Skills Portfolio — Mounaim BELKHOUMALI

AI Engineer / Data Scientist — LLMs · Generative AI · RAG · MLOps

pro.belkhoumali.mounaim@gmail.com · +33 7 48 51 51 01 · Paris, France · LinkedIn: BELKHOUMALI Mounaim

1. Summary

Versatile AI Engineer, graduate of École polytechnique (MSc Artificial Intelligence, 2018–2023) and Sorbonne University (MSc Data Science, 2023–2024), with 3+ years of hands-on experience delivering production Generative AI, Machine Learning and Computer Vision systems.

Worked with large groups (Publicis Sapient / Publicis Group), mid-size companies (EPSILON France) and fast-growing SMBs / start-ups (MapBrain, Zaivio, Biopharm, Jumia). End-to-end delivery capability: business framing, state-of-the-art research, prototyping, industrialization (Docker, Kubernetes, CI/CD) and frontend (React / Next.js).

Core focus areas: LLMs (OpenAI, Azure OpenAI, open-source), hybrid RAG pipelines, autonomous agents (LangGraph), real-time voice chat (Azure Realtime API), multimodal Computer Vision, MLOps.

2. Technical Skills

2.1 Artificial Intelligence & LLMs

- **Proprietary models:** OpenAI API (GPT-4, GPT-4o, GPT-4o-vision, GPT-4.1-mini/nano), Azure OpenAI, Whisper (Speech-to-Text), gpt-4o-mini-tts (Text-to-Speech), gpt-4o-realtime-preview (bidirectional voice chat), DALL·E, FLUX 1.1 Pro (Azure AI Foundry).
- **Open-source models:** Hugging Face Transformers, open LLMs (Mistral, Llama, etc.), text-embedding-3-large embeddings (3072 dim).
- **Orchestration:** LangChain 0.3, LangGraph (agent state machine), Pydantic.
- **RAG patterns:** vanilla RAG, hybrid RAG (semantic + BM25), HyDE, RAG Fusion, reranking (rank-bm25), RAGAS evaluation.
- **Prompt engineering:** chain-of-thought, few-shot, role-based prompting, content-specific prompts (quiz, flashcards, podcast, mind map).
- **Voice AI:** PCM16 24 kHz streaming, barge-in, real-time tool-calling via Azure Realtime API.

2.2 Machine Learning / Deep Learning

- Scikit-Learn (regression, classification, clustering, PCA, model selection).
- TensorFlow / Keras (deep networks, time series).
- Computer Vision: OpenCV, YOLOv7.
- Evaluation: cross-validation, metrics (accuracy, F1, precision/recall), hyperparameter benchmarking.

2.3 Backend & Data Engineering

- **APIs:** FastAPI, Flask, Flask-RESTX (Swagger), Flask-SocketIO (WebSocket).
- **Async:** SQLAlchemy 2.0 async, asyncio, AnyIO, Gunicorn (workers/threads).
- **Relational databases:** PostgreSQL, MySQL, Alembic (idempotent migrations).
- **Vector databases:** Milvus / Zilliz Cloud (IVF_FLAT index, L2 distance, 3072 dimensions).
- **Cache & messaging:** Redis (distributed TTL cache, async job tracking across workers).
- **Big Data:** Spark, Hadoop.
- **Web scraping & ingestion:** Selenium, BeautifulSoup, yt-dlp, PyPDF2, python-pptx, moviepy, pydub, pytesseract, pdf2image (OCR fallback).

2.4 DevOps / Cloud

- **Containers:** Docker (Python 3.10-slim + ffmpeg + poppler + tesseract), Docker Compose.
- **Orchestration:** Kubernetes (GCP Artifact Registry, per-pod CPU/RAM sizing).
- **Cloud:** Azure (OpenAI, Cognitive Services, AI Foundry), GCP, OVH (Object Storage), AWS (Zilliz EU Central), Render, Heroku, Vercel (Blob).
- **CI/CD & versioning:** Git, GitHub.

- **Security:** Fernet-encrypted OAuth tokens, secret management, fine-grained multi-tenant scoping.
- **Reliability:** circuit breaker, exponential retry, idempotency, zero-downtime deployments.

2.5 Frontend

- React 19, Vite, TypeScript, Tailwind CSS v4, shadcn/ui, React Router 7, Motion (animations).
- Next.js (SSR, App Router), Streamlit (data dashboards).

2.6 Methodologies & soft skills

- Agile / Scrum, code review, unit and integration testing (Pytest, httpx ASGITransport).
- Business stakeholder communication (scoping workshops, demos).
- Distributed remote teamwork (1 year with US-based Zaivio team).
- Technical pedagogy (translating AI concepts for non-technical audiences).

3. Detailed Experience (STAR method)

3.1 MapBrain (Freelance) · AI Engineer / Full-Stack

Period: December 2025 · May 2026 (6 months) · Location: Paris · Freelance · End-to-end Azure architecture

End-to-end design and development of **MapBrain**, an AI-powered educational SaaS platform that turns any source content (PDF, YouTube video, audio, slides, Word documents) into a personalized interactive learning path.

