

Yaghmoracen Belmir

AI Engineer

 Yaghmo |  Yaghmo |  yaghmo.belmir@gmail.com |  +33.753.52.32.39 |  Paris, France

AI Engineer with hands-on experience building and deploying ML and Generative AI systems — from GAN-based medical imaging research (MRI-to-CT translation for radiotherapy) to production-style LLM applications (RAG pipelines, agentic workflows, multimodal inference, streaming APIs). Comfortable owning the full pipeline: data preprocessing, model training, API design, containerization, and deployment. Whether working independently or as part of a larger engineering team.

WORK EXPERIENCE

- AI Engineer Intern - Medical Imaging Research (GAN-based Image Translation)** LaTIM - 2025
- Designed and trained deep learning models (PyTorch) for MRI-to-CT image translation to support radiotherapy dose planning.
 - Built scalable preprocessing pipelines for large MRI/CT datasets; evaluated with PSNR, SSIM, and structural metrics.
 - Implemented GAN architectures and improved anatomical preservation accuracy by 2.5% over baseline.

PROJECTS

- Agentic RAG Assistant with Self-Hosted LLM Inference**
- Built retrieval-augmented conversational agent with citation-grounded memory (retrieval tool vs. full-context injection chosen dynamically per model's context budget), agentic tool-calling loop for self-editing knowledge base, and pluggable LLM/embedding/TTS backends (llama.cpp, OpenAI/Anthropic adapters, Gemini, Fish Audio). Wrote hardware-probing model-fitting engine that auto-selects stacks against live VRAM/RAM (validated on 8GB GPU, up to 1M-token context) and containerized GPU-aware multi-service deployment (Docker Compose).
- Multi-Model Local AI Assistant**
- Designed offline multimodal AI assistant with local LLM/VLM inference, multimodal RAG, multilingual semantic search over ChromaDB (e5-small, CPU), int8-quantized Faster-Whisper speech recognition, dynamic context-window budgeting (69% reserved for RAG history), streaming FastAPI APIs, and Docker/CUDA deployment.
- Multilingual Sentiment Analysis System**
- End-to-end NLP pipeline combining language detection, translation, sentiment classification, and context-aware response generation using Mistral-7B.
- Medical Image Segmentation Benchmark**
- Comparative study of U-Net, DeepLabV3, and FCNN architectures for lung region segmentation under identical training conditions. Evaluated with Dice, IoU, and boundary metrics using PyTorch and MONAI.

EDUCATION

- Master 2** Optics, Image, Vision, Multimedia (SIM) — **Paris-East Créteil University** 2025
- Master 2** Visual Computing — **University of Science and Technology Houari Boumédiène** 2020

CERTIFICATIONS

IBM Data Science

SKILLS

Programming	Python, C, C++
Core AI	PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face, MONAI
Generative AI	LLMs, VLMs, RAG, AI Agents, Prompt & Context Engineering, Fine-tuning
Computer Vision	OpenCV, CNNs, GANs, Diffusion Models, Segmentation, Object Detection, OCR, Tracking
Backend & Deployment	FastAPI, REST APIs, Docker, Git, CI/CD
Data & Vector Search	Pandas, NumPy, ChromaDB, Semantic Search, ML Pipelines, Model Evaluation
Cloud	AWS, Azure, GCP

LANGUAGES

French Fluent **English** Fluent **Arabic** Native