Abdelatif TOUATI - AI Engineer · Local-First Generative AI & RAG

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

Professional Summary

AI Engineer (3+ years in software, 2+ in AI) specializing in local-first, on-premise generative AI for data-sensitive environments. Deterministic-first by design: rules, structured parsing, and typed schemas handle everything predictable, with LLMs invoked only where they add real value. Most recently shipped a self-hosted AI assistant for a legal case management platform (hybrid RAG, document-analysis agents) on CPU-only infrastructure with no external AI APIs. Creator and maintainer of RAGnarok-AI, an open-source, local-first RAG evaluation framework on PyPI.

Professional Experience

AI Engineer (Freelance) · Confidential Client, Legal-Tech Software Vendor *Feb 2026 – Sep 2026*
End-to-End Delivery: Architected and shipped an AI assistant for a legal case management platform, self-hosted on a 16-core CPU-only server (no GPU, no external AI APIs); set up CI/CD and Scrum delivery in a 5-person team (tech lead, 2 devs, 2 PMs), including GDPR and RGAA accessibility compliance.
Latency Optimization: Cut end-to-end RAG response latency by 35% (~51s to ~33s) on quantized local models (Ollama / Qwen) through deterministic filtering, context-aware responses, and parallelized business API calls (asyncio).
Architecture & Agents: Engineered a FastAPI gateway (Symfony backend, Qdrant + Elasticsearch hybrid search, Redis, Celery, SSE streaming) with a deterministic pre-router and LLM fallback across document-analysis, multi-tenant statistics, and assistant agents, plus a fault-tolerant async document pipeline (structure-aware chunking, map-reduce summarization into typed schemas with page-level citations).
Evaluation-Driven Routing: Built a 30-question golden set (84 automated checks: routing, JSON format, citations, language, crash-free runs) and benchmarked 3 local LLMs; the deterministic pre-router raised routing accuracy from 52% to 85% and the overall pass rate from 77% to 93%.
Production Hardening & Security: 1,000+ automated tests (87% coverage), pre-LLM guardrails (out-of-scope, prompt injection), RBAC, circuit breakers, and Prometheus / Grafana observability (13 metrics, 4 dashboards).

Creator & Maintainer, RAGnarok-AI (Open Source) *Jan 2026 – Present*
Local-First Evaluation Engine: Architected a 100% local RAG and agent evaluation framework (pip install ragnarok-ai): retrieval metrics (Precision@K, Recall@K, MRR, NDCG), a local LLM-as-a-Judge (Prometheus 2), and calibration tooling against human labels (Cohen's kappa, bootstrap CIs).
CI/CD & Production Monitoring: CLI with blocking quality gates (--fail-under) and a dedicated GitHub Action, PSI-based drift detection, live A/B testing, root-cause failure diagnosis, a Kubernetes Helm chart, and 20+ integrations (Qdrant, pgvector, LangChain, LangGraph, LlamaIndex, Ollama, OpenAI).

Data Science & Machine Learning Lecturer · Sup de Pub (OMNES Education) *Oct 2026*
Designed and delivered a hands-on Python ML course (decision trees, model evaluation, customer segmentation) for non-programmer Master's students in Data Marketing & AI, with a final group project.

Teaching Assistant – Data Science (Apprenticeship) · Wild Code School *Sep 2024 – Jan 2026*
Data Engineering: Contributed to building Iliad, the school's internal data platform, unifying 4 heterogeneous data sources with Prefect, dbt, MongoDB, and DuckDB / MotherDuck.
Teaching: Delivered daily lectures, live-coding sessions, project architecture reviews, and mock exams to Data Analytics & Data Science cohorts in Python, SQL, and Machine Learning.

Research Experience & Competitions

LLM-as-a-Judge Reliability · RAGnarok-AI / RAGnarok Judge (Open Source) *Oct 2026 – Present*
Judge Calibration Study: Ran a human-annotation study measuring agreement between local LLM judges (Prometheus 2, Mistral 7B, Llama 3.1 8B) and expert labels, classifying every disagreement with a fixed decision tree validated by blind re-classification. Surfaced a silent parsing bug that defaulted every judge score to 3/5, fixed in v1.12.2. Results forthcoming.

