

Achref SOUA

Ingénieur Machine Learning

Paris, France | achref.soua@outlook.com | linkedin.com/in/achref-soua | github.com/achref-soua

Profil

Ingénieur Machine Learning (≈ 3 ans) qui conçoit et déploie des **systèmes ML en production** : agents LLM & RAG, vision par ordinateur (imagerie médicale 3D) et infrastructure de recherche vectorielle haute performance.

Expérience

Data Scientist, PrediSurge – Saint-Étienne, France Oct. 2025 – Présent

- Industrialisation du pré-traitement d'imagerie : pipeline automatisé DICOM → NIfTI (sélection de série par heuristique de qualité) supprimant le tri manuel des examens CT en entrée du produit **ISizing**.
- Conception et entraînement de modèles de **segmentation 3D multi-labels** (25 structures vasculaires), **Dice global de 94%**, pour la planification endovasculaire.
- Pipeline de **sizing** automatique (**VTK**) : maillages 3D + centerlines, puis par label diamètres (min/max/moyen tous les 1 mm), volumes, angulation et tortuosité en **~1 min/cas**, alimentant un **moteur de recommandation de stents** (artères viscérales).
- Registre de données central (scans NIfTI + métadonnées) versionné avec **DVC** : traçabilité et reproductibilité des jeux d'entraînement.

Data Scientist, Moovto – Paris, France Janv. 2025 – Sept. 2025

- Assistant **FinOps multi-agents** (LLaMA 3, LangGraph) pour l'analyse des coûts cloud, déployé sur **AWS** (Lambda, S3, Bedrock) avec CI/CD, ramenant l'analyse des coûts de plusieurs heures à quelques minutes.
- Optimisation de la mémoire long-terme des dialogues multi-tours : **+23% de cohérence** des réponses.

Data Scientist, Intellect Core SARL – Grombalia, Tunisie Nov. 2023 – Déc. 2024

- Base vectorielle et chatbot **RAG** conçus de bout en bout (LangChain), déployés en **serverless** sur AWS ; modèles open-source (**LLaMA 2**, **LeoLM**) et propriétaires (**GPT-4**, **Mistral**), **–50% de tickets** de support.

Projets

Quiver — Base de données vectorielle security-first (Rust) Lien GitHub

- Moteur écrit **from scratch en Rust** : ANN (HNSW, DiskANN, IVF + quantization), noyaux SIMD, chiffrement de bout en bout (XChaCha20, RBAC, audit). **32× moins de RAM** à recall égal via DiskANN+PQ (10 M×768-d : ~1 Go vs ~31 Go). Sur SIFT1M, **2^e derrière FAISS**, devant Qdrant, Milvus, Chroma, pgvector, LanceDB & Weaviate.

Pulse — Aide à la décision clinique (chirurgie endovasculaire) Lien GitHub

- Application **full-stack** (Next.js 15, FastAPI) sur données synthétiques : **RAG** (hybride + reranking) sur Quiver, copilote **multi-agents LangGraph** en streaming SSE (Groq), 6 scores cliniques validés, moteur de compatibilité (IFU-fit) et rapports PDF.

Helio — Plateforme growth marketing IA-native (open-source) Lien GitHub

- Monorepo **multi-tenant** (Postgres RLS, isolation au niveau base) : pipeline d'événements Redpanda → ClickHouse, parcours orchestrés sur **Temporal** (workflows durables), scoring de churn **ONNX/XGBoost** et copilote IA (segments & parcours en langage naturel). SSO/SCIM, serveur MCP.

Compétences

Langages: Python, Rust, TypeScript, SQL, C++

LLM & IA générative: LangChain, LangGraph, RAG, agents multi-agents, MCP, embeddings, vector search, Hugging Face (LLaMA, Mistral, GPT-4), Groq

ML & Vision: PyTorch, TensorFlow, scikit-learn, segmentation 3D, imagerie médicale (DICOM/NIfTI, VTK), ONNX, XGBoost, OpenCV

MLOps & Cloud: AWS (Lambda/S3/Bedrock), Docker, Kubernetes, CI/CD, DVC, Temporal, ClickHouse, PostgreSQL, Redis, FastAPI

Formation

Diplôme d'Ingénieur en Informatique — ESPRIT, Tunis

2023