

Ahmed Doghri

SUMMARY

Senior AI Engineer with 10 years shipping end-to-end production ML systems. Specialize in RAG pipelines, multi-agent orchestration, and LLM evaluation harnesses (RAGAS, DeepEval). Track record of measurable impact in regulated domains such as pharma, fintech, and B2B.

PROFESSIONAL EXPERIENCE

Idea Boundary | Remote USA

Founder and Principal AI Engineer (Independent Consulting) 01/2025 - Present

- Built production Slack RAG assistant with citation grounding, hybrid retrieval with reranking (FAISS, BM25), and graph-based Q&A (Neo4j, LangChain); achieved 94% citation precision on RAGAS eval harness, reduced response latency 35% and manual lookups 50%.
- Reduced manual product tagging 60% (92% accuracy) by designing a multi-agent orchestration system for e-commerce catalog ops; implemented confidence-gated human-in-the-loop workflows using LangGraph and custom prompt strategies.
- Increased customer retention 12% by deploying a churn prediction system (XGBoost, dbt, MLflow) with 78% precision and automated retraining pipelines.

Weave | USA

Senior AI Engineer 03/2024 - 12/2024

- Reduced hallucination rate 60% (automated fact-checking + human QA) by architecting 7 production LLM agent pipelines for FDA IND submissions; implemented function calling, retrieval grounding, citation validation, and AI safety guardrails.
- Deployed RAG pipeline for table extraction (Elasticsearch, Spark); cut parsing time 40%.
- Cut inference costs 30% and p95 latency 25% on production SageMaker endpoint (50K+ requests/day) via model distillation and vLLM batched serving.

BenchSci | Canada

Senior Machine Learning Engineer 04/2023 - 03/2024

- Improved biomedical NER accuracy 30% and throughput 25% by deploying transformer + LLM ensemble on Vertex AI with DeepEval evaluation framework and CI/CD automation.
- Increased ML pipeline success rate 35% and reduced p95 runtime 20% by productionizing end-to-end MLOps (Vertex AI, Kubeflow, MLflow) with CI/CD and defined SLOs.

Machine Learning Engineer 01/2022 - 04/2023

- Architected real-time ML inference pipeline (Kafka, Flink) processing 50K+ events/hour; reduced latency 35% and improved throughput 20% via distributed GPU clusters (GCP).
- Achieved 95% F1 on biomedical document classification by fine-tuning transformers (BERT, BioBERT, LoRA) deployed production serving on Vertex AI at 1M+ predictions/week.

IBM | Canada

Senior Data Scientist and Consultant 05/2021 - 01/2022

- Drove \$4M contact-center savings by deploying 3 bilingual conversational AI assistants (Watson, Azure Bot Service) for enterprise banking clients.

- Built production PII masking pipeline processing 100K+ chats/day using spaCy NER, PyTorch, and async task queues (Flask/Celery); ensured regulatory compliance for banking data.

Data Scientist and Consultant

08/2019 - 05/2021

- Scaled NLP clustering across 1.2M chat logs for intent discovery and routing optimization.
- Developed portfolio optimization models for Canadian pension funds (XGBoost, Spark); achieved 5% better risk-adjusted returns vs. historical worst-sector benchmarks.

Ford Motor Company | Canada

Graduate Research Assistant - Embedded ML

05/2017 - 05/2019

- Achieved 98% precision on real-time engine misfire detection by building embedded ML pipeline (Wavelet Transform, PCA, RNN classifier)—deployed to production vehicle systems
- Designed and shipped custom Data Acquisition system processing automotive sensor signals at 100MHz with signal conditioning for under-hood deployment [C, Linux]

EDUCATION

- **McMaster University** | Canada 05/2017 - 05/2019
Master of Science in Mechatronics (Summa Cum Laude) | **GPA:** 12/12
- **KAIST University** | South Korea 09/2012 - 02/2017
Bachelor in Electrical Engineering (Summa Cum Laude) | **GPA:** 4.05/4.3

SKILLS & OTHER

- **LLM & AI Frameworks:** LangChain, LangGraph, LlamaIndex, OpenAI, Azure OpenAI, Anthropic, vLLM, TensorRT, Hugging Face Transformers
- **ML & NLP:** PyTorch, TensorFlow, spaCy, scikit-learn, XGBoost, BERT, LoRA/PEFT, Fine-Tuning
- **Techniques:** RAG, Hybrid Retrieval, Reranking, Agentic AI, Multi-Agent Orchestration, Prompt Engineering, Function Calling, Context Engineering, AI Safety, Distributed training
- **Vector & Data Stores:** FAISS, Pinecone, Milvus, Neo4j, Elasticsearch, PostgreSQL
- **MLOps & Infrastructure:** MLflow, Kubeflow, Vertex AI, SageMaker, Bedrock, Docker, Kubernetes, Kafka, Flink, Spark, dbt, CI/CD
- **Cloud Platforms:** AWS, GCP, Azure
- **Evaluation & Observability:** RAGAS, DeepEval, OpenTelemetry, Phoenix, Eval Harnesses, Guardrails, Hallucination Detection, Token Cost Monitoring
- **Languages:** French (Fluent), Arabic (Fluent), Spanish (Intermediate)