

CONTACT

moussalfa94@gmail.com

06 29 53 10 31

Remote / Île-de-France

TJM : 650 EUR/jour

Disponible immédiatement

IA & LLM

RAG avancé

Chunking adaptatif, Reranking

Cohere, RRF, hybrid search

Agents & Orchestration

Juge agent, routing, CrewAI

Prompt engineering avancé

LLM

OpenAI, Mistral, Claude, Llama

Frameworks

Langchain, LlamaIndex, DSPy

Vector Stores

Qdrant, ChromaDB, Pinecone

OCR & Extraction

Unstructured, Tesseract, PDF

PERFORMANCE & QUALITÉ

Optimisation mémoire

Profiling, garbage collector

Détection memory leaks

Tests

Non-régression, TDD, pytest

Tests intégration, E2E

Observabilité

Logging structuré, monitoring

DÉVELOPPEMENT

Backend

Python, FastAPI, Laravel, Node

Frontend

Angular, React, TypeScript

BDD

PostgreSQL, MongoDB, Qdrant

API

REST, GraphQL, WebSocket, SSE

DEVOPS & CLOUD

Docker / Compose / Swarm

Scaleway, OVH, Linux, Nginx

GitHub Actions, GitLab CI

Reverse proxy, SSL, monitoring

LANGUES

Français — natif

Anglais — professionnel B2+

SOFT SKILLS

Leadership technique

Gestion de projet Agile

Mentorat & pédagogie

Autonomie totale

Communication client

VEILLE & CERTIFICATIONS

Veille IA quotidienne

Contributeur open-source

Lecture papers ArXiv

Communautés HuggingFace

DIVERS

Permis A & B, véhicule

Mobilité Île-de-France

CENTRES D'INTÉRÊT

IA générative & RAG

Trading algorithmique

Mécanique auto / aéro

Électronique & IoT

Moussa S.

Ingénieur IA & LLM — RAG, Agents, Pipelines

6 ans d'expérience | Fullstack Python/TS | DevOps Docker | Disponible immédiatement

Ingénieur IA spécialisé en systèmes RAG avancés, agents LLM et pipelines de traitement documentaire. 6 ans d'expérience en développement fullstack et intégration IA, de la conception au déploiement en production. Autonome sur l'ensemble de la chaîne : architecture, développement, DevOps, tests et mise en production. Disponible pour missions courtes, POC et audits IA.

EXPÉRIENCES PROFESSIONNELLES

Ingénieur IA & Développement Fullstack

Entreprise IT | Île-de-France

Mars 2020 – Présent (6 ans)

> Chatbot documentaire RAG : FastAPI + WebSocket streaming, chunking sémantique adaptatif, ChromaDB + reranking Cohere RRF, OCR Tesseract/Unstructured, évaluation RAGAS continue — -80% temps recherche, 50+ users, latence <2s

> Pipeline extraction documentaire IA : OCR multi-moteurs, classification LLM (type facture/contrat/rapport), routing intelligent, structuration JSON, déployé Scaleway Docker — 500+ docs/jour, 95% précision, remplacement 3 process manuels

> Génération de rapports automatiques : Langchain + GPT-4, templates Jinja2 dynamiques, export PDF/DOCX WeasyPrint, planification Celery Beat — 15h/sem économisées, 200+ rapports/mois

> Orchestration d'agents IA : multi-agents avec juge évaluateur, routing zero-shot par intention, prompt engineering avancé (few-shot, chain-of-thought), fallback GPT-4/Mistral/Claude — satisfaction 92%, coût API -60%

> 4 applications métier fullstack (Angular 17 + FastAPI + PostgreSQL + Redis) : gestion réparations, facturation automatisée, landing pages + Google Ads A/B testing — CPA -35%, leads +120%

> Système trading algorithmique : Docker 8 services, XGBoost + walk-forward 4000+ périodes, paper trading safety limits, 135+ tests pytest, monitoring Telegram 24/7

> Optimisation & CI/CD : profiling mémoire (tracemalloc, py-spy), fix memory leaks, GitHub Actions (lint/test/build/deploy), Nginx SSL, backups pg_dump auto — mémoire -40%, latence API /3

> Mentoring : 3 devs juniors, revues de code, guidelines archi/sécurité, formations RAG internes, gestion Agile (sprints 2 semaines)

Administrateur Réseaux & Systèmes

ESN | Val-de-Marne

2016

> Sécurisation et administration infrastructures IT, déploiement serveurs, support technique N2

PROJETS PERSONNELS

> Chatbot RAG multi-sources : chunking adaptatif, reranking Cohere RRF, Qdrant, streaming WebSocket, évaluation RAGAS

> Pipeline extraction documentaire : OCR multi-moteurs, classification LLM, structuration JSON automatique

> Agents IA orchestrés : multi-agents + juge évaluateur, routing zero-shot, mémoire contextuelle persistante

> Système embarqué : firmware C/STM32, filtre Kalman, PID temps réel, calibration IMU

FORMATION

Diplôme d'Ingénieur — Systèmes d'Information & Cybersécurité

ESIEA, Paris / Ivry-sur-Seine | IA, Cloud, Réseaux, Sécurité, Systèmes embarqués

2020 – 2024

DUT Génie Électrique et Informatique Industrielle (GEII)

IUT d'Évry | Electronique, informatique industrielle, PCB

2018 – 2020

Baccalauréat STI2D

Lycée Christophe Colomb, Sucy-en-Brie

2018

COMPÉTENCES TECHNIQUES

PythonTypeScriptFastAPIAngularReactLaravelNode.jsPostgreSQLMongoDBRedisDocker

Docker ComposeLangchainLlamaIndexOpenAIMistralClaudeRAGQdrantChromaDBPinecone

CohereDSPyCrewAIRAGASOCRTesseractUnstructuredScalewayNginxGitHub Actions

GitLab CIpytestXGBoostCeleryGraphQLWebSocketSSEGitLinuxAgile/ScrumJira

NotionFigmaPostmanSwaggerVS CodeCursor