Mustapha Unubi Momoh

Ontario, Canada

(548) 255-4426 linkedin.com/in/mustaphaunubi mustaphaunubi@gmail.com github.com/mustaphau

Summary

Machine Learning Engineer with experience in building production ML infrastructure and applied generative AI.

Skills

- **Infrastructure:** Kubernetes, Sagemaker pipelines and endpoint, BigQuery, Elasticsearch, NVIDIA Triton Inference Server, Kubeflow, Docker, Amazon EKS, Redis, FAISS (ANN index), Feast, Prometheus, Grafana
- **ML Frameworks:** TensorFlow, PyTorch, NVIDIA Merlin
- **GenAI/LLM:** RAG, LLM fine-tuning, prompt engineering, TensorRT-LLM, quantization, LLM evaluation, Amazon Bedrock, LangChain. LlamaIndex

Featured Machine Learning projects

Multistage Multimodal Recommender System on Amazon Elastic Kubernetes Service *March - May 2026*

Medium Article | *TDS Article* | *Codebase* | *Demo*

- Built and deployed an end-to-end multistage multimodal recommender system with Two-Tower candidate generation, DLRM ranking, diversity-based reranking, and candidate filtering on Amazon EKS
- Handled user cold-start through feature masking and context-aware recommendations; improved content based signals and item cold-start using CLIP image and Sentence-BERT item embeddings.
- Reduced item feature lookup latency by 99.7% (from 195 ms to 0.5 ms), end-to-end latency by 54%, and improved throughput by 310% all through in-memory feature caching at model initialization.
- Deployed the full serving stack on Amazon EKS: NVIDIA Triton for model serving, Feast for feature serving, FAISS for candidate retrieval, Kubeflow for pipeline orchestration, and Valkey-backed Bloom filters for seen-item filtering.

Recommender System with Continuous Retraining on Amazon EKS *Jan - March 2026*

Medium Article | *Codebase*

- Built and deployed a DCN-based ads-ranking recommender system on Amazon EKS that automatically retrains when model performance (based on AUC-ROC) drifts below a defined threshold.
- Designed the project around production ranking concerns, including model serving, performance monitoring, retraining triggers, and server autoscaling.

KaraamAI — AI-Powered Documentation Assistant with TensorRT-LLM *2024*

Codebase| *Demo* *(2024 NVIDIA Generative AI on RTX PCs Contest)*

- Built a RAG-powered documentation assistant that connects to Atlassian Confluence, scrapes and indexes team documentation to a vector store, and enables conversational Q&A over the knowledge base.
- Implemented content generation capabilities including article writing grounded in existing documentation style and automated generation of PowerPoint presentation.
- Compiled and deployed Llama 2 on TensorRT-LLM with quantization for low-latency local GPU inference.
- Wrote a publish-back workflow that pushes AI-generated articles directly to the user's Confluence space via the Confluence API.

Relevant Experience

Pixite Inc. (Contract) *November 2024 – July 2025*

Machine Learning Engineer, Recommender Systems and Search *Remote*

- Designed and proposed recommender-system architecture options on AWS and GCP, evaluating trade-offs in training speed, inference latency, delivery timelines, and operating costs across data ingestion, model training, and inference.
- Collaborated with engineering to train recommendation models for *Pigment app* enabling homepage content personalization for millions of users.
- Built an offline evaluation pipeline from scratch to compute relevance metrics (MRR, NDCG, Precision@K), facilitating the comparison of models.

Other Contracts (EveryRate & Upwork clients) *March 2023 – present*
Data Science, GenAI, and ML Engineering *Remote*

- Implemented an AI shopping assistant integrating Amazon Kendra, Amazon Neptune, and Bedrock-hosted LLMs for retrieval-augmented product discovery for a client
- At EveryRate, I designed and deployed an ETL pipeline to extract mortgage rates from structured documents using Azure AI Document Intelligence, Azure Functions, and Blob triggers.
- At EveryRate, I benchmarked OCR pipeline tools (Amazon Textract, Google Document AI, Azure AI Document Intelligence) and vision-language models for tabular data extraction.
- Consulted for various Upwork clients on AI-powered shopping assistant, LLM-powered medical records parsing, injury claim summarization.

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.
Skills: Data science, causal inference, regression discontinuity design, beta regression, thematic analysis.

University of Port Harcourt, Port Harcourt **2010 – 2015**
Bachelor of Engineering, Petroleum Engineering; Second Best Graduating Student; Second Class Honours, Upper Division
Relevant coursework: Computer Programming for Engineers, Numerical Computation, Probability and Statistics.

Leadership, Membership, and Awards

- *AWS Community Builder in Machine Learning and GenAI* *2024 – 2026*
- *International Master's Research Award of Excellence, University of Waterloo* *2023 – 2024*
- *Graduate Teaching Assistant: Linear Algebra, University of Waterloo* *2023 – 2024*
- *Teaching Assistant / Section Leader, Stanford Code in Place* *April 2021 – May 2021*

Open-Source Contributions and Hackathons

- **AWS Retail Demo Store:** Contributed to personalization and experimentation components, including updates to the Thompson Sampling multi-armed bandit experiment and product-service updates for full-item retrieval.
- **Hybrid Search Application:** Hybrid search application that integrates traditional full-text (lexical/BM25) search with semantic (neural sparse embeddings) search built with OpenSearch
- **2022 SAS Hackathon, Top 3 in Retail:** Built an end-to-end marketing analytics system for customer retention programs.
- **2021 Blackathon, 1st Place:** Built an AI assistive-technology project for blind and visually impaired users.