



MAIA BAZERJI

AI Engineer | RAG, LLM & evaluation in production

Seeking a permanent role (CDI) - Available immediately

AI Engineer (MSc, EQF level 7). I put RAG systems into production and measure that they work: LLM-as-judge evaluation, golden datasets, regression detection, CI (pytest, ruff, mypy). Fine-tuning and French NLP (CamemBERT). Stack: Qdrant, FastAPI, Docker, GCP Cloud Run, PyTorch. French & English: bilingual.

CONTACT

☎ +33 6 12 39 74 50
✉ bazerjimaia@gmail.com
📍 Malakoff, France
🌐 maiabazerji.com
🐙 github.com/maiabazerji
in linkedin.com/in/maiabazerji
M medium.com/@maiabazerji

EDUCATION

MSc Information Technology

Epitech Paris | 2024-2026 | Graduated September 2026

AI & Data specialisation, EQF level 7

MSc 1 Information Systems

Lebanese University | 2023-2024

BSc Computer Science

Lebanese University | 2020-2023

CERTIFICATIONS

Deep Learning Specialization

Coursera | 2024

Evaluating and Debugging Generative AI

DeepLearning.AI | 2026

Claude with Google Cloud Vertex AI

Anthropic | 2026

Introduction to Subagents

Anthropic | 2026

LANGUAGES

French & English : bilingual

SKILLS

Generative AI, LLM & Agents

RAG pipelines, RAGAS evaluation, LLM-as-judge, golden datasets, regression detection, cross-encoder reranking, BM25, fine-tuning (LoRA/QLoRA), LangChain, LlamaIndex, LangGraph, Qdrant, embeddings, pgvector, agentic workflows, function calling, structured output (Pydantic), OpenAI API, Anthropic Claude API, Hugging Face, Langfuse

MLOps & Production

FastAPI, Docker, Kubernetes, CI/CD, GitHub Actions, pytest, ruff, mypy, MLflow, PostgreSQL, GCP, model monitoring, observability

Machine Learning & Vision

Scikit-learn, XGBoost, PyTorch, TensorFlow, feature engineering, time series, OpenCV, Tesseract, OCR, Computer Vision

Data Engineering & Dev

Python, JavaScript, SQL, pandas, NumPy, Git, Spark, Streamlit, Flask, Weights & Biases, React, Flutter, Electron, Firebase

ACTIVITIES & AWARDS

Epitech Ambassador (2024-2026)

Hult Prize committee, Lebanese Univ. (2022-2023)

PROFESSIONAL EXPERIENCE

BazarSoft

Apr 2025 - Present

AI & Full-Stack Engineer

Remote freelance (France) • Client in Dubai, UAE

- Customer-support assistant (RAG + agent) grounded in product documentation: ingestion, chunking, embeddings, search, tool calls into internal APIs, human escalation when confidence is low. Bilingual EN/AR, RTL-correct. *FastAPI, PostgreSQL, Claude API, Docker.*
- 44-case evaluation suite** (22 EN / 22 AR): grounded answers, mandatory tool calls, out-of-scope refusals, escalation, with fabrication checks (a price quoted without a tool call, a promised refund, an invented date).
- Full-stack delivery of the product suite**: self-serve licensing platform (React, Stripe), marketing site, B2B mobile app for field sales (Flutter, offline-first) and a display app; RTL/LTR multilingual, deployed with Docker/Caddy. Defined the AI product roadmap.

Kabeen

Jan 2025 - Apr 2026

Data & AI Engineer (apprenticeship)

Saint-Cloud, France

- Architected and deployed a hybrid-search RAG/ETL pipeline as sole engineer: **-80% document processing time** and catalogue scaling to **100,000+ entries**, unlocking real-time search and classification.
- Industrialised scraping pipelines for continuous metadata ingestion: eliminated manual updates and enabled daily refreshes of 500+ entries with no performance degradation.
- Built a data quality and anomaly detection system, reducing errors in indexed data and enforcing validation at the ingestion layer (FastAPI, Qdrant).
- Took LLM modules to production with real-time frontend synchronisation and a zero-downtime deployment strategy.

Buy Any Insurance

Mar - Sep 2024

Software Engineer (Computer Vision / OCR)

Dubai, UAE

- Built a document extraction engine (OCR + computer vision) automating client onboarding: **+15% extraction precision** through image preprocessing (OpenCV) and CNN fine-tuning, saving ~100 hours of manual review per month. Inference pipeline deployed to production (*PyTorch, TensorFlow, FastAPI, Firebase*).

PROJECTS

• RAG System

Three retrieval strategies compared over one corpus (classic / graph / agentic). Evaluation harness: golden dataset, LLM-as-judge, regression detection. GitHub Actions CI (ruff, mypy, pytest). *FastAPI, Qdrant, Langfuse, Docker.*

• E-invoicing AI

QR PDP-compliant invoicing: LLM-assisted data entry and structured-output validation. *FastAPI, Pydantic, Electron, Claude API.*

• NER Travel Order

Structured extraction from French travel orders (BIO tagging, annotated corpus). spaCy **F1 0.84** (~5 ms, 15 MB) vs CamemBERT **F1 0.89** (~50 ms, 440 MB); **spaCy chosen for production** on the latency/size trade-off. *Hugging Face, spaCy, Flask, W&B.*

• Moderation Service

On-device image moderation (messaging app): MobileNetV3Small, macro-F1 93.1% over 9,714 images. TFLite export in three quantised variants (1.24 / 1.96 / 3.89 MB) targeting edge, mobile and server inference. *TensorFlow, Keras, TFLite, Docker.*