OMAR HAJJOUJI

AI Systems Engineer focused on architecting and deploying production AI platforms powered by LLMs, RAG pipelines, and agent-based automation. Experienced in backend system design, structured output generation, real-time speech processing, and scalable containerized deployments. Proven ability to take AI products from 0→1 and integrate them into live production environments.

CONTACT DETAILS

+34 645554703, +216 52607213

omarhajjouji@gmail.com

https://www.linkedin.com/in/omar-hajjouji-661b95169/

https://omarhajjouji.tn

https://www.upwork.com/freelancers/~0182da1716d558065f

SKILLS

AI & LLMs

- LLM Integration (OpenAI, Gemini)
- Prompt Engineering
- MCP tools
- Voice & Real-Time AI Systems
- RAG Architectures
- Embedding Pipelines
- Real-Time Speech AI
- Structured Outputs

Backend & APIs

- Python,FastAPI, Flask,
- REST APIs, Streaming, gRPC
- Microservices Architecture,
- Async Processing,
- API Integrations
- Docker, Kubernetes
- Linux servers, GCP
- Data Ingestion Pipelines

PROFESSIONAL EXPERIENCE

TopDoctors - AI Systems & Backend Engineer

DECEMBER 2022 - TODAY

- Designed and implemented multiple AI MVPs from concept to production, owning system architecture, backend development, AI integration, and production deployment.
- Built and maintained backend microservices (Python, FastAPI) integrating LLM systems, RAG pipelines, speech processing, and analytics into existing healthcare applications end-to-end.
- Architected scalable ingestion pipelines processing structured and unstructured clinical data including smart medical device data via mobile applications storing and transforming datasets in Google BigQuery to power analytics and AI-driven chatbot systems.
- Developed an LLM-based clinical interoperability engine converting unstructured medical reports into structured, schema-constrained FHIR-compliant outputs, implementing advanced prompting strategies, structured output validation, and SNOMED CT terminology standardization.
- Built a real-time AI transcription platform generating structured clinical documentation from live consultations.
- Built a production-grade multi-agent voice AI appointment booking system, integrating speech pipelines, LLM reasoning with structured tool-calling, and backend MCP services.
- Deployed and managed containerized AI microservices using Docker and Kubernetes, supporting scalable and reliable production environments.

BeHIT Spain - AI & Speech Systems Engineering Intern

JUNE 2022 - DECEMBER 2022

- As part of NVIDIA Early access Program, we developed a multilingual (EN/ES) real-time speech-to-text system using NVIDIA Riva and achieving ~7% WER through domain-specific fine-tuning.
- Configured GCP GPU-powered Linux servers and built low-latency streaming pipelines using gRPC, proxy servers, and backend APIs.
- Deployed containerized services with Docker and Kubernetes to support scalable real-time inference.
- Built full MVPs and integrated AI speech systems into a production clinical management platform and hospital kiosk avatars for patient guidance.

Data CoLab - Head of Mentoring Department

FEBRUARY 2020 - JUNE 2022

- Led and managed a data science mentoring team, overseeing curriculum design, workshop planning, and quality assurance across AI and machine learning training programs.
- Supervised student-led ML and AI projects, providing technical guidance on model development, evaluation, and real-world implementation.

EDUCATION

National Institute of applied Science and Technology

2018 - 2024

Engineering Degree (Systems & Industrial Engineering)

El Mhamdia High School

2014 - 2018

Computer Science Baccalaureate with Honors