Ben Alla Ismail

Data & AI Engineer · LLM Systems · RAG & Retrieval · Data Pipelines

Paris, France | +33 7 73 12 05 88 | ismailbenalla52@gmail.com | linkedin.com/in/ismail-ben-alla-bai | github.com/benisalla

SUMMARY

AI Engineer with 2+ years delivering production systems across agentic AI, multilingual RAG, real-time deep learning, and data engineering. Achieved sub-80 ms voice conversion and sub-350 ms retrieval across 100k+ legal documents; experienced in LLM orchestration, evaluation, and cloud deployment.

TECHNICAL SKILLS

AI & LLMs: RAG, prompt engineering, fine-tuning (LoRA, QLoRA), embeddings, semantic and hybrid retrieval, LLM reranking and evaluation, agentic and multi-agent systems, structured outputs, guardrails

Machine Learning & Deep Learning: Supervised learning, feature engineering, CNNs, Transformers, generative and multimodal models, model optimization, PyTorch, TensorFlow, scikit-learn, ONNX

Data Engineering: ETL/ELT, data modeling, batch and stream processing, Python, SQL, Airflow, BigQuery, PySpark, Kafka, PostgreSQL

MLOps & DevOps: CI/CD, containerization, infrastructure as code, model tracking and monitoring, Docker, Kubernetes, GitHub Actions, Terraform, MLflow, AWS, GCP, Azure

Web Applications: REST API development, FastAPI, React, Git/GitHub

EDUCATION

Centrale Lyon, Lyon **Sept. 2024 – Oct. 2025**

Master's in Artificial Intelligence (Deep Learning, NLP, MLOps)

National School of Applied Sciences, Fez **Sept. 2019 – Aug. 2024**

Computer Science Engineering Degree

EXPERIENCE

Dailymotion · Ray AI Platform **June 2026 – Present**

Data & AI Engineer

- Built Airflow pipelines enriching web, app, and CTV sitelists with Gemini and Google Search grounding, generating descriptions, IAB categories, and language metadata.
- Developed BigQuery-to-Qdrant ingestion and multilingual retrieval for Ray using English-normalized context, semantic search, and LLM reranking.
- Designed web scraping workflows (HTTPX, Playwright, Trafilatura, Selectolax, Extruct) and LangSmith evaluation with human-reviewed golden datasets and ranking metrics.
- Developed an LLM-powered Alert Agent for automated incident investigation, integrating Gemini with MCP clients and dedicated Airflow, GCP, and BigQuery MCP servers to diagnose pipeline, Dataflow, Pub/Sub, and SQL issues from Slack alerts.
- Led a data lineage technology assessment across BigQuery, Airflow, dbt, Pub/Sub, Dataflow, GCS, and BI, comparing OSS solutions including DataHub, OpenMetadata, and Marquez, and prototyped OpenLineage-based runtime lineage integration.

Monce.ai · LLM-Powered Data Extraction **September 2025 – June 2026**

AI Engineer

- Designed a human-in-the-loop feedback system for continuous agent improvement, increasing document extraction accuracy by approximately 18%.
- Built a modular agent evaluation framework covering 10+ metrics, including precision, recall, hallucination rate, and latency.
- Refactored the data extraction pipeline using LangChain, LlamaIndex, and asynchronous processing, reducing average processing time by approximately 35%.

SteelSeries · Real-Time Deep Learning Voice Conversion **April 2025 – September 2025**

AI Engineer

- Built a real-time voice conversion system using PyTorch and ONNX, achieving end-to-end latency below 80 ms.
- Developed a multi-voice blending module based on the LLVC architecture, supporting simultaneous voice-profile mixing in real time.
- Improved MOS by 0.3 points over the StreamVC baseline through spectrogram smoothing, Griffin-Lim phase refinement, and targeted fine-tuning.

Juridia · RAG-Powered Agentic Legal AI **January 2024 – July 2024**

AI Engineer

- Designed and deployed a production RAG system for Arabic and French legal documents using FastAPI, LlamaIndex, PostgreSQL, and Qdrant, reducing average query latency to below 350 ms.
- Improved retrieval precision by 18–22% using cross-encoder reranking, hybrid search (BM25 + dense vectors), and multilingual embeddings.
- Built a full-stack search interface using React, supporting keyword, semantic, and metadata search across 100k+ legal documents.

PROJECTS

Asliia · Document Extraction Platform **2024 – Present**

Independent AI Engineer

- Built a configurable multi-agent extraction platform with custom schemas, API/email/file workflows, FastAPI, PostgreSQL, React, Terraform, and AWS/Azure deployment.

Qlaver · AI Email Management Agent **2024 – Present**

Independent AI Engineer

- Built an AI email agent generating personalized audio summaries with ElevenLabs; delivered using FastAPI, PostgreSQL, React, AWS, Terraform, and GitHub Actions.

LANGUAGES

English & French