

Mustapha Unubi Momoh

Ontario, Canada

☎ (548) 255-4426  [linkedin.com/in/mustaphaunubi](https://www.linkedin.com/in/mustaphaunubi)  mustaphaunubi@gmail.com  github.com/mustaphaunubi



Summary

Applied Scientist and Machine Learning Engineer with expertise in Generative AI, recommender systems, search, and causal inference. Led teams in various endeavours, including deploying low-latency large-scale personalization for millions of users. Experienced in fine-tuning and deploying diffusion models and large language models in real-world applications, and proficient across the entire machine learning lifecycle. Active open-source contributor, published researcher, and winner of some data science hackathons.

Skills

- **Programming:** Python, R, Java, SQL, C++
 - **ML/AI Frameworks:** Pytorch, TensorFlow, Nvidia Merlin, Hugging Face, LangChain, LlamaIndex, FAISS, Weaviate, Pinecone
 - **GenAI/LLM:** Amazon Bedrock, OpenAI GPTs, VertexAI Agent Builder, RAG
 - **Search & RecSys:** Elasticsearch, OpenSearch, BM25, semantic search, hybrid search, DLRMs, session-based recommendations
 - **Cloud/MLOps:** AWS (Sagemaker, Lambda, Bedrock, Personalize, CloudFormation), GCP (Vertex AI, BigQuery), Azure (Document Intelligence, Azure ML), Terraform, Docker, Kubernetes, MLFlow, GitHub Actions
 - **Systems & Tools:** Unix/Linux, Bash, Git, GraphRAG, Knowledge Graphs (Neo4j, Amazon Neptune), Distributed GPU Training, Experimentation & A/B testing
-

Education

- **University of Waterloo, Kitchener, Ontario** **Sept 2022 – Dec 2024**
Master of Applied Science, Systems Design Engineering
Thesis: [Remote Medical Diagnosis in Virtual Reality: A Mixed-methods Approach to Understanding Patients' and Physicians' Perceptions through Thematic Analysis and Regression Discontinuity Design.](#)
Tools: Data Science, Causal Inference, Beta regression, Thematic Analysis
 - **Relevant Coursework:**
 - * Advanced Topics in Pattern Recognition: Graphical Deep Learning (SYDE 770)
 -  **Project Paper:** [Comparative Analysis: Real-World Weight Cross-Entropy Loss Function Across Various Activation Functions](#)
The project aims to evaluate a real-world weighted cross-entropy loss function (RWWCE) using different activation functions on challenging problems such as credit card fraud detection and Fashion MNIST classification.
 - * Information Visualization for AI Explainability (CS 889):  **Project Paper**
 - * Data Structure in Health Informatics (CS 792), Time Series Analysis (SYDE 631)
 - **University of Port Harcourt, Port Harcourt** **2010 – 2015**
Bachelor of Engineering, Petroleum Engineering (second best grad, 2nd Class Honours, Upper Division)
Publication: [Experimental Evaluation of Particle Sizing in Drilling Fluids to Minimize Filtrate Losses and Formation Damage.](#)
Relevant courses: Computer Programming for Engineers, Numerical Computation, Probability and Statistics.
-

Work Experience

Pixite Inc.

Remote

Machine Learning Engineer (Recommender systems and Search)

November 2024 – July 2025

- Collaborated with the product team to define data requirements for personalized search and recommendation features for [Pigment app](#).
- Preprocessed the user interaction events and items metadata to support ingestion and training of the recommendation models.
- Trained and deployed recommendation models for [Pigment app](#) enabling homepage content personalization for millions of users.
- Led strategic decisions around recommendation request/response caching to optimize performance, including evaluating trade-offs between edge caching and API Gateway layers, and deciding which recommendation types to cache.
- Built an offline evaluation pipeline from scratch to compute relevance metrics (MRR, NDCG, Precision@K), facilitating the comparison of models.
- Conducted offline experiments to assess the effect of feature augmentation (for example, the effect of new event attributes and item metadata) on ranking relevance; documented findings to inform subsequent modelling efforts.
- Documented A/B testing requirements to measure engagement uplift from homepage personalization.

