

Mehdi BOUBNAN

SENIOR AI ENGINEER · AI PLATFORM & ML SYSTEMS

Paris, France

+33 6 58 55 13 00 | mehdi.boubnan@gmail.com | mehdi.boubnan.com | [mehdi-boubnan](https://www.linkedin.com/in/mehdi-boubnan)



Senior AI & Platform Engineer with 6+ years of experience building and scaling production ML systems. Specialized in LLMs, AI agents, high-performance inference, backend architecture, and MLOps.

Professional Experience

Mindee

Paris, France

SENIOR AI ENGINEER (PLATFORM)

Nov 2025 - Present

- Architected and deployed Mindee's customer-facing AI support agent using Pydantic AI and AG-UI, exposing tools across the frontend and backend, integrating GitBook MCP for direct documentation access, and Intercom for seamless human handoff.
- Designed scalable backend architecture for AI agents, unifying dependency injection across FastAPI services and Taskiq workers through lifecycle-managed shared resources.
- Designed and implemented a clean data access layer using Repository and Unit of Work patterns (SQLModel), enabling short-lived sessions, efficient connection pooling, and reliable transactional behavior.
- Integrated Redis for caching and transient data management (e.g., inference results), with robust serialization strategies for complex Pydantic objects.
- Strengthened production operations by standardizing structured logging and context propagation and integrating Kibana for debugging and monitoring.

SENIOR MACHINE LEARNING ENGINEER

Nov 2023 - Nov 2025

- Rebuilt the ML deployment pipeline from the ground up, leveraging ONNX and TensorRT for optimized model formats and Triton Server for scalable, high-performance inference.
- Modernized asynchronous backend infrastructure with FastAPI, Celery, RabbitMQ, and multithreaded workers, improving modularity, throughput, and response time under load.
- Partnered with DevOps on CI/CD and observability using Helm, Docker, Argo CD, Grafana, and Prometheus, covering both Celery workers and Triton inference servers.
- Shipped full-stack features, from frontend interfaces (TypeScript) to backend APIs and PostgreSQL database migrations, enabling end-to-end product ownership.

MACHINE LEARNING ENGINEER

Nov 2022 - Nov 2023

- Fine-tuned Qwen-VL, a vision-language model, on proprietary datasets for structured document understanding with JSON outputs.
- Deployed large language models using vLLM, benchmarking inference latency across diverse hardware for optimal production readiness.
- Built a custom Retrieval-Augmented Generation (RAG) pipeline combining image and text embeddings, with fast similarity search backed by PostgreSQL (pgvector).
- Productionized ML services through optimized model exports, Docker, and GitHub Actions, publishing images to Harbor and AWS ECR.
- Created internal Python libraries for optimized geometry processing (Shapely) and synthetic data generation (Pillow/OpenCV), accelerating experimentation.

DATA SCIENTIST

Mar 2020 - Nov 2022

- Trained deep learning models for computer vision and NLP, including semantic segmentation, OCR, transformers, and LayoutLM-based architectures.
- Curated and engineered large-scale datasets by blending real-world annotated data with automatically labeled synthetic data for high-performance training pipelines.
- Built and deployed document information extraction systems by combining ML models with custom rule-based post-processing, achieving high accuracy in real-world use cases.

Outscale (formerly Fortia Financial Solutions)

Paris, France

MASTER'S THESIS INTERN - MACHINE LEARNING RESEARCHER

May 2019 (6 months)

- Developed an automated financial document generation tool in LaTeX, using reinforcement learning (DQN and PPO) to optimize the ordering of financial elements.

Société Générale

DATA SCIENCE INTERN

- Created an interactive Python dashboard using Dash for advanced credit database analysis.
- Implemented an anomaly detection algorithm to identify irregularities in credit data.

Paris, France

Mar 2018 (6 months)

Total France

DATA SCIENCE INTERN

- Forecasted petroleum product sales volumes in France using XGBoost.

Paris, France

Sep 2017 (6 months)

Education

École Normale Supérieure Paris-Saclay

MASTER OF SCIENCE - MVA (MATHEMATICS, VISION, LEARNING)

- Coursework:** Probabilistic graphical models, computer vision, object recognition, reinforcement learning, speech, and natural language processing.

Paris, France

2019 - 2020

CentraleSupélec (formerly École Centrale Paris)

ENGINEERING DEGREE - APPLIED MATHEMATICS, MAJOR IN DATA SCIENCE

- Coursework:** Advanced machine learning, optimization, statistics, deep learning, mathematics, and stochastic processes.

Paris, France

2015 - 2019

Lycée Moulay Youssef

PREPARATORY CLASSES - MPSI & MP

Rabat, Morocco

2013 - 2015

Projects

DEEP REINFORCEMENT LEARNING FOR DOOM  GitHub

- Applied reinforcement learning techniques (DQN, A3C, curiosity-driven exploration) to play various Doom game scenarios.

KAGGLE BIRD CLASSIFICATION - 4TH OF 126  GitHub

Languages & Technical Skills

Programming	Python, TypeScript, Bash, SQL
AI Coding Assistants	Codex, Claude, GitHub Copilot
AI & Machine Learning	Pydantic AI, PyTorch, TensorFlow, RAG
Inference & Deployment	Triton Inference Server, ONNX, TensorRT, vLLM, Docker
Backend & Data	FastAPI, Taskiq, Celery, Redis, PostgreSQL, pgvector, SQLAlchemy
MLOps & CI/CD	Kubernetes, Helm, Argo CD, GitHub Actions, Grafana, Prometheus, Kibana
Languages	French (native), English (fluent)