

CONTACT

moussalfa94@gmail.com
06 29 53 10 31
Remote / Ile-de-France
TJM : 500 EUR/jour
Soirs & week-ends OK

IA & LLM

RAG avance
Chunking adaptatif, Reranking
Cohere, RRF, hybrid search
Agents & Orchestration
Juge agent, routing, CrewAI
Prompt engineering avance
LLM
OpenAI, Mistral, Claude, Llama
Frameworks
Langchain, LlamaIndex, DSPy
Vector Stores
Qdrant, ChromaDB, Pinecone
OCR & Extraction
Unstructured, Tesseract, PDF

PERFORMANCE & QUALITE

Optimisation memoire
Profiling, garbage collector
Detection memory leaks
Tests
Non-regression, TDD, pytest
Tests integration, E2E
Observabilite
Logging structure, monitoring

DEVELOPPEMENT

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

Francais -- natif
Anglais -- professionnel B2+

SOFT SKILLS

Leadership technique
Gestion de projet Agile
Mentorat & pedagogie
Autonomie totale
Communication client

VEILLE & CERTIFICATIONS

Veille IA quotidienne
Contributeur open-source
Lecture papers ArXiv
Communautes HuggingFace

DIVERS

Permis A & B, vehicule
Mobilité Ile-de-France

CENTRES D'INTERET

IA generative & RAG
Trading algorithmique
Mecanique auto / aero
Electronique & IoT

Moussa S.

Ingenieur IA & LLM -- RAG, Agents, Pipelines

6 ans d'experience | Fullstack Python/TS | DevOps Docker | Disponible soirs & week-ends

Ingenieur IA specialise en systemes RAG avances, agents LLM et pipelines de traitement documentaire. 6 ans d'experience en developpement fullstack et integration IA, de la conception au deploiement en production. Autonome sur l'ensemble de la chaine : architecture, developpement, DevOps, tests et mise en production. Disponible pour missions courtes, POC et audits IA.

EXPERIENCES PROFESSIONNELLES

Ingenieur IA & Developpement Fullstack

Entreprise IT | Ile-de-France

Mars 2020 -- Present (6 ans)

- > **Chatbot documentaire RAG** en production : architecture micro-services FastAPI avec WebSocket streaming temps reel et Redis cache. Pipeline d'indexation multi-formats (PDF, DOCX, images, tableaux Excel). OCR avance via Tesseract + Unstructured API avec pre-traitement image (rotation, skew, binarisation). Embeddings OpenAI ada-002, chunking semantique adaptatif (recursive + semantic splitting, overlap 20%), stockage vectoriel ChromaDB avec metadata filtering, reranking Cohere via Reciprocal Rank Fusion (RRF). Evaluation continue qualite via RAGAS (faithfulness, relevancy, answer correctness). Sources citees dans chaque reponse -- **recherche documentaire -80%, 50+ utilisateurs, latence <2s**
- > **Pipeline IA extraction documentaire** cloud-native : extraction automatisee de textes, tableaux et images depuis PDF/scans multi-pages. Detection automatique du type de document (facture, contrat, rapport). Normalisation et classification par LLM Mistral Small avec prompts specialises. Routing intelligent vers les bons pipelines selon le type detecte. Structuration JSON normalisee. Deploye sur Scaleway avec Docker Compose, auto-restart, logs structures et monitoring -- **500+ docs/jour, 95% precision, remplacement de 3 processus manuels**
- > **Generation de rapports automatiques** via Langchain + GPT-4 : analyse contextuelle de donnees metier issues de PostgreSQL, templates Jinja2 dynamiques, generation multi-sections avec table des matieres, export PDF/DOCX via WeasyPrint, integration API REST interne avec authentification JWT, planification automatique via Celery Beat -- **gain de 15h/semaine pour l'equipe metier, 200+ rapports genres/mois**
- > **Orchestration d'agents IA** : systeme multi-agents avec juge evaluateur et scoring de confiance, routing dynamique par intention (classification zero-shot), prompt engineering avance (few-shot, chain-of-thought, system prompts dynamiques), gestion de la memoire contextuelle par session, evaluation qualite RAGAS automatisee, fallback multi-modeles (GPT-4 > Mistral > Claude) -- **satisfaction utilisateur 92%, cout API -60% grace au routing**
- > **4 applications metier** en production (Angular 17 + FastAPI + PostgreSQL + Redis) : application de gestion de reparations en boutique (CRUD complet, suivi workflow multi-etapes, notifications push, historique client, export CSV), facturation automatisee avec TVA et relances, landing pages responsive + campagnes Google Ads avec tracking conversions GA4 et A/B testing -- **CPA -35%, leads +120%, 3 clients en production**
- > **Systeme de trading algorithmique autonome** : architecture Docker 8 services (FastAPI, PostgreSQL, Redis, Celery worker/beat, monitoring Telegram, backup auto). Pipeline ML complet : collecte OHLCV multi-sources (CCXT, yfinance), feature engineering (40+ indicateurs), entraînement XGBoost avec walk-forward validation (4000+ periodes testees), paper trading avec safety limits (drawdown -15%, daily loss -5%). Suite de 135+ tests unitaires et non-regression (pytest). Deploiement automatise 24/7 avec notifications Telegram temps reel sur trades et alertes
- > **Optimisation performances Python** : profiling memoire production (tracemalloc, objgraph, memory_profiler), detection et correction de memory leaks sur services long-running, tuning garbage collector (gc.set_threshold, gc.freeze), analyse flamegraph avec py-spy pour identifier les hotspots CPU, optimisation requetes SQL (EXPLAIN ANALYZE, index, connection pooling) -- **empreinte memoire -40%, temps reponse API /3, zero OOM en production**
- > **Infrastructure & CI/CD** : mise en place Docker Compose multi-services avec healthchecks, reverse proxy Nginx avec SSL Let's Encrypt auto-renew, CI/CD GitHub Actions complet (linting flake8/black, tests pytest, build Docker multi-stage, deploy automatise via SSH), monitoring applicatif avec logs structures JSON, alertes Telegram sur erreurs critiques, backups PostgreSQL pg_dump quotidiens avec rotation 7 jours
- > **Mentoring & leadership** : encadrement de 3 developpeurs juniors (onboarding, pair programming, montee en competence), mise en place de revues de code systematiques sur chaque PR, redaction de guidelines et bonnes pratiques (conventions, architecture, securite), formation interne sur RAG et prompt engineering (3 sessions), gestion de projet Agile (sprints 2 semaines, daily standups, retrospectives, backlog Notion)

