Otmane El Bourki

Ingénieur IA - Modèles de Langage, Systèmes Multi-Agents, MLOps
otmane.elbourki@gmail.com | github.com/oelbourki | linkedin.com/in/oelbourki | Paris, France | +33775735751

Compétences

LLM & Multi-Agents : RAG, Systèmes Multi-Agents, Prompt Engineering, Évaluation LLM, GPT, ChatGPT, Google Gemini
Optimisation : LoRA, QLoRA, Quantification 4-bit (AWQ, GPTQ), vLLM, ONNX Runtime, CUDA
MLOps & Cloud : MLOps, GCP (Vertex AI, Cloud Run), AWS (SageMaker, EC2), Docker, Kubernetes, CI/CD, GitOps, Prometheus, Grafana
Données : Neo4j, Qdrant, ChromaDB, Redis, Recherche Vectorielle, SQL, Pandas, NumPy
Développement : Python, PyTorch, TensorFlow, Hugging Face, FastAPI, APIs REST, Microservices

Expérience

Indépendant - Ingénieur IA — Sep 2025 – Présent, Paris, France

- Conception d'une plateforme IA multi-agents prête pour la production dans l'immobilier, orchestrant 6 agents avec 15+ outils, exploitant des pipelines RAG et l'évaluation de LLM.
- Développement de *codibox*, moteur d'exécution Python sécurisé basé sur Docker avec sandboxing et récupération d'artéfacts pour l'exécution sécurisée de code généré par l'IA (PyPI).

Impactera - Ingénieur IA — Fév 2025 – Août 2025, Paris, France

- Livraison d'un système multi-agent en production fournissant des analyses financières en temps réel, recherche sémantique et support à la décision automatisé avec 99,9% de disponibilité.
- Implémentation de pipelines d'intégration de données connectant les API CCH Tagetik et bases internes, permettant des requêtes en langage naturel, analyses SQL et reporting automatisé.

42 Paris - Ingénieur R&D IA — Juin 2024 – Fév 2025, Paris, France

- Conception d'un moteur d'algèbre linéaire optimisé CUDA et d'un pipeline BCI (Interface Cerveau-Machine) pour classification EEG en temps réel (90% de précision, latence <2s).
- Mise en place d'une infrastructure distribuée résiliente avec K3s, GitOps et CI/CD.

Tetricks - Ingénieur IA — Juin 2023 – Juin 2024, Palaiseau, France

- Chatbot RAG en production sur GCP (Vertex AI, Cloud Run), réduction des coûts de 47%.
- Optimisation de l'inférence LLM via quantification, fine-tuning et vLLM, améliorant le débit par 3.
- Intégration Vision LLM pour description automatique de pièces, financement Bpifrance obtenu.

Université Ibn Tofail - Ingénieur Recherche ML — Fév 2022 – Juil 2022, Kénitra, Maroc

- Reconnaissance faciale FER2013 : 90% de précision avec modèles deep learning à l'état de l'art.
- Amélioration de la robustesse des modèles (+20%) via augmentation et régularisation.

1337AI - Fondateur — Oct 2019 – Fév 2022, Benguerir, Maroc

- Animation d'un hub IA pour 300+ développeurs : ateliers et hackathons ML/GenAI.

Éducation

École 42 Paris – RNCP7, Architecture IT : Données & Bases de Données
Université Ibn Tofail – Master Intelligence Artificielle
Université Ibn Tofail – Licence Mathématiques & Informatique

Certifications

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

Langues

Français (Courant), Anglais (Courant), Arabe (Courant)