-agent AI platform for real estate, orchestrating 6 agents with 15+ tools, leveraging RAG pipelines and LLM evaluation.
- Built *codibox*, a secure Docker-based Python execution engine with sandboxing and artifact recovery for safe AI-generated code execution (PyPI).

Impactera - AI Engineer — Feb 2025 – Aug 2025, Paris, France

- Delivered production multi-agent system providing real-time financial insights, semantic search, and automated decision support with 99.9% uptime.
- Implemented data integration pipelines connecting CCH Tagetik APIs and internal databases, enabling natural language queries, SQL analysis, and automated reporting.

42 Paris - AI R&D Engineer — Jun 2024 – Feb 2025, Paris, France

- Engineered CUDA-optimized linear algebra engine and Brain-Computer Interface (BCI) pipeline for real-time EEG classification (90% accuracy, <2s latency).
- Built resilient distributed infrastructure using K3s, GitOps, and CI/CD automation for scalable workloads.

Tetricks - AI Engineer — Jun 2023 – Jun 2024, Palaiseau, France

- Deployed production RAG chatbot on GCP (Vertex AI, Cloud Run), reducing infrastructure costs by 47%.
- Optimized LLM inference using quantization, fine-tuning, and vLLM, achieving 3x throughput.
- Led Vision LLM integration for automated room description generation, securing Bpifrance funding.

Ibn Tofail University - ML Research Engineer — Feb 2022 – Jul 2022, Kenitra, Morocco

- Conducted applied research and benchmarking on facial expression recognition using FER2013, achieving 90% precision and evaluating state-of-the-art deep learning models for robust emotion classification.
- Enhanced model generalization and stability by 20% through advanced data augmentation, regularization, and ablation analysis to identify optimal architectures.

1337AI - Founder — Oct 2019 – Feb 2022, Benguerir, Morocco

- Led AI hub for 300+ developers, delivering workshops and hackathons on ML and GenAI projects.

Education

École 42 Paris - Paris, France — Master's-Level Diploma (RNCP7) in IT Architecture: Data & Database
Ibn Tofail University - Kenitra, Morocco — Master's in Artificial Intelligence
Ibn Tofail University - Kenitra, Morocco — Bachelor's in Mathematics and Computer Science

Certifications

GCP AI Infrastructure, GenAI Winter School (École Polytechnique), AWS Practical Data Science, TensorFlow Developer

Languages

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