Administrateur Reseaux & Systemes

ESN | Val-de-Marne

2016

- > Automatisation et securisation des infrastructures IT, administration reseaux, deploiement serveurs
- > Depannage serveurs et peripheriques, recherche de pannes reseau, support technique niveau 2

PROJETS TECHNIQUES

- > **Chatbot RAG multi-sources** : chunking semantique adaptatif (recursive + semantic splitting), reranking hybride Cohere RRF, Qdrant vector store, streaming WebSocket, evaluation qualite RAGAS, sources citees dans les reponses
- > **Pipeline extraction documentaire IA** : OCR multi-moteurs (Tesseract, Unstructured API), traitement images automatise (rotation, skew, binarisation), classification par LLM, normalisation et structuration JSON automatique
- > **Systeme d'agents IA orchestres** : architecture multi-agents avec juge evaluateur, routing dynamique par intention, prompt engineering avance, chaines multi-etapes avec memoire contextuelle persistante
- > **Systeme embarque stabilisation camera** : firmware C sur STM32, filtre de Kalman, controle PID temps reel, communication serie UART, calibration IMU

FORMATION

Diplome d'Ingenieur -- Systemes d'Information & Cybersecurite

ESIEA, Paris / Ivry-sur-Seine | 2020-2024 | IA, Cloud, Reseaux, Securite, Systemes embarques

DUT Genie Electrique et Informatique Industrielle (GEII)

IUT d'Evry | 2018-2020 | Electronique, informatique industrielle, design carte PCB

Baccalaureat STI2D

Lycee Christophe Colomb, Sucy-en-Brie | 2018

COMPETENCES TECHNIQUES

PythonTypeScriptFastAPIAngularReactLaravelNode.jsPostgreSQLMongoDBRedisDockerDocker Compose

LangchainLlamaIndexOpenAIMistralClaudeRAGQdrantChromaDBPineconeCohereDSPyCrewAIRAGAS

OCRTesseractUnstructuredScalewayNginxGitHub ActionsGitLab CipytestXGBoostCeleryGraphQLWebSocket

SSEGitLinuxAgile/ScrumJiraNotionFigmaPostmanSwaggerVS CodeCursor