EveryRate

Remote

Data Engineer

May 2024 – December 2024

- Evaluated the performance of OCR pipeline tools such as Amazon Textract, Google Document AI, Azure AI Document Intelligence, and Vision language models, including Open-AI GPT-4o for tabular data extraction.
- Developed an ETL pipeline to extract mortgage rates from structured documents using Azure AI Document Intelligence, Azure Functions, and Blob triggers.
- Implemented a workflow that automates the parsing of rate sheets, extracts essential metadata, and inserts the data into a PostgreSQL instance. This solution significantly minimized manual efforts.
- Refactored PostgreSQL database to ensure efficient data ingestion and retrieval.

OKRFI (currently inactive)

Remote

Machine Learning Engineer (ML and MLOps)

February 2024 – May 2024

- Trained and deployed a hybrid (CNN-RNN) model for detecting deepfakes in identity verification videos common in financial KYC.
- Deployed an event-driven system consisting of AWS Lambda, API Gateway, and SNS to support fast and cost-effective inference.
- Automated the versioning and deployment of models using SageMaker Model Registry.
- Led data science operations in partnership discussions with [Brex](#) and [Tyfone](#), providing detailed benchmarking reports to support the pitch.

Capgemini

Remote

Freelance Data Scientist and Generative AI/MLOps Engineer

March 2023 – February 2024

- Secured early access to AWS services, including Amazon Bedrock and LLMs such as Amazon Titan and Titan text Embeddings, to prototype GenAI proof of concepts rapidly.
- Fine-tuned and deployed Stable Diffusion and Chat-Bison models on Vertex AI to support a GenAI proof-of-concept for a beauty retail use case.
- Built an AI shopping assistant POC using vector and graph databases to improve product discovery on an e-commerce tire platform. It is similar to Shopify's assistant but tailored to the client's inventory.

- Developed a GenAI-based automation application with Atlassian Confluence and Amazon Bedrock to improve documentation generation and review.

Founder and Data Analytics Trainer, Karaam Analytics

April 2020 – December 2022

- Operated as an independent contractor, collaborating with clients and organizations to assess and address their data analytics literacy needs.
- Delivered training on introduction to data analytics to staff of ExxonMobil, Diageo, and individual clients in Nigeria.

Data Science & STEM instructor, [Data Scientists Network](#) and [Tuteria Limited](#)

April 2018 – Dec. 2021

- Delivered classes on Python and SQL for Data science.
 - Volunteered as a tutor, delivering AI instruction to learners in Lagos, Nigeria
 - Provided lessons on STEM subjects and standardized tests.
-

Teaching, Leadership, Membership, and Awards

- *AWS Community Builder in Machine Learning and GenAI* 2024 – 2025
 - *Graduate Teaching Assistant: Linear Algebra, 3D Viz in AutoCAD (both @UWaterloo)* 2023 - 2024
 - *International Master's (research) Award of Excellence, University of Waterloo.* 2023 - 2024
 - *Code of Conduct Volunteer (Django 2022 Conference, Porto)* July 2022 – Sept 2022
 - *Teaching Assistant/Section leader: Stanford University (Code in Place)* April 2021 – May 2021
-

Open-Source Contributions, Projects, and Hackathons

AWS Retail Demo Store

- Specific Contributions to **Personalization and Experimentation**:
 - [PR: Updates to the Thompson Sampling Algorithm in the Multi-armed Bandit Experiment](#)
 - [PR: Updates to the products service for full items retrieval, Others](#)

2024 Nvidia's Generative AI on RTX PCs Contest: [Simplifying Documentation review on Atlassian Confluence with TensorRT-LLM and LLAMA2](#)

- Tools: Quantization, Compiling LLMs to TensorRT-LLM, Optimized LLAMA-2 deployment for low-latency GPU inference, Docker, Streamlit

2022 SAS Hackathon (top 3 in retail) :[End-to-End Marketing Analytics to facilitate the development of customer retention programs](#)

2021 Blackathon (1st place): [AI Assistive Technology for the Visually Impaired and Blind](#)