

Amos Bunde

Senior Machine Learning / MLOps Engineer | Production ML Platforms and Infrastructure

Nairobi, Kenya (open to relocation) | +254 713 131 604 | amos_bunde@yahoo.com

linkedin.com/in/amos-bunde | iam-amos.io | github.com/AmosBunde

SUMMARY

Machine learning engineer with 12 years designing, deploying, and operating production ML systems on AWS, GCP, and Azure. Build end to end: data ingestion and feature pipelines, model training, and real-time and batch serving, with the MLOps automation and observability that keep systems reliable in production. Took a clinical AI system from prototype to production with containerized serving, GitHub Actions CI/CD, and an evaluation and monitoring pipeline that caught a silent failure mode; shipped a LightGBM system on SageMaker behind live APIs that reduced program cost by 60 percent. Hold an MSc in Computing and Information Systems, with a dissertation on optimizing compute workloads for cost and energy efficiency. Mentor engineers and set engineering standards while staying hands-on in Python and SQL.

TECHNICAL SKILLS

- **ML and MLOps:** end-to-end ML systems, training pipelines, model serving, deployment workflows, model observability and monitoring, evaluation harnesses, CI/CD (GitHub Actions), MLflow, SageMaker
- **Inference and serving:** real-time and batch inference, FastAPI, gunicorn, containerized model serving
- **Cloud and infrastructure:** AWS (SageMaker, EC2, Lambda, S3, Glue, Redshift, Step Functions), GCP (Vertex AI, Dataflow, BigQuery), Azure; Docker, Kubernetes, Terraform, CloudFormation
- **Data engineering:** PySpark, Spark, Kafka, Flink, Airflow, dbt, feature and ingestion pipelines, ClickHouse
- **Languages and ML:** Python (production depth), SQL (expert), scikit-learn, PyTorch, TensorFlow, LightGBM
- **Optimization and governance:** cost and latency optimization, GPU and CUDA parallel computing, data sovereignty, security and compliance for regulated data

PROFESSIONAL EXPERIENCE

Aga Khan University

Sep 2022 – Present

Senior → Lead → Manager, ML and Data Infrastructure Engineering

Distributed teams, Kenya and Pakistan

- Design and operate end-to-end ML systems: data ingestion and feature pipelines, model deployment and serving, monitoring, and CI/CD, across research and clinical use cases.
- Took Afya Gemma, a clinical AI system, from prototype to production with containerized serving (Docker Compose, gunicorn), GitHub Actions CI/CD, and a ChromaDB vector store (6,193 chunks, 1024-dimensional embeddings); Google GenAI Accelerator Award.
- Built the evaluation and observability pipeline that surfaced a silent topic-mismatch failure mode after four version iterations, then added the monitoring and escalation paths that made daily production use reliable.
- Built multi-country ingestion including real-time Fitbit streaming for 615 healthcare workers, with explicit handling of partner, network, and platform failure modes.
- Mentor engineers and set architectural, release, and review standards across AWS, Azure, and GCP; identified and developed two technical leads now owning components of the platform.

Zubale

Oct 2021 – Jul 2022

Senior Big Data and ML Engineer (Contract)

Mexico City (Remote)

- Re-engineered real-time streaming and batch pipelines on GCP (Kafka, Pub/Sub, Dataflow, GCS, BigQuery), reducing cloud data service cost by approximately 25 percent while improving throughput and reliability.
- Built automated ML infrastructure and the production data stack (Airflow, dbt, Looker, Superset), with DataOps practices that cut deployment lead time and improved release reliability.

Copia Global

Feb 2020 – Aug 2022

Senior Data Engineer → Data Platform Manager (AWS)

Nairobi

- Built and deployed a LightGBM recommendation system on AWS SageMaker behind live referral APIs (real-time inference), reducing referral program cost by 60 percent.
- Built an AWS batch pipeline (Lambda, S3, Glue, Spark, Redshift, Step Functions, SNS) processing 2.5M-plus rows per day with change data capture, monitoring, and alerting so failures were caught before downstream teams noticed.

- Established the data engineering function, release practices, and MLOps standards from the ground up; separated analytical and transactional workloads to eliminate a recurring source of downtime.
- Zealtech Data Solutions**
Senior BI and Analytics Consultant

May 2019 – Jan 2020
Nairobi
- Delivered ML models, dashboards, and customer 360 data marts; integrated new sources to strengthen analytics accuracy.
- Earlier Roles**
Data and Software Engineering

2014 – 2019
Cropnuts, Skytop, KNEC, Redbric
- Built an ML calibration model for soil analysis at Cropnuts that expanded access across Kenya, Senegal, and Nigeria, working with US and Germany engineering teams; earlier roles built the foundation in production data work.

SELECTED PROJECTS

- Afya Gemma** (Google GenAI Accelerator Award): production clinical AI with two-stage RAG, ChromaDB, FastAPI, Docker, GitHub Actions CI/CD, and an evaluation and monitoring harness.
- Agent Arena**: open-source LLM agent evaluation platform with cost-per-correct-answer metrics and deterministic replay, built for measurable model observability.
- Grounding-Conditional Evaluation**: arXiv technical report (cs.CL) on measuring honest abstention in retrieval-augmented systems, applying evaluation rigor to a production failure mode.

EDUCATION

MSc, Computing and Information Systems | Liverpool John Moores University
Dissertation: reinforcement learning and Bayesian optimization for HPC workload scheduling and energy and cost efficiency.

BSc, Applied Statistics with Computing | University of Kabianga

Awarded Feb 2026

CERTIFICATIONS

Data Engineering on GCP | AWS Cloud Practitioner | GPU Programming Using CUDA, University of Stuttgart | Advanced Parallel Programming with MPI and OpenMP, University of Stuttgart | Human-Centered Generative AI, Stanford