

ABRAHAM AJIBADE

Louisville Metropolitan Area | 859-693-3366 | abraham0ajibade@gmail.com | [LinkedIn](#) | [GitHub](#) | [Website](#)

PROFESSIONAL SUMMARY

Artificial Intelligence and Machine Learning Solution Architect with 5+ years of experience architecting, deploying, and operating production AI systems across healthcare, finance, retail, and insurance. Expertise spans generative AI, agentic systems, LLM inference, machine learning, MLOps/LLMOps, and cloud-native AI infrastructure. Proven ability to design high-performance inference platforms, resilient AI workflows, and enterprise observability and governance frameworks across Azure, AWS, and GCP. Experienced in translating complex business and technical requirements into secure, scalable AI architectures and leading solutions from experimentation through production deployment and operational ownership.

TECHNICAL SKILLS

- **Programming & Development:** Python, SQL, R, Groovy, Bash
- **Generative AI & Agentic Systems:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Agentic AI, Multi-Agent Systems, Model Context Protocol (MCP), Prompt Engineering, Vector Databases, LangChain, LangGraph, Semantic Kernel, Pydantic AI, FastMCP
- **LLM & Model Serving:** NVIDIA Triton Inference Server, SGLang, vLLM, ONNX, OpenVINO, PyTorch, Hugging Face Transformers, Azure Kubernetes Service (AKS), Azure Container Apps (ACA)
- **Machine Learning & Deep Learning:** PyTorch, Scikit-learn, PySpark MLlib, XGBoost, Supervised & Unsupervised Learning, NLP, Transformer Models, Computer Vision, Hyperparameter Optimization, Azure Machine Learning, Azure Databricks, AWS SageMaker
- **MLOps, LLMOps & Observability:** MLflow, Arize AI, Prometheus, Grafana, Tempo, Model Monitoring, Model & Request Tracing, Drift Detection, Ground-Truth Performance Monitoring, Model Lifecycle Management
- **Cloud & AI Platforms:** Azure, AWS, GCP
- **Containers & Infrastructure:** Docker, Kubernetes, Terraform, Ansible, Databricks Asset Bundles
- **Data Engineering:** Apache Spark, PySpark, Databricks, Apache Airflow, ETL/ELT, Distributed Data Processing, AWS EMR, AWS Glue
- **DevSecOps & Engineering:** Git, GitHub, Bitbucket, Azure DevOps, GitHub Actions, Azure Pipelines, Jenkins, Drone CI, pytest, unittest, Ruff, SQLFluff, pre-commit, Dev Containers

WORK EXPERIENCE

Machine Learning Engineer

December 2024-Present

Northwestern Medicine | Chicago, Illinois

- Architected and deployed a high-concurrency AI inference platform on Azure Kubernetes Service (AKS) using NVIDIA Triton Inference Server and ONNX-optimized models, sustaining a sub-2-seconds P99 latency for real-time radiology report analysis under high-volume, event-driven workloads.
- Engineered production observability and instrumentation for Triton inference endpoints using native metrics, Prometheus, and Grafana, providing operational visibility into request latency, throughput, concurrency, GPU utilization, and model-serving health.
- Engineered a reusable model conversion framework to migrate models across inference backends, abstracting backend-specific requirements and standardizing model transformation, validation, and deployment workflows to improve model portability and reduce migration effort across production serving platforms.
- Implemented end-to-end model tracing for the primary client-facing Business Logic Scripting (BLS) model served through the Triton Inference Server, enabling request-level analysis of inference execution paths, downstream model interactions, and latency bottlenecks to accelerate production incident diagnosis.
- Architected a resilient LLMOps pipeline across Azure Databricks and GPU-accelerated Azure infrastructure, orchestrating daily batch inference for Mixtral 8x7B through Triton Inference Server on NVIDIA A100 GPUs with automated retries, failure recovery, and operational alerting.
- Engineered and deployed reusable project templates using *Copier* to standardize repository architecture, development tooling, CI/CD configuration, and deployment patterns across multiple ML/AI projects, reducing project setup overhead and accelerating deployments by approximately 20% compared with non-templated implementations.
- Engineered fault-tolerant batch inference and data integration pipelines on Azure Databricks, implementing state-based recovery, automated retry strategies, failure notifications, and validated downstream synchronization to maintain data integrity during transient infrastructure and upstream-system failures.
- Established closed-loop MLOps monitoring with Arize AI, integrating feature and prediction drift detection, ground-truth performance evaluation, threshold-based alerting, and automated audit trails to identify model degradation across production deployments.
- Led a DevSecOps transformation within Azure DevOps by integrating automated code, dependency, and secret

scanning into CI/CD pipelines, enabling the identification and remediation of 150+ critical vulnerabilities before production deployment.

- Authored scalable ML architecture and deployment standards, driving adoption of Terraform, containerized inference, and standardized MLOps patterns across Azure AKS, Container Apps, and Databricks to improve deployment consistency, governance, and infrastructure reproducibility.
- Established a shift-left engineering quality strategy using Devcontainers, pre-commit hooks, Ruff, pytest, and SQLFluff, reducing failed remote CI/CD runs by 40%+ while decreasing unnecessary build execution and associated infrastructure costs.
- Served in the production maintenance and incident-response rotation, owning the operational health of deployed ML services, troubleshooting cross-platform inference and pipeline failures, and driving continuous improvements to reliability, observability, and CI/CD stability.