Hallucination-Detection Benchmark: Integrated HHEM and MiniCheck as local judges and built a paired benchmark harness on RAGTruth (grouped bootstrap, exact McNemar): HHEM kappa 0.47 vs. 0.07 for Prometheus 2 7B on 200 sentences (p = 8e-10).
Pre-Registered Small-Judge Training: Froze and published the evaluation protocol prior to training (targets, statistical plan, data provenance, and hashes), then initiated LoRA fine-tuning of sub-1B models (Qwen3-0.6B, Pleias-RAG-350M) on MLX for sentence-level hallucination detection under a 3 GB memory budget, with results committed for publication on Hugging Face regardless of outcome.

Invited Participant, Drawerrr AI Safety & Control Challenge · Ullr-AI *Nov 2026*
Authorization Gateway for AI Agents (Ullr-AI): Built an independent deterministic enforcement layer preventing irreversible actions by tool-calling agents: risk-tier policy, Ed25519-signed approvals bound to exact action arguments, expiry and replay protection, and a tamper-evident audit chain (FastAPI, Pydantic v2, SQLite, Docker Compose, Ollama).
Pre-Registered Three-Layer Evaluation: Designed and ran per-threat conformance tests, a scripted in-container attacker (0 failures across 170 adversarial checks), and a local LLM-agent benchmark with a hidden test set hashed (SHA-256) before execution. Zero policy bypasses in protected conditions; false denials matched predictions.

Technical Skills

Generative AI & RAG: Hybrid RAG (Dense + Lexical), Agentic Workflows, Deterministic Pre-Routing, Structure-Aware Chunking, Map-Reduce Summarization, Structured Outputs (JSON Schema Constrained Decoding), Prompt Engineering, Guardrails (Prompt Injection, Out-of-Scope Filtering), Local LLM Inference (Ollama, MLX, Quantized Qwen Models).
RAG & Agent Evaluation: RAGnarok-AI, LLM-as-a-Judge (Prometheus 2), Hallucination Detection (HHEM, MiniCheck), Retrieval Metrics (Precision@K, Recall@K, MRR, NDCG), Judge Calibration (Cohen's kappa, Bootstrap CI), Golden Set Benchmarking, Drift Detection (PSI).
AI Frameworks & ML: PyTorch, LoRA Fine-Tuning, Hugging Face Transformers, scikit-learn, pandas, NumPy, NLP, Time Series, Contextual Bandits (LinUCB); integrations with LangChain, LangGraph, LlamaIndex.
Vector & Data Stores: Qdrant, Elasticsearch, pgvector, ChromaDB, FAISS, Redis, PostgreSQL, MongoDB, DuckDB / MotherDuck.
Backend & Data Engineering: Python (asyncio, Pydantic), SQL, FastAPI, Celery, REST APIs, SSE (Server-Sent Events), Prefect, dbt.
DevOps, MLOps & Quality: Docker, Kubernetes (Helm), GitHub Actions (CI/CD), GCP, Prometheus, Grafana, pytest, mypy, ruff, pre-commit, PyPI Packaging, Git, Linux.
Methods & Compliance: Agile / Scrum, GDPR, Accessibility (RGAA).

Education

ASCENCIA - Collège de Paris *France | Sep 2024 – Jan 2026*
Master's-Level Degree in AI Engineering & AI Project Management (EQF Level 7, RNCP 36129)

- Apprenticeship at Wild Code School · Thesis: Design of Space Coders Agency, a multi-agent EdTech platform.
- Coursework: Machine Learning, Deep Learning, NLP, LLMs & RAG, Agentic AI, Time Series, MLOps, AI Project Management, GDPR & AI Ethics.

Wild Code School *France | Mar 2023 – Aug 2023*
Bachelor's-Level Degree in Data Analytics (EQF Level 6) · 5-month intensive bootcamp

Université Lumière Lyon 2 - Master's Degree (M2) in Urban and Regional Planning · Dual Bachelor's Degrees in Geography and History

- **Thesis:** Industrial and Natural Risk Assessment in Lyon's Chemical Valley (GIS, spatial analysis, statistics).

Languages

French (Native) · **English** (Professional & Technical Proficiency) · **Spanish** (Intermediate) · **Arabic** (Conversational)