Situation

MapBrain wanted to offer an AI-powered educational platform able to turn any document into an interactive learning path (summaries, mindmaps, quizzes, flashcards, generated podcasts and real-time voice conversational tutor), on a full Azure architecture end-to-end.

Task

Continuously design and harden the backend architecture, the RAG pipeline, the real-time voice tutor and the parallel content generators.

Action

- **Multi-format ingestion pipeline** (PDF, DOCX, PPTX, MP3, MP4, YouTube) with text extraction, fallback OCR (Tesseract + Poppler) and audio transcription via Azure OpenAI Whisper, supporting content up to 200 MB.
- **Full RAG system** on Azure Cognitive Search + Milvus / Zilliz Serverless (text-embedding-3-large embeddings) with hybrid dense + BM25 retrieval and contextual reranking, enabling students to query their full course library in natural language.
- **Real-time voice tutor** via Azure OpenAI Realtime API (GPT-4o Realtime) over WebSocket / Socket.IO, with function calling, barge-in, streamed transcripts and silence detection — < 800 ms latency.
- **Parallel pedagogical content generators** (quizzes, mindmaps, flashcards, course outlines, Pretext slides) orchestrated through LangGraph, cutting the time to generate a full deck from 25 s to 10–12 s through outline-splitting + N concurrent LLM calls.
- **Bi-voice educational podcast generator** (Alex / Sara) via Azure OpenAI TTS (gpt-4o-mini-tts) with parallel chunking, voice steering and MoviePy audio fusion.
- **Azure AI Foundry (FLUX 1.1 Pro) and gpt-image-1** integration for automatic contextual illustration generation in slides and mindmaps, with smart model routing.
- **Hybrid Flask + FastAPI architecture** with Flask-RESTX (Swagger), Flask-SocketIO (live chat) and async FastAPI endpoints, deployed on Kubernetes (AKS) with multi-worker Gunicorn and Redis for cross-worker job_manager state sharing.
- **Circuit breaker**, persistent conversation managers, Redis caching, Next.js revalidation from the backend, and structured JSON logs for Azure Monitor observability.
- **Containerized deployment** via Azure Container Registry + AKS (YAML manifests, Kubernetes secrets), multi-origin CORS and blob storage via Vercel Blob for generated media.

Result

Stabilized, performant platform: full deck generated in **10–12 s (down from 25 s)**, **voice latency < 800 ms**, horizontal scalability, fluid voice UX through barge-in and RAG tool-calling, full Azure end-to-end.

Stack: Python · Flask · FastAPI · LangChain · LangGraph · Azure OpenAI (GPT-4.1-mini/nano, GPT-4o Realtime, Whisper, TTS, Embeddings) · Azure AI Foundry (FLUX 1.1 Pro, gpt-image-1) · Azure Cognitive Services (Bing Search, Translation) · Azure Kubernetes Service (AKS) · Azure Container Registry · Milvus / Zilliz · Redis · Socket.IO · WebSockets · Tesseract OCR · MoviePy · Pydub · Gunicorn · Docker · Pytest · Git.

3.2 Zaivio (Freelance, USA Remote) · Generative AI Engineer

Period: December 2024 · November 2025 (1 year) · Mode: Remote Freelance, distributed team

Project A — E-reputation monitoring platform

- **Situation:** Companies needed a unified tool to monitor their online image (web, social media, forums) and flag reputational risks early.
- **Task:** Build a real-time multi-source analysis system with automated, actionable reporting.
- **Action:** Multi-source scraping pipeline (Selenium + Tavily API), LLM-based sentiment analysis, automated email reports (SMTP) including concrete recommended actions.
- **Result:** Deployed on Heroku and used to monitor client entities' online image.

Stack: Python · LLMs · LangChain · prompt engineering · Git · OpenAI · Docker · Heroku · Tavily API · Selenium · SMTP / Email APIs.

Project B — AI assistant for medical documents

- **Situation:** Healthcare professionals must analyze long, complex medical documents and require sourced answers for clinical decision-making.
- **Task:** Deliver an assistant that answers precisely on the documents using autonomous agents and RAG.
- **Action:** Agent architecture with LangGraph (orchestration), RAG pipeline over medical documents, FastAPI backend, Next.js frontend.
- **Result:** Sourced answers improving the reliability and speed of clinical decisions.

Stack: LangGraph · LangChain · Git · OpenAI · Python (FastAPI) · Next.js.

Project C — Content automation platform

- **Situation:** Publishers and media spend massive amounts of time on trend tracking, writing and multi-channel publishing.
- **Task:** Automate the full cycle (trend detection → article + visual generation → multi-channel publishing).
- **Action:** Trend detection via scraping, text generation (OpenAI API) and visual generation (DALL·E), automated publishing to LinkedIn / X / WordPress through their APIs.
- **Result:** Massive time savings for editors; 3 products shipped to production over the engagement.

