Abdelatif TOUATI - Ingénieur IA · IA générative locale & RAG

Lyon, France | abdel.touati@gmail.com | +33 6 51 38 55 04
2501project.dev | github.com/2501Pr0ject | huggingface.co/The2501Pr0ject

Profil

Ingénieur IA (3+ ans en développement logiciel, dont 2+ en IA), spécialisé en IA générative locale (local-first) et on-premise pour les environnements à données sensibles. Approche deterministic-first : règles, parsing structuré et schémas typés traitent tout ce qui est prévisible, et les LLM n'interviennent que là où ils apportent une vraie valeur. Dernière réalisation : un assistant IA auto-hébergé pour une plateforme de gestion de dossiers juridiques (RAG hybride, agents d'analyse documentaire), sur une infrastructure 100 % CPU, sans aucune API d'IA externe. Créateur et mainteneur de RAGnarok-AI, framework open source d'évaluation RAG local-first.

Expérience professionnelle

Ingénieur IA (freelance) · **Client confidentiel, éditeur de logiciel legal tech** *Feb 2026 – Sep 2026*
Livraison de bout en bout : conception et mise en production d'un assistant IA pour une plateforme de gestion de dossiers juridiques, auto-hébergé sur un serveur 16 cœurs sans GPU (aucune API d'IA externe) ; mise en place de la CI/CD et des sprints Scrum dans une équipe de 5 personnes (tech lead, 2 développeurs, 2 chefs de projet), avec conformité RGPD et accessibilité RGAA.
Optimisation de la latence : réduction de 35 % de la latence de bout en bout des réponses RAG (~51 s à ~33 s) sur modèles locaux quantisés (Ollama / Qwen), grâce au filtrage déterministe, à des réponses adaptées au contexte et à la parallélisation des appels à l'API métier (asyncio).
Architecture & agents : développement d'une passerelle FastAPI (back-end Symfony, recherche hybride Qdrant + Elasticsearch, Redis, Celery, streaming SSE) avec pré-routeur déterministe et repli LLM vers des agents d'analyse documentaire, de statistiques multi-tenant et d'assistance, ainsi que d'un pipeline documentaire asynchrone résilient (chunking structurel, synthèse map-reduce en schémas typés, sources citées à la page).
Routage piloté par l'évaluation : création d'un golden set de 30 questions (84 contrôles automatisés : routage, format JSON, citations, langue, absence de crash) et benchmark de 3 LLM locaux ; le pré-routeur déterministe a porté la précision du routage de 52 % à 85 % et le taux de réussite global de 77 % à 93 %.
Robustesse & sécurité : plus de 1 000 tests automatisés (87 % de couverture), garde-fous en amont du LLM (hors périmètre, injection de prompt), contrôle d'accès par rôle (RBAC), circuit breakers, observabilité Prometheus / Grafana (13 métriques, 4 tableaux de bord).

Créateur & mainteneur, RAGnarok-AI (open source) *Jan 2026 – Aujourd'hui*
Moteur d'évaluation local-first : conception d'un framework d'évaluation de RAG et d'agents 100 % local (pip install ragnarok-ai) : métriques de retrieval (Precision@K, Recall@K, MRR, NDCG), LLM-as-a-Judge local (Prometheus 2) et outils de calibration sur annotations humaines (kappa de Cohen, intervalles bootstrap).
CI/CD & monitoring de production : CLI avec seuils de qualité bloquants (--fail-under) et GitHub Action dédiée, détection de dérive (PSI), A/B testing en production, diagnostic des causes d'échec, chart Helm Kubernetes et plus de 20 intégrations (Qdrant, pgvector, LangChain, LangGraph, LlamaIndex, Ollama, OpenAI).

Intervenant Data Science & Machine Learning · Sup de Pub (OMNES Education) *Oct 2026*
Conception et animation d'un module pratique de machine learning en Python (arbres de décision, évaluation de modèles, segmentation client) pour des étudiants non développeurs de M2 Data Marketing & IA.

Teaching Assistant – Data Science (alternance) · Wild Code School *Sep 2024 – Jan 2026*
Data engineering : contribution à la création d'Iliad, la plateforme de données interne de l'école, unifiant 4 sources hétérogènes (Prefect, dbt, MongoDB, DuckDB / MotherDuck).
Enseignement : cours quotidiens, sessions de live coding, revues d'architecture de projets et examens blancs pour des promotions Data Analytics et Data Science (Python, SQL, machine learning).

