

# IBRAHIM MUSAH

London, UK | +44 7442829882 | [a.ibrahimmusah@gmail.com](mailto:a.ibrahimmusah@gmail.com) | [www.linkedin.com/in/ibrahim-musah](https://www.linkedin.com/in/ibrahim-musah)

Microsoft-Certified Azure AI Engineer specializing in designing and deploying agentic AI solutions and fine-tuned LLMs. Delivered end-to-end ML lifecycle, with a strong focus on scalable and governable AI systems. Seeking to leverage technical skills in LLM orchestration, retrieval-augmented generation, and Azure AI infrastructure to drive innovation in enterprise AI solutions.

## PROJECTS

**RAG Document Assistant** | FastAPI, Streamlit, LangChain, ChromaDB, AWS

*Production-ready RAG system enabling intelligent document Q&A with source citations.*

- Developed production-grade RAG system processing 1000+ documents with 89% answer accuracy using LangChain, ChromaDB, and GPT-5-nano.
- Architected microservices backend with FastAPI supporting 50+ concurrent users, reducing query latency by 40% through optimized chunking strategies.
- Implemented CI/CD pipeline with GitHub Actions, Docker containerization, and AWS ECS deployment achieving 99.9% uptime.

**Enterprise AI Agent Prototype on Azure AI Foundry** | Azure AI Studio, Prompt Flow, Azure OpenAI, Azure ML

*Developed a prototype for an enterprise-grade AI agent using the Microsoft Azure AI ecosystem.*

- Orchestrated multi-step AI workflows using Azure Prompt Flow for agentic automation.
- Deployed and managed Azure OpenAI models with responsible AI governance controls.
- Implemented RAG architecture with Azure AI Search for knowledge-enhanced responses.

**LLM-based Email Automation Agent** | Hugging Face, LLaMA-3.2, Llama.cpp, FAISS, DuckDB

*Collaborated in repurposing LLaMA model for intelligent email handling*

- Fine-tuned LLaMA-3.2-3B-Instruct and reduced model size by 85% using Llama.cpp for lightweight device deployment.
- Engineered a high-performance data layer with DuckDB, managing 1K+ records with 2x faster query response than SQLite.
- Implemented a semantic search system using FAISS, achieving sub-100ms email retrieval across 1K+ threads

**Sentiment Analysis on Edge Devices** | TensorFlow Lite, Keras, Android, Google Colab

*Transformer model optimized for edge deployment.*

- Quantized transformer model using TensorFlow Lite, gaining 90% model size reduction for mobile deployment.
- Delivered 80% faster inference with on-device sentiment analysis, removing cloud dependency.
- Optimized the training pipeline leveraging GPU acceleration in Google Colab, cutting iteration time by 40%.

**Predictive Modeling for Strategic Decision-Making** | Python (XGBoost, Pandas, Numpy, Scikit-learn)

*Built a machine learning model to forecast and mitigate customer churn for a telecom provider*

- Trained XGBoost classifier achieving 90% accuracy, improving retention strategies by 18%.
- Insights contributed to 25% reduction in churn.

## WORK EXPERIENCE

**AI Engineer Intern** | Gen AI Agency, London | 06/2024 – 12/2024

- Architected and developed LLM-based multi-agent systems, leveraging prompt engineering and OpenAI API, boosting productivity by 25% and enhancing coordination by 40%.
- Integrated AI solutions into products with 90% client alignment.
- Streamlined web data ETL using Python and Excel, cutting manual effort by 50% and raising efficiency by 30

**Full-stack Developer** | Sorphise Limited, Accra | 12/2022 – 06/2023

- Built scalable Django backend with optimised PostgreSQL queries, accelerating data retrieval by 40%.
- Developed a responsive front end using Bootstrap, enhancing user experience and increasing engagement by 20%.
- Implemented REST APIs with 99% success rate, ensuring reliable communication across services.

## EDUCATION

**M.Sc. Data Science** | Teesside University, Middlesbrough | 09/2023 – 05/2025 | **Distinction** (Achieved)

## SKILLS

**AI/ML Frameworks & LLMs:** Transformers, LLM Fine-Tuning (LLaMA, Hugging Face), TensorFlow/TF Lite, Keras, PyTorch, OpenAI API, LangChain

**ML Algorithms:** XGBoost, Random Forests, Neural Networks, Transformer Architectures, RAG Systems, Multi-Agent Orchestration

**Data & Vector Stores:** PostgreSQL, MS SQL, DuckDB, ChromaDB, SQLite

**Cloud & MLOps:** Microsoft Azure (AI, ML, DevOps), Git/GitHub, CI/CD (GitHub Actions), Docker

**Development:** Python, FastAPI, Django, REST APIs, Streamlit, Bootstrap

## CERTIFICATION

**Microsoft Certified:** Azure AI Engineer Associate | 09/2025