salmaouardi1998@gmail.com
+33758224169

Salma OUARDI

LinkedIn
Paris, France

Machine Learning Engineer

SKILLS

- **Languages:** Python (NumPy, Pandas, Scikit-learn, spaCy, PySpark, Streamlit), SQL, Bash, Linux
- **French** (Fluent), **English** (Fluent)
- **Machine Learning & Deep Learning:** Supervised/Unsupervised Learning, Probabilistic Modeling, Representation Learning, Self-Supervised Learning, Transformer Architectures.
- **LLM & NLP:** RAG (advanced), Legal AI, Transformers (GPT, LLaMA, Mistral), Entity Resolution, Embeddings & Retrieval (Elasticsearch), LangChain, LangGraph
- **AI Systems & APIs:** Multi-Agent Architectures, LLM Evaluation & Robustness, Prompt Engineering, FastAPI
- **Version Control:** Git, Github
- **MLOps & Deployment:** MLflow, DVC, vLLM, AWQ (Model Quantization), CI/CD, Docker, Kubernetes
- **Cloud & Big Data:** GCP (Vertex AI, BigQuery), Azure (Databricks, DevOps, Azure Openai), Apache Spark, Airflow

WORK EXPERIENCE

OnePoint | Paris, France **September 2025 – Present**
Machine Learning Engineer – AI Department

- ARIA – Legal Risk Assessment RAG (VINCI Construction)
  - Led optimization of **ARIA**, a production-grade Legal RAG system designed to assist lawyers in assessing contractual and tender-related risks
  - Improved retrieval and answer robustness using Elasticsearch + LLM orchestration for high-stakes legal document analysis
  - Strengthened reliability and traceability across asynchronous architecture (Azure Service Bus, Promptflow, Azure Functions)
  - Enhanced error handling, logging, and evaluation strategies for legal risk scoring workflows
  - Contributed to experimentation and evaluation pipelines to improve response precision in legal reasoning contexts
- Site Reporter – Voice-to-Report AI Assistant (VINCI Construction)
  - Designed and built an end-to-end AI MVP replacing manual construction site reporting
  - Developed a full voice-to-report pipeline: Speech-to-text via Azure OpenAI / Structured extraction using Mistral LLM / DOCX generation via FastAPI backend
  - Architected modular system (Streamlit frontend + REST API + LLM services)
  - Automated full workflow and validated output quality with domain users
- DocuScore – Enterprise RAG Readiness & Corpus Intelligence Platform
  - Built DocuScore, an AI platform scoring RAG-readiness and detecting corpus anomalies (duplicates, outdated docs, outliers)
  - Designed embedding + similarity pipelines for large-scale document evaluation
  - Implemented custom scoring algorithms to assess structure, quality, and retrieval suitability

Ryte AI – HealthTech Startup | Paris, France **September 2023 – March 2025**
Machine Learning Engineer – NLP & LLMs in Healthcare

- Built transformer-based intent classification models using RoBERTa on synthetic medical queries; deployed to production on Vertex AI with >90% accuracy
- Created a multi-strategy entity extraction pipeline using LLaMA 3 and Mistral 7B for semantic parsing and spaCy for NER, with span-level F1 tracking and IOB tagging on a GPT-annotated medical dataset.
- Developed a semantic entity linking system using BioLORD embeddings and cross-encoder reranking, improving precision in mapping to standardized medical concepts
- Curated and annotated 1,000+ synthetic medical questions using GPT-assisted generation and Prodigy annotation, in collaboration with medical professionals
- Reworked core patient-provider entity resolution pipeline by combining fuzzy logic, TF-IDF, and custom heuristics; improved F1 from 77% to 98% and reduced runtime from 3 days to ~1 hour

- Engineered Spark data pipelines (~10+ TB) with Scala UDFs for large-scale batch processing, deployed via Azure Databricks
- Set up end-to-end ML delivery stack with CI/CD on Azure DevOps; managed experiment tracking with MLflow and DVC, and optimized inference for quantized models using vLLM + AWQ

Orange | Paris, France **March 2023 – September 2023**

Data Scientist Intern – GCP & Complaint Classification

- Built a customer complaint classification model using Vertex AI and BigQuery, reaching ~86% precision and ~75% recall
- Applied active learning on 200K+ unlabeled records to iteratively boost model performance and reduce manual labeling cost.
- Designed a scalable GCP prediction pipeline (BigQuery + Cloud Functions), processing 1,000 records in under 2min.
- Facilitated data science knowledge transfer to internal teams through presentations and training sessions

EDUCATION

Université Paris-Saclay | Paris, France **Sep 2023**

MSc in Artificial Intelligence

Ecole des Sciences de l'Information | Rabat, Morocco **June 2022**

Engineering Degree in Data & Knowledge

PROJECTS

- **MediGuide – RAG-Based Drug Q&A App –** Developed a local RAG assistant using LangChain, FAISS, and Ollama to answer medical queries from drug PDFs; deployed with FastAPI, Streamlit, and Docker, with CI/CD via GitHub Actions.