Marwan Mashra

Paris, France | marwan.mashra@gmail.com | mashra.dev

linkedin.com/in/marwanmashra | github.com/MarwanMashra

Summary

Lead AI Engineer who **launched Europe's first airport AI Agent** and delivered **14+ public-facing AI products**, including AI avatars deployed in **5 museums** and **4 government institutions**.

Combines technical experience in **RAG/LLM systems** with a background in end-to-end **product development in early-stage startups**.

Experience

Lead AI Engineer, Jumbo Mana – France Oct. 2023 – present

Joined as part of the **initial engineering team** and later **promoted to lead R&D**, building AI-powered avatars for high-profile clients such as VINCI Airports, the **European Parliament** and the **French Embassy** in the US.

- **Launched Europe's first airport AI Agent** at Lyon Airport, handling over **20k queries/month** and providing real-time assistance to travelers.
- **Led an R&D team of 5–7** to deliver **14+ public-facing AI products**, featured in **30+ media articles**, and deployed in high-stakes environments such as airports, museums, and government institutions.
- **Defined product strategy and R&D roadmap**, aligning development with client needs, and coordinating across teams to ensure the product's long-term success.
- Led the design and development of a **battle-tested RAG system**, vetted for factual accuracy and reliability by historians from partner museums.
- Built a **scalable data processing pipeline** for RAG, serving as the **primary data pipeline for six-figure client contracts** involving high volumes of complex, unstructured museum archives.
- Achieved a **3x speedup** in the RAG data processing pipeline, reaching a throughput of **2k pages/hour** through data-parallel distributed inference on a 4x RTX 4090 cluster.
- **Owned and operated AI infrastructure** for production inference of LLM, TTS and STT models, managing GPU clusters and handling model deployment, optimization, and scaling.
- Led the design of an **in-house evaluation framework** for the RAG system, measuring retrieval, faithfulness, factual accuracy, and custom metrics to assess verbosity and conciseness.
- Organized **workshops to promote software engineering best practices** and encourage high standards in design patterns, code quality, and reviews.

Computer Vision Engineer, Freelance, Flaggr – France Sep. – Oct. 2023

- Reduced by 60% inference time of the main vehicle detection algorithm, while losing less than 4% accuracy.
- Implemented their first version of a night vehicle detection algorithm.

Skills & Technologies

Applied AI/ML Areas: RAG/LLM systems, LLM evals, TTS & STT models

Frameworks & Tools: PyTorch, Transformers, Vector Databases, SpaCy

AI Inference stack: Triton Inference Server, TensorRT-LLM, DeepSpeed, vLLM

Cloud & DevOps: Kubernetes, Docker, Helm, Grafana, Microsoft Azure

Languages: **English**(Fluent), **French**(Fluent)

Education

MSc in Artificial intelligence, University of Paris-Saclay – France

BSc in Computer Science, University of Montpellier – France