Recherche & compétitions

Fiabilité des LLM-as-a-Judge · RAGnarok-AI / RAGnarok Judge (Open Source) *Oct 2026 – Aujourd'hui*
Étude de calibration des juges : étude d'annotation humaine mesurant l'accord entre des juges LLM locaux (Prometheus 2, Mistral 7B, Llama 3.1 8B) et des annotations expertes ; chaque désaccord classé selon un arbre

de décision fixe, validé par reclassification à l'aveugle. Découverte d'un bug de parsing silencieux qui ramenait chaque score à 3/5, corrigé dans la v1.12.2. Résultats à paraître.
Benchmark de détection d'hallucinations : intégration de HHEM et MiniCheck comme juges locaux et création d'un banc de comparaison appariée sur RAGTruth (bootstrap groupé, test exact de McNemar) : kappa de 0,47 pour HHEM contre 0,07 pour Prometheus 2 7B sur 200 phrases (p = 8e-10).
Entraînement pré-enregistré de petits juges : protocole d'évaluation figé et publié avant tout entraînement (cibles, plan statistique, provenance et empreintes des données), puis lancement du fine-tuning LoRA de modèles de moins d'un milliard de paramètres (Qwen3-0.6B, Pleias-RAG-350M) sur MLX, pour la détection d'hallucinations au niveau de la phrase avec un budget mémoire de 3 Go ; résultats publiés sur Hugging Face.

Participant invité, challenge Drawerrr « AI Safety & Control » · Ullr-AI *Nov 2026*
Passerelle d'autorisation pour agents IA (Ullr-AI) : développement d'une couche de contrôle déterministe indépendante qui empêche les actions irréversibles d'agents à outils : politique par niveau de risque, approbations humaines signées Ed25519 et liées aux arguments exacts de l'action, protection contre l'expiration et le rejeu, journal d'audit infalsifiable (FastAPI, Pydantic v2, SQLite, Docker Compose, Ollama).
Évaluation pré-enregistrée en trois couches : conception et exécution de tests de conformité par menace, d'un attaquant scripté dans le conteneur de l'agent (0 échec sur 170 contrôles adverses) et d'un benchmark d'agents LLM locaux avec un jeu de test caché dont l'empreinte SHA-256 a été publiée avant l'exécution. Aucun contournement de la politique dans les conditions protégées ; faux refus conformes aux prévisions.

Compétences Techniques

IA générative & RAG : RAG hybride (dense + lexical), workflows agentiques, pré-routage déterministe, chunking structurel, synthèse map-reduce, sorties structurées (décodage contraint par JSON Schema), prompt engineering, garde-fous (injection de prompt, filtrage hors périmètre), inférence LLM locale (Ollama, MLX, modèles Qwen quantisés).
Évaluation RAG & agents : RAGnarok-AI, LLM-as-a-Judge (Prometheus 2), détection d'hallucinations (HHEM, MiniCheck), métriques de retrieval (Precision@K, Recall@K, MRR, NDCG), calibration de juges (kappa de Cohen, IC bootstrap), benchmark sur golden set, détection de dérive (PSI).
Frameworks IA & ML : PyTorch, fine-tuning LoRA, Hugging Face Transformers, scikit-learn, pandas, NumPy, NLP, séries temporelles, bandits contextuels (LinUCB) ; intégrations LangChain, LangGraph, LlamaIndex.
Bases vectorielles & données : Qdrant, Elasticsearch, pgvector, ChromaDB, FAISS, Redis, PostgreSQL, MongoDB, DuckDB / MotherDuck.
Back-end & data engineering : Python (asyncio, Pydantic), SQL, FastAPI, Celery, API REST, SSE (Server-Sent Events), Prefect, dbt.
DevOps, MLOps & qualité : Docker, Kubernetes (Helm), GitHub Actions (CI/CD), GCP, Prometheus, Grafana, pytest, mypy, ruff, pre-commit, packaging PyPI, Git, Linux.
Méthodes & conformité : Agile / Scrum, RGPD, accessibilité (RGAA).

Formation

ASCENCIA - Collège de Paris *France | Sep 2024 – Jan 2026*
Titre RNCP niveau 7 « Chef de projet en intelligence artificielle / Ingénieur IA » (RNCP 36129)
Alternance à la Wild Code School · Mémoire : conception de Space Coders Agency, une plateforme EdTech multi-agents.
Enseignements : machine learning, deep learning, NLP, LLM & RAG, IA agentique, séries temporelles, MLOps, gestion de projet IA, RGPD & éthique de l'IA.

WILD CODE SCHOOL *France | Mar 2023 – Aug 2023*
Titre Data Analyst (niveau 6) · Formation intensive de 5 mois

UNIVERSITÉ LUMIÈRE LYON 2 - Master 2 Aménagement du territoire · Licences de Géographie et d'Histoire
Mémoire : risques industriels et naturels dans la vallée de la chimie lyonnaise (SIG, analyse spatiale).

Langues

Français (langue maternelle) · **Anglais** (professionnel et technique) · **Espagnol** (intermédiaire) · **Arabe** (oral)