

# Ben Alla Ismail

AI Engineer | Agentic AI | RAG & LLM Evaluation | GCP / BigQuery

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

## PROFILE

AI Engineer with 2+ years of experience building and deploying production AI systems across agentic workflows, multilingual RAG, retrieval quality, LLM evaluation, and data platforms. Strong Python and SQL background with hands-on experience in GCP, BigQuery, Airflow, LangChain, LlamaIndex, MCP clients, reranking, golden datasets, observability, and cost- and latency-aware delivery. Fluent in French and English.

## CORE CAPABILITIES

|                  |                                                                                                                          |
|------------------|--------------------------------------------------------------------------------------------------------------------------|
| Agentic AI       | Tool calling, MCP clients, multi-agent workflows, structured outputs, guardrails, human-in-the-loop feedback             |
| RAG & Retrieval  | Hybrid search, multilingual embeddings, metadata search, cross-encoder reranking, grounding, citation-oriented retrieval |
| Quality & LLMOps | Golden datasets, offline/online evaluation, precision/recall, hallucination and latency metrics, LangSmith, MLflow       |
| Cloud & Data     | GCP, BigQuery, Airflow, Dataflow, Pub/Sub, GCS, Qdrant, PostgreSQL, Docker, Terraform, CI/CD                             |

## PROFESSIONAL EXPERIENCE

**Data & AI Engineer, Dailymotion - Ray AI Platform** Jun 2026 - Present

- Built Airflow pipelines using Gemini and Google Search grounding to enrich web, app, and CTV sitelists with descriptions, IAB categories, and language metadata.
- Developed BigQuery-to-Qdrant ingestion and multilingual retrieval with English-normalized context, semantic search, and LLM reranking.
- Designed web-scraping workflows and LangSmith evaluation using human-reviewed golden datasets and ranking metrics.
- Developed an LLM-powered Alert Agent integrating Gemini with MCP clients and dedicated Airflow, GCP, and BigQuery MCP servers to investigate pipeline, Dataflow, Pub/Sub, and SQL incidents from Slack alerts.
- Led a data-lineage technology assessment across BigQuery, Airflow, dbt, Pub/Sub, Dataflow, GCS, and BI; prototyped OpenLineage runtime integration.

**AI Engineer, Monce.ai - LLM-Powered Data Extraction** Sep 2025 - Jun 2026

- 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 LangChain/LlamaIndex extraction workflows with asynchronous processing, reducing average processing time by approximately 35%.

**AI Engineer, SteelSeries - Real-Time Deep Learning Voice Conversion** Apr 2025 - Sep 2025

- Built a PyTorch and ONNX voice-conversion system with end-to-end latency below 80 ms.
- Developed real-time multi-voice blending based on the LLVC architecture and improved MOS by 0.3 points over the StreamVC baseline.

**AI Engineer, Juridia - RAG-Powered Agentic Legal AI** Jan 2024 - Jul 2024

- Designed and deployed a production Arabic/French RAG system with FastAPI, LlamaIndex, PostgreSQL, and Qdrant, achieving average query latency below 350 ms.
- Improved retrieval precision by 18-22% through cross-encoder reranking, BM25+dense hybrid search, and multilingual embeddings.
- Built a React interface supporting keyword, semantic, and metadata search across 100k+ legal documents.

## SELECTED PROJECTS

**Independent AI Engineer, Qlaver - AI Email Management Agent** 2024 - Present

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

## EDUCATION

**Master's in Artificial Intelligence** Centrale Lyon, Lyon | Sep 2024 - Oct 2025

**Computer Science Engineering Degree** National School of Applied Sciences, Fez | Sep 2019 - Aug 2024

## TECHNICAL TOOLKIT

Python, SQL, FastAPI, React, LangChain, LlamaIndex, PyTorch, TensorFlow, scikit-learn, ONNX, Airflow, BigQuery, PySpark, Kafka, PostgreSQL, Qdrant, Docker, Kubernetes, Terraform, GitHub Actions, MLflow, LangSmith, AWS, GCP, Azure.

Languages: French (fluent) | English (professional fluent)