Stack: OpenAI API · Git · Selenium · DALL·E · LinkedIn / X / WordPress APIs · FastAPI · Docker.

3.3 EPSILON France — Data Scientist / AI Engineer

Period: April 2024 – September 2024 (6 months) · Paris · End-of-studies internship

End-of-studies internship at EPSILON France, including 3 months in-house (April – June 2024) and 3 months on a consulting assignment at Publicis Sapient for the Publicis Group (June – September 2024), engagement contracted through EPSILON France.

Part 1 (April – June 2024) — EPSILON France (in-house): RFP analysis RAG system

- **Situation:** Manual analysis of RFP PDFs (extracting criteria, dates, amounts) consumed significant time from sales teams and was error-prone.
- **Task:** Automate the extraction of RFP criteria and ensure analysis quality.
- **Action:** Built a RAG system with LangChain to parse and query RFP PDFs; optimized precision with HyDE and RAG Fusion; implemented RAGAS metrics to evaluate quality and reliability; systematic benchmarking to pick the best RAG parameters (chunking, top-k, embedding model, prompt).
- **Result: –50% manual workload** on RFP analysis; answer quality validated through RAGAS.

Stack: Python · LangChain · RAG Systems · Prompt Engineering · Hugging Face · open-source LLMs · Azure OpenAI · Git.

Part 2 (June – September 2024) — Consulting assignment at Publicis Sapient (Publicis Group), on behalf of EPSILON France: multimodal generative AI for ethical compliance

- **Situation:** The Publicis Group wanted a tool able to automatically assess the ethical compliance of ads (text, image, video, PDF) before broadcast, to avoid reputational and regulatory risk.
- **Task:** Design a multimodal generative AI tool for ethics review and integrate it into the validation workflow, as part of a consulting engagement contracted through EPSILON France.
- **Action:** Added new multimodal capabilities (images, videos, PDFs) to the existing tool; used OpenAI LLMs (GPT-4, GPT-4o) and open-source LLMs combined with advanced prompt engineering; comparative model benchmarking.
- **Result:** Reached **90% classification accuracy** on ad ethical compliance.

Stack: Python · LLMs (OpenAI, GPT-4, GPT-4o) · Git · LangChain · Selenium.

3.4 Biopharm — Machine Learning Engineer

Period: February 2023 (6 months)

- **Situation:** Biopharm needed to optimize the procurement of spare parts (a demand-forecasting problem).
- **Task:** Deliver an end-to-end ML solution usable by business teams.
- **Action:** Business scoping workshops; state-of-the-art review (suitable ML / time-series models); model testing and evaluation; model selection; deployment via a Tkinter UI tailored to business needs.
- **Result:** Solution deployed and used by Biopharm's business teams.

Stack: Python · Machine Learning · Deep Learning · Scikit-Learn · TensorFlow · Pandas · Tkinter.

3.5 Jumia — Machine Learning Engineer

Period: March 2022 (6 months)

- **Situation:** Seller churn on the Jumia marketplace directly hurt GMV, and no operational predictive model existed.
- **Task:** Build a seller churn prediction model and a tool to drive retention actions.
- **Action:** Exploratory analysis (PCA, clustering); classification model (Scikit-Learn); automated real-time seller-data pipeline; Streamlit dashboard for predictions and recommended actions.
- **Result:** Model reaching **80% accuracy**; dashboard rolled out to Vendor Success teams.

Stack: Python · Deep Learning · OpenCV · TensorFlow · YOLOv7 · Scikit-Learn · Streamlit.

4. Education

Period	Degree	Institution
2023 – 2024	MSc Data Science	Sorbonne University (Paris)
2018 – 2023	MEng + MSc Artificial Intelligence	École polytechnique

5. Certifications & languages

- Dataiku certification.
- English: professional working proficiency (1-year engagement with a US team).
- French: native speaker.

6. Skills × experience coverage matrix

Skill / Technology	MapBrain	Zaivio	EPSILON (incl. Publicis)	Biopharm	Jumia
Python	●	●	●	●	●
OpenAI / GPT-4 / GPT-4o	●	●	●		
Azure OpenAI	●		●		
LangChain	●	●	●		
LangGraph	●	●			
RAG (hybrid / HyDE / RAG Fusion / BM25)	●	●	●		
Voice AI / Realtime API	●				
Multimodal Computer Vision	●		●		●
Machine Learning (Scikit-Learn / TensorFlow)				●	●
Time series				●	
FastAPI / Flask	●	●			
WebSocket / SocketIO	●				
SQLAlchemy / PostgreSQL / MySQL	●				
Milvus / Vectors	●	●			
Redis	●				
Docker	●	●			
Kubernetes / AKS / GCP	●				
Render / Heroku / Vercel	●	●			
React / TypeScript / Tailwind / shadcn	●	●			
Next.js	●	●			
Streamlit / Tkinter				●	●
Git / GitHub	●	●	●	●	●

Skills portfolio updated May 2026.