Yassine Bentayfor

Évry-Courcouronnes, France · bentayfor.yassine@gmail.com · phone on request
LinkedIn · GitHub · Kaggle · Available for CDI Sep. 2026

PROFILE

Final-year double-degree engineer (ENSIIE × EMI) building reliable ML and data systems. Currently interning at P&G (Data Foundations FBNL); four prior internships across data engineering, machine learning, applied GenAI, and MLOps (Thales, Oracle Labs, Integrative Phenomics, Oracle R&D). Comfortable shipping production PySpark/Delta Lake pipelines and LLM systems with RAG, fine-tuning, and rigorous evaluation. Solid foundation in statistics, optimisation, and statistical learning. Available for CDI Sept. 2026.

EDUCATION

ENSIIE – National School of Computer Science for Industry and Business *Sep. 2024 – Present*
M.Eng. in Machine Learning, Applied Mathematics & Statistics (Double Degree) Évry, France

Mohammadia School of Engineers (EMI) *Sep. 2022 – Jun. 2026*
Engineering Degree in Computer Science and Digitalization Rabat, Morocco

EXPERIENCE

Procter & Gamble *Feb. 2026 – Aug. 2026*
Data Engineering Intern – Data Foundations France-Benelux

- Developed and maintained PySpark/Delta Lake ETL pipelines on Azure Databricks for commercial reporting across France-Benelux markets, covering two product lines from ingestion to Power BI delivery.
- Automated deployment with Databricks Asset Bundles + full CI/CD pipeline (dev → prod promotion, quality gates); migrated legacy pipelines to Unity Catalog and resolved technical debt across 15 production notebooks (schema drift, SCD corruption, Hive Metastore dependencies).

Integrative Phenomics *Jun. 2025 – Sep. 2025*
Data & Applied ML Intern Paris, France

- Modelled clinical nutrition planning as a MILP, coupled with a guardrailed LLM generative layer for plan synthesis; 99.2% feasibility on 120 stress tests, sub-second solve latency.
- Re-platformed a single-user prototype into a multi-user service (FastAPI, PostgreSQL, async workers). Cut clinical-team iteration time by 22% by integrating LLM-based report generation and PCA-based longitudinal biomarker profiling into production.

Oracle Labs *Jul. 2024 – Sep. 2024*
Machine Learning Engineer Intern Casablanca, Morocco

- Designed a production semantic search system over large codebases: E5-Large-v2 bi-encoder, ColBERT reranker, RAG. Retrieval quality up 47% on nDCG and precision, with retrieval and generation benchmarked separately so gains were attributable.
- Built a parametrised Pytest evaluation harness integrated into CI; test coverage went from 56% to 86% and p95 latency dropped 41%.

Oracle *Jun. 2023 – Sep. 2023*
R&D Intern – AI Code Assistance Casablanca, Morocco

- Fine-tuned StarCoderPlus 15B with LoRA inside a custom RAG pipeline (FAISS, sentence-transformers); +17% factual alignment on BLEU and AlignScore over 20+ structured prompts. Touched the full LLM lifecycle: dataset curation, PEFT, prompt design, structured offline evaluation.

Thales Group *Jan. 2024 – Jul. 2024*
Cloud Engineering Apprentice – MLOps Remote

- Shipped event-driven data and MLOps pipelines on AWS (Lambda, API Gateway, EventBridge, SNS, S3) with Docker, CI/CD, structured observability, and reusable IaC modules. Deployment fully automated, least-privilege IAM across services.

PROJECTS

Neural Network Volatility Modeling *(TensorFlow, Black-Scholes, Heston, COS method)* 2025
Replicated Buehler et al. (2019) "Deep Learning Volatility": 4-layer ANN for implied-volatility prediction under BS (1M LHS samples) and Heston (100K samples via COS + Brent IV inversion). Reached BS MSE 1.55e-8 and Heston MSE 1.14e-6, matching the paper.

Big Data Accident Analytics Pipeline *(Kafka, Spark, MinIO, Trino, OpenSearch, Airflow)* 2025
Six-stage distributed pipeline on 1M+ accident records: Kafka → MinIO → Spark → Trino + OpenSearch, orchestrated by Airflow with optional Spark ML classification.

TECHNICAL SKILLS

Data Engineering	PySpark, Spark, Azure Databricks, Delta Lake, Unity Catalog, Lakehouse, Medallion Architecture, Dimensional Modelling, SCD (Type 1/2), ETL/ELT, Airflow, Data Quality, Lineage, DataOps
Stats / Quant / Math	Hypothesis Testing, Bayesian Inference, Causal Inference, Probabilistic Graphical Models, Time Series, Stochastic Processes, Black-Scholes / Heston, COS Method, Implied Volatility, Monte Carlo, MILP (CPLEX/CBC), Convex Optimisation
ML / GenAI	PyTorch, TensorFlow, Hugging Face Transformers, RAG, LoRA / PEFT, Embeddings (E5, sentence-transformers), Semantic Search (FAISS), Rerankers (ColBERT), Prompt Engineering, Evaluation (BLEU, AlignScore, nDCG, ablations), FastAPI, LangChain
Cloud & Infra	AWS (Lambda, Glue, Redshift, S3, API Gateway, EventBridge, IAM), Azure (Databricks, Data Lake), GCP (BigQuery), Docker, Kubernetes, Linux, CI/CD, Infrastructure-as-Code
Languages & Storage	Python (advanced), SQL (PostgreSQL, Redshift, BigQuery, T-SQL), Scala, Java, Bash, R | Delta, Parquet, NoSQL, REST APIs
Certs & Awards	OCI GenAI Pro · OCI Foundation · IBM DevOps Pro · **Kaggle** House Prices Top 10/5000+ (2025) · **1st** E-Toufoula Hackathon (2024) · **2nd** Think AI Morocco (2024) · **2nd MA / 5th AF** ODC (2024) · NASA Space Camp Commander's Cup (2019)
Languages	Arabic, French (native), English (fluent)