

DANIEL EGBO

AI ENGINEER | LLMs | MLOps | DATA PLATFORMS

Cape Town, WC, SA • +27784539717 • danoegbo@gmail.com • [LinkedIn](#) • [GitHub](#) • [Website](#)

SUMMARY

AI Engineer with experience designing, building, and deploying production AI systems across cloud and on-premise environments. I have experience in LLM fine-tuning, retrieval-augmented generation (RAG), conversational AI, speech AI, and MLOps. Skilled in developing end-to-end AI applications spanning data engineering, model development, workflow orchestration, cloud infrastructure, and production deployment using Python, PyTorch, Docker, Kubernetes, AWS, and GCP.

CORE SKILLS

Machine Learning

Model Development | Feature Engineering | Model Training | Model Evaluation | Hyperparameter Tuning | Predictive Modeling

LLMs & Generative AI

LLM Finetuning | RAG Systems | Prompt Engineering | Vector Databases | Embeddings | Agentic AI Systems

MLOps & Deployment

Docker | Kubernetes | CI/CD (Github Actions) | Model Serving | AWS (Sagemaker, ECR, Bedrock) | GCP (Vertex AI, BigQuery)

Deep Learning Frameworks

PyTorch | Scikit-learn | XGBoost | HuggingFace Transformers | Unsloth | NVIDIA Nemo Speech

Data & Distributed Systems

Python | SQL | Spark/Dask | Data Engineering | Distributed Computing

AI Applications

Chatbots | Speech AI | Computer Vision | NLP Systems | Recommendation Systems | Agentic RAG, Agentic Systems

AI Infrastructure

Langchain/LangGraph | Strands
Agents | Agno | NVIDIA AI Stack |
Vector Databases (FAISS, Pinecone, OpenSearch)

PROFESSIONAL EXPERIENCE

Machine Learning Engineer/Data Scientist

Consultant

Jan 2025 – Present

Cape Town, South Africa

- Built LLM fine-tuning and RAG systems for production AI applications using vector databases and embedding pipelines.
- Developed conversational AI and speech models, including multilingual and low-resource language systems.
- Designed and deployed ML pipelines using Docker, Kubernetes, and CI/CD across AWS and GCP.
- Curated large-scale text and speech datasets for model training and fine-tuning.
- Delivered end-to-end AI solutions from model development to cloud deployment and monitoring.

Machine Learning Engineer (Research Intern)

[Nabafat AI](#)

June 2023 – December 2024

Remote, US

- Built and optimised ML systems for NLP, computer vision, and LLM applications.
- Developed LLM fine-tuning pipelines (LoRA/PEFT) across multiple foundation models.
- Implemented RAG chatbot systems using vector databases and embedding pipelines.
- Deployed production ML systems using Docker, AWS ECR, and CI/CD workflows.
- Improved dataset quality and model performance through large-scale data engineering.

Research Data Scientist / Machine Learning Engineer

[SAAO](#)

February 2021 – September 2024

Cape Town, South Africa

- Built ML models for classification and cross-matching using 500k+ records and multi-catalogue datasets.
- Designed scalable pipelines for astronomical data processing and feature extraction.
- Developed reproducible ML workflows supporting scientific publications and research output.
- Applied ML techniques for large-scale pattern recognition in observational datasets.

Machine Learning Intern

[Ubenwa Health](#)

September 2020 – January 2021

Remote, Canada

- Built ML pipelines for clinical audio feature extraction and preprocessing.
- Developed and evaluated classification models using Scikit-learn and XGBoost.
- Improved dataset quality and model performance through feature engineering and cleaning.

SELECTED PROJECTS

SAAO Research RAG Analytics Systems

Python | APIs | ETL | RAG | MinIO | Postgres | Ollama | OpenSearch | Docker

[GitHub Repo](#)

- Built an end-to-end RAG system for scientific literature and catalogue analysis using LLMs and vector search.
- Designed ETL pipelines integrating external APIs, research papers, and structured astronomical datasets.
- Implemented a retrieval layer with OpenSearch and embeddings for semantic search over research data.
- Containerised the system with Docker, enabling reproducible deployment workflows in research environments.

Weather Disease Prediction Platform

Python | MLFlow | Prefect | DVC | Terraform | AWS | GCP | Docker | Kubernetes

- Built an end-to-end MLOps platform for predicting weather-sensitive diseases using machine learning.
- Implemented experiment tracking with MLflow, data versioning with DVC, and workflow orchestration with Prefect.
- Developed CI/CD pipelines with GitHub Actions and containerised deployment workflows using Docker.
- Provisioned cloud infrastructure on AWS and GCP using Terraform, supporting ECS Fargate, Cloud Run, and Kubernetes deployments.
- Integrated Evidently and Grafana for model performance monitoring and data drift detection.

GitHub Repo

Brain MRI Tumour Classification System

Python | TensorFlow | Keras | CNN | GCS | Computer Vision | Docker

- Developed a deep learning pipeline for brain MRI classification into multiple tumour categories using convolutional neural networks (CNNs).
- Implemented an end-to-end image preprocessing pipeline, including resizing, normalisation, and augmentation for medical imaging data.
- Designed and trained CNN architectures for multi-class classification of brain tumour types (glioma, meningioma, pituitary, and no tumour).
- Evaluated model performance using standard classification metrics and visualised the distributions of prediction confidence.
- Deployed a Streamlit-based web application for real-time inference and interactive medical image prediction.

GitHub Repo

EDUCATION		
University of Cape Town		
Doctor of Philosophy, Astronomy		Expected 2026
University of Nigeria, Nsukka		
Master of Science, Physics & Astronomy		2020
University of Nigeria, Nsukka		
Bachelor of Science, Physics & Astronomy		2014

PROFESSIONAL TRAININGS	
Machine Learning & MLOps - MLOps Zoomcamp (DataTalks.Club) - Machine Learning Zoomcamp (DataTalks.Club)	Large Language Models & Generative AI - Building RAG Agents with LLMs (NVIDIA) - Conversational AI Applications (NVIDIA) - Rapid Application Development with LLMs (NVIDIA) - Introduction to NVIDIA NIM Microservices (NVIDIA)
Deep Learning & Data Science - Getting Started with Deep Learning (NVIDIA) - Accelerating End-to-End Data Science Workflows (NVIDIA) - Fundamentals of Accelerated Data Science (NVIDIA)	

TECHNICAL HIGHLIGHTS
- Built production RAG and agentic AI systems using modern LLM frameworks. - Fine-tuned open-source foundation models using LoRA and PEFT. - Deployed AI workloads on AWS and GCP with Docker and Kubernetes. - Developed multilingual speech AI solutions and curated large-scale datasets. - Built end-to-end MLOps pipelines for training, deployment, and monitoring.

LEADERSHIP & COMMUNITY
- Arize AI Ambassador - Tavily AI Ambassador - BRICS/IDIA Data Analytics Trainer - Mentor, DARA Big Data Africa School - Mentor, SARAO Open Data for Social Impact Challenge