

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



Plus de 6 ans d'expérience dans la conception et la mise à l'échelle de systèmes ML en production. Expertise en LLM, agents IA, inférence haute performance, architecture backend et MLOps.

Expérience professionnelle

Mindee

Paris, France

SENIOR AI ENGINEER (PLATFORM)

nov. 2025 - aujourd'hui

- Conçu et déployé l'agent IA d'assistance client de Mindee avec Pydantic AI et AG-UI, en exposant des outils côté frontend et backend, en intégrant GitBook MCP pour accéder directement à la documentation et Intercom pour une reprise fluide par un conseiller humain.
- Conçu une architecture backend évolutive pour les agents IA, en unifiant l'injection de dépendances entre les services FastAPI et les workers Taskiq grâce à des ressources partagées gérées sur leur cycle de vie.
- Conçu et mis en œuvre une couche d'accès aux données fondée sur les patrons Repository et Unit of Work (SQLModel), avec des sessions courtes, un pool de connexions efficace et des transactions fiables.
- Intégré Redis pour le cache et les données temporaires, notamment les résultats d'inférence, avec une sérialisation robuste des objets Pydantic complexes.
- Amélioré l'exploitation en production en standardisant la journalisation structurée et la propagation du contexte, et en intégrant Kibana pour le diagnostic et la supervision.

SENIOR MACHINE LEARNING ENGINEER

nov. 2023 - nov. 2025

- Refondu la chaîne de déploiement ML avec ONNX et TensorRT pour optimiser les formats de modèles, et Triton Server pour une inférence performante à grande échelle.
- Modernisé l'infrastructure backend asynchrone avec FastAPI, Celery, RabbitMQ et des workers multithreads, améliorant la modularité, le débit et les temps de réponse sous charge.
- Collaboré avec l'équipe DevOps sur la CI/CD et l'observabilité avec Helm, Docker, Argo CD, Grafana et Prometheus, pour les workers Celery et les serveurs Triton.
- Livré des fonctionnalités de bout en bout, des interfaces TypeScript aux API backend et aux migrations PostgreSQL.

MACHINE LEARNING ENGINEER

nov. 2022 - nov. 2023

- Adapté le modèle de vision et langage Qwen-VL par fine-tuning sur des jeux de données propriétaires pour comprendre des documents structurés et produire des sorties JSON.
- Déployé de grands modèles de langage avec vLLM et comparé leur latence d'inférence sur différents matériels pour préparer leur mise en production.
- Construit un pipeline RAG combinant des embeddings d'images et de texte, avec une recherche de similarité rapide dans PostgreSQL (pgvector).
- Mis en production des services ML grâce à l'optimisation des exports de modèles, Docker et GitHub Actions, avec publication des images sur Harbor et AWS ECR.
- Développé des bibliothèques Python internes pour optimiser le traitement géométrique (Shapely) et générer des données synthétiques (Pillow/OpenCV), accélérant l'expérimentation.

DATA SCIENTIST

mars 2020 - nov. 2022

- Entraîné des modèles d'apprentissage profond en vision par ordinateur et en traitement automatique du langage, notamment pour la segmentation sémantique, l'OCR, les transformers et les architectures LayoutLM.
- Constitué et préparé de grands jeux de données en associant des données réelles annotées et des données synthétiques étiquetées automatiquement pour les chaînes d'entraînement.
- Conçu et déployé des systèmes d'extraction d'informations documentaires associant modèles ML et post-traitement fondé sur des règles, avec une grande précision en conditions réelles.

Outscale (anciennement Fortia Financial Solutions)

Paris, France

MASTER'S THESIS INTERN - MACHINE LEARNING RESEARCHER

mai 2019 (6 mois)

- Développé un outil de génération automatique de documents financiers en LaTeX, utilisant l'apprentissage par renforcement (DQN et PPO) pour optimiser l'ordre des éléments financiers.

Société Générale

DATA SCIENCE INTERN

- Créé un tableau de bord interactif en Python avec Dash pour l'analyse approfondie d'une base de données de crédit.
- Développé un algorithme de détection d'anomalies pour repérer les irrégularités dans les données de crédit.

Paris, France

mars 2018 (6 mois)

Total France

DATA SCIENCE INTERN

- Prévu les volumes de vente de produits pétroliers en France avec XGBoost.

Paris, France

sept. 2017 (6 mois)

Formation

École Normale Supérieure Paris-Saclay

MASTER MVA (MATHÉMATIQUES, VISION, APPRENTISSAGE)

- Enseignements** : modèles graphiques probabilistes, vision par ordinateur, reconnaissance d'objets, apprentissage par renforcement, parole et traitement automatique du langage.

Paris, France

2019 - 2020

CentraleSupélec (anciennement École Centrale Paris)

DIPLÔME D'INGÉNIEUR - MATHÉMATIQUES APPLIQUÉES, SPÉCIALISATION SCIENCE DES DONNÉES

- Enseignements** : apprentissage automatique avancé, optimisation, statistiques, apprentissage profond, mathématiques et processus stochastiques.

Paris, France

2015 - 2019

Lycée Moulay Youssef

CLASSES PRÉPARATOIRES - MPSI ET MP

Rabat, Maroc

2013 - 2015

Projets

APPRENTISSAGE PAR RENFORCEMENT PROFOND APPLIQUÉ À DOOM

 GitHub

- Appliqué des techniques d'apprentissage par renforcement (DQN, A3C et exploration guidée par la curiosité) à différents scénarios du jeu Doom.

CLASSIFICATION D'OISEAUX SUR KAGGLE - 4E SUR 126

 GitHub

Langues et compétences techniques

Programmation	Python, TypeScript, Bash, SQL
Assistants de code IA	Codex, Claude, GitHub Copilot
IA et ML	Pydantic AI, PyTorch, TensorFlow, RAG
Inférence et déploiement	Triton Inference Server, ONNX, TensorRT, vLLM, Docker
Backend et données	FastAPI, Taskiq, Celery, Redis, PostgreSQL, pgvector, SQLAlchemy
MLOps et CI/CD	Kubernetes, Helm, Argo CD, GitHub Actions, Grafana, Prometheus, Kibana
Langues	Français (langue maternelle), anglais (courant)