JULIEN RABAULT

Applied AI / ML Engineer

+33 7 81 16 46 29
julienrabault@icloud.com
linkedin.com/in/julienrabault
github.com/JulienRabault
Toulouse, France
julienrabault.github.io

Four years at CNRS, training and fine-tuning deep learning models in production (PyTorch, multi-GPU, Jean Zay supercomputer), with two peer-reviewed publications. Currently at Berger-Levrault designing Athena, an agentic AI platform powered by LangGraph, RAG, and MCP.

TECHNICAL SKILLS

Agentic & RAG :	RAG / GraphRAG, Multi-agent systems, MCP Protocol, LangChain / LangGraph, Prompt Engineering, Structured Outputs, Embeddings, LLM Routing, Fine-tuning
Deep Learning :	PyTorch, Transformers, Computer Vision, Diffusion Models (DDPM), NLP, CNNs / U-Net / YOLOv5, Generative models (GAN, VAE)
MLOps & Infra :	Docker, AWS, CI/CD, MLFlow, Airflow, Kubernetes, Celery, Langfuse, HPC / Slurm, Linux
Development :	Python, FastAPI, HuggingFace, Qdrant / pgvector, Mistral / OpenAI API, Git, SOLID / Architecture, C#, SQL

WORK EXPERIENCE

AI ENGINEER

Jan. 2026 - present

Berger-Levrault | Toulouse | AI R&D team, 12 people

Designing and building **Athena**, Berger-Levrault's agentic platform: multi-agent orchestration, intelligent routing, integrated with documents and business APIs across verticals including local government, industry, and maintenance. Currently in pilot with ~30 users. Cross-functional team (designer, frontend dev, DevOps), client workshops, Langfuse observability.

Multi-agent architecture
Designed the multi-agent architecture (LangGraph), built question-type routing, orchestrated RAG agents and MCP API agents, integrated source attribution into responses. **Platform in production**, delivering grounded and actionable answers across business domains.

Content extractor - OCR/PDF/DOCX
Redesigned and rebuilt the extraction service: OCR, images, PDF, DOCX. Async batch processing (Celery + Mistral batch API), factory/registry patterns for extensibility. **Cut extraction costs by 50%.**

Airflow pipelines
Inherited and improved document ingestion pipelines (PDF, technical manuals, work orders, equipment documentation). **5 operational DAGs.**

MCP Builder
Developed an LLM-powered pipeline that pre-processes OpenAPI specs (endpoint grouping, masking, description generation), with human-in-the-loop review for route validation and domain knowledge. **120+ internal APIs mapped**, progressively integrated at runtime.

MACHINE LEARNING ENGINEER

Dec. 2021 - Jan. 2026

CNRS - National AI Research Programme (PNRIA) | Toulouse

Collaborated with research teams across France on applied AI projects. Led two projects in parallel (6-12 months each), delivering to major French research institutions (Meteo France, CNES, CEA, INEE). Training and fine-tuning on Jean Zay (multi-GPU DDP, up to 8 GPUs, Slurm).

GENS / MetScore - Meteo France
Evaluated production weather models; performed multi-GPU optimization and fine-tuning of a diffusion model (DDPM) in PyTorch on Jean Zay. Built MetScore (YAML config, Python library). **Library still in production**; diffusion POC achieved **20% compute savings** with no loss in quality.
Co-authored AMS 2025 paper.

DeepFaune - CNRS/INEE
Fine-tuned YOLOv5 on a custom dataset (1.5M images, 24 classes), multi-GPU training, addressed class imbalance, optimized for inference speed on CPU. **93% accuracy across 24 species, 3x faster.** Peer-reviewed publication.

Other contributions
AUTOFILL (CEA, PairVAE, nanomaterial data generation, MAE 0.98), BIGSF (CNES, tech lead and architecture refactor, galactic filament image analysis library), MORPHOGAN (Univ. Lorraine, StyleGAN2 code overhaul, automated pipeline).
Taught: *Introduction to LLMs* (3-hour course, ~25 PhD students and CNRS researchers).

SOFTWARE ENGINEER (apprenticeship)

Aug. 2020 - Sept. 2021

Agileo Automation | Montauban

Built a supervision and control framework for robotic machinery in semiconductor manufacturing. C#, object-oriented architecture, HMI, CI/CD. Team of 5 engineers, Agile/Scrum.

EDUCATION

MSc - ARTIFICIAL INTELLIGENCE & PATTERN RECOGNITION (IARF) | Universite Paul Sabatier Toulouse III - IRIT | Deep Learning, Computer Vision, NLP — 2019 - 2021

BSc - COMPUTER SCIENCE | Universite Paul Sabatier Toulouse III — 2016 - 2019

Languages: French (native) | English (full professional proficiency, scientific writing)

PUBLICATIONS

"Enriching Operational High-Resolution Ensemble Forecasts with StyleGAN-2" - AIES, 2025. **3rd author**, peer-reviewed.
"The DeepFaune initiative: automatic identification of European fauna" - **Julien Rabault** et al. **Co-author**, peer-reviewed.

OPEN SOURCE PROJECTS

LLMock (PyPI) - LLM mock server for testing retries, fallbacks and rate limiting. Python, FastAPI, 10+ providers, OpenAI-compatible.
DDPM-weather - Probabilistic diffusion model for weather image denoising, 20% compute cost savings. PyTorch, Meteo France.
BananaML - End-to-end ML pipeline deployed on AWS. Computer Vision + REST API. FastAPI, Docker, CI/CD.