Data Scientist

May 2023-November 2024

Blue Lambda Technologies | Atlanta, Georgia

- Architected and scaled end-to-end ML systems across three distinct domains (real estate, retail, finance), delivering PyTorch, XGBoost, and Scikit-Learn models to production with 95%+ uptime via AWS SageMaker for scalable batch inference and automated retraining.
- Collaborated on the development and deployment of an internal RAG platform using LangChain and Qdrant Vector Database with Gemma 3.4B, resolving up to 7,000 weekly employee queries and reducing support ticket volume by 42% with 88% first-pass accuracy.
- Owned the data engineering backbone for 15+ concurrent models; built and maintained robust ETL pipelines using PySpark on Databricks/AWS EMR, strictly enforcing data lineage and quality gates to ensure production-grade, validated datasets.
- Instituted proactive MLOps monitoring using Arize and MLflow to detect data and concept drift; integrated auto-alerts and rollback protocols, successfully sustaining less than 5% metrics degradation across models over 18-month lifecycles.
- Championed and enforced engineering standards by establishing reproducible ML environments using Docker and Docker-Compose configurations via Devcontainers.
- Instituted static analysis, secret scanning, and automated linting in CI/CD workflows, reducing technical debt by 60% in training and deployment pipelines.
- Designed and implemented scalable feature engineering workflows for structured and unstructured data, which improved model generalization across the three distinct business domains.
- Authored living documentation standards for data flows, model cards, and inference APIs, adopted as team-wide templates to accelerate cross-team collaboration and streamline compliance audits.

Graduate Research Fellow

January 2021-May 2023

University of Kentucky | Lexington, Kentucky

Tyler B. Mark Industrial Hemp Research Lab

- Implemented optimization softwares (AIMS, R and Python) and prescriptive analytics on resource allocation problems for bourbon manufacturing firms, culminating in published research and improving operational efficiency by 22%.
- Conducted rigorous statistical modeling for academic research papers, utilizing R and Python to validate complex datasets and ensure 95% confidence intervals of predictions.
- Authored research findings based on predictive modeling techniques (market trends for industrial hemp firms).

Data Scientist

January 2019-November 2020

Fiverr & UpWork | Victoria Island, Lagos, Nigeria

- Delivered custom ML solutions (Scikit-learn, XGBoost) for classification and regression across diverse client portfolios, achieving 15% average R^2 uplift and 22% precision gain by directly aligning model metrics to business-critical KPIs.
- Architected a real-time incremental data ingestion system using PySpark on AWS EMR, processing over 1.5M rows per run with sub-minute latency, reducing data-to-model latency by 87% and enabling live predictive scoring.
- Engineered production-grade feature pipelines, incorporating automated hyperparameter optimization and SHAP-based explainability to enable non-technical stakeholders to interpret key model drivers via interactive dashboards.
- Refactored legacy code (Jupyter notebooks) into modular, PEP8-compliant packages; instituted comprehensive pytest/unittest suites (unit/integration/E2E), achieving 99% test coverage and guaranteeing audit-ready client handoffs.
- Automated and scaled preprocessing workflows across multi-gigabyte datasets, implementing schema validation, drift detection, and fault-tolerant retries to ensure always-ready data delivery with zero manual intervention.
- Drove model governance and transparency by integrating SHAP values and model cards directly into production APIs, empowering clients to trust and act on predictions with verifiable confidence.

Research and Content Intern

January 2018-December 2018

Gidimo | Lagos, Nigeria

- Optimized mobile learning content for a WAEC prep application serving 10,000+ students, ensuring 100% alignment with national education policy and quality standards.
- Leveraged user feedback analysis and research to drive a 15% improvement in content engagement, providing data-backed insights for iterative platform enhancements.
- Streamlined content delivery workflows and coordinated large-scale uploads to support organizational strategies for expanding educational access.
- Standardized quality assurance protocols for mobile content uploads, reducing content errors and improving the overall user interface experience for students.
- Coordinated cross-functional workflows between the research and technical teams to streamline the deployment of high-priority exam prep features.

RECENT PROJECTS

Retail Chat Agent Backend – Multimodal Product Discovery

- Developed and deployed an AI-powered chat assistant for multimodal retail product discovery, enabling users to search via natural-language descriptions or photo uploads. Integrated OpenCLIP (for image embeddings), BLIP (for image captioning), text embeddings, and vector database searches; served via FastAPI and AWS infrastructure to deliver fast, relevant matches and streamline the shopper experience.

Retail Chat Agent MCP Server – Multifunction Tools Support

- Designed and deployed a Model Context Protocol (MCP) server to provide real-time product, vendor, and inventory search capabilities for multimodal retail agents. Implemented modular tool endpoints for vector search, database queries, and workflow automation, enabling chat-based and autonomous agents to perform high-precision product discovery, validate inventory, and trigger downstream operational actions. Integrated seamlessly with the multimodal search backend to unify text, image, and metadata retrieval across retail catalogs.

Human Resources Policy RAG Chatbot

- Architected and deployed a specialized RAG (Retrieval-Augmented Generation) chatbot, using FastAPI and Qdrant to provide context-grounded answers on human resources policy-related questions within the organization. The entire system was engineered for data privacy by ensuring local language model integration and storing proprietary documentation exclusively within a self-hosted Vector Database, eliminating reliance on external APIs.

Retail Chat Agent ETL – Automated Multimodal Data Pipeline

- Engineered and deployed a production-grade ETL pipeline to automate the ingestion and vectorization of large-scale e-commerce catalogs. Built with Apache Airflow and Python, the system orchestrates web scraping and API calls, AWS S3 image hosting, and metadata storage in PostgreSQL. Integrated OpenAI CLIP and BLIP to generate synchronized text and image embeddings, populating a Qdrant vector database to power high-performance semantic search for downstream AI chat agents.

EDUCATION

Master of Science in Agricultural Economics

2021-2023

University of Kentucky | Lexington, Kentucky

Bachelor of Science in Agricultural Economics

2012-2017

University of Benin | Benin-City, Nigeria