ASHTONE ONYANGO

AI Engineer | LLM Training Systems (RL)

Nairobi, Kenya | Tel: +254740497975
ashtone@wanailabs.org | LinkedIn | GitHub | My Portfolio

PROFESSIONAL SUMMARY

Machine Learning Engineer with 6+ years designing, training, and deploying production AI systems, including transformer models trained from scratch and fine-tuned with reinforcement learning (RLHF/SFT). Experienced in building RL feedback environments, designing reward functions, and implementing LLM evaluation pipelines. Proven ability to read ML research papers and convert them into working implementations. Strong command of PyTorch, Docker, distributed training, and MLOps; with hands-on exposure to model internals, including attention mechanisms, tokenizers, and inference optimization.

SKILLS AND COMPETENCIES

- **LLM Internals**: Attention Mechanisms (MHA, FlashAttention), KV Caches, Tokenizers, Infer
- **Frameworks**: PyTorch (from-scratch training & fine-tuning), HuggingFace Transformers, TensorFlow, JAX (familiar), Triton, CUDA Kernels (exposure)
- **LLM Internals**: Attention Mechanisms (MHA, FlashAttention), KV Caches, Tokenizers, Inference Optimization, Distributed Training (DDP), Hyperparameter Tuning
- **MLOps & Infra**: Docker, Kubernetes, AWS SageMaker, GCP Vertex AI, CI/CD (GitHub Actions, Jenkins), Reproducible Environments, Git
- **Languages**: Python (advanced), JavaScript/Node.js, SQL, Bash.
- **Data & Pipelines**: Apache Spark, Kafka, Airbyte, BigQuery, PostgreSQL, Pinecone, FAISS, Redis
- **Research & Eval**: ML Paper Implementation, LLM Evaluation Design, STEM Problem Engineering, Ground-truth Solution Verification, Automated Scoring Pipelines
- **Certifications**: AWS ML Engineer (Udacity), IBM AI Analyst Mastery Award, Microsoft IoT/AI, Data Engineer (DataCamp), Full-Stack Developer (Udacity).

WORK EXPERIENCE

iMerit Technology, Remote (US) Nov 2025 – March 2026

RL Environment & LLM Evaluation Engineer (Contract)

- Design and build high-precision RL environments and annotation workflows for GenAI training, with a focus on complex reasoning, STEM problem engineering, and LLM boundary testing.
- Engineer original, computationally intensive STEM and data science problems simulating real-world scientific workflows, directly analogous to designing challenging RL tasks for LLM training.
- Develop non-trivial reasoning chains and 'fair but hard' evaluation cases to test GPT-4, Claude, and Gemini, identifying failure modes and reward-hacking vectors.
- Configure automated scoring criteria and verify ground-truth solutions using Python (NumPy, Pandas) and SQL, implementing judge logic equivalent to continuous RL reward signals.
- Deploy and manage reproducible Docker environments and CI/CD pipelines on AWS/GCP for delivery of high-fidelity training datasets.
- Serve as a link between expert annotators and AI systems, improving workflows and strategies to create high-quality RL training data.

Wan AI Labs, Nairobi, Kenya Oct 2024 – Sept 2025

Lead AI Engineer (Agentic Systems & LLM Deployment)

- Led AI division building production-grade agentic systems using LLMs (GPT-4, Claude, Gemini), with fine-tuned models deployed across healthcare, legal, and enterprise operations.
- Designed and deployed 20+ AI systems using LangChain and CrewAI, automating complex workflows, resulting in 40% productivity gains and 60% reduction in manual operations.
- Oversaw integration with enterprise systems via GraphQL APIs and Kafka pipelines, managing the full lifecycle from model selection through production deployment and monitoring.
- Facilitated iterative model improvement cycles with structured human feedback — practical experience directly applicable to RLHF pipeline design and reward model training.

Rightsify Group LLC, Pasadena, California (Remote). Jun 2023 – Aug 2024

AI Engineer (Audio Generative LLMs)

- Architected and trained custom transformer LLMs from scratch for music generation, serving thousands of daily users, building full training pipelines including data preprocessing, model architecture design, and evaluation.
- Used reinforcement learning during model training by applying feedback loops and rewards to make the output better, resulting in a 20% improvement in model performance.
- Designed distributed training pipelines on AWS with PyTorch and HuggingFace, reducing training time by 60% while processing 1M+ audio samples through custom preprocessing workflows.
- Created tools to continuously check and analyze the model's performance, similar to how feedback works in reinforcement learning.
- Implemented CI/CD for ML model deployment with Docker and Kubernetes, reducing deployment time from 4 hours to 15 minutes in reproducible, containerized environments.
- Automated dataset migration pipelines (AWS S3 → GCP) for 10TB+ datasets, establishing scalable MLOps infrastructure for model serving across Paperspace and DigitalOcean.

ZURI HEALTH, Nairobi, Kenya. Jan 2023 – April 2023

Software Developer

- Collaborated with platform engineers to design analytics pipelines that extract actionable insights from multi-country user interaction data.
- Built a conversational AI triage system using NLP and RAG architecture, serving 5,000+ users, achieving 92% diagnostic accuracy, and reducing query resolution time by 30%.
- Created a way to search through medical information that made answers more relevant by 40% using improved methods for finding data.

Google Crowdsource, Nairobi, Kenya. Aug 2021 – Present

AI Facilitator/Trainer - Part-time

- Trained participants on AWS SageMaker and GCP Vertex AI for deploying production ML models.
- Facilitated 11+ hands-on ML workshops covering LLM fine-tuning, NLP, and cloud ML deployment for 33+ university students across Africa.
- Managed community of 5,000+ AI contributors; trained participants on AWS SageMaker and GCP Vertex AI for deploying production ML models.

ANALYTICS VIDHYA, Gurgaon, India Sep 2022 – Dec 2022

Technical Writer – Data (Seasonal)

- Wrote and published 3 tutorial articles for the monthly data science blogathon on creating ETL pipelines for over 2000 rows of data extracted from OLAP and OLTP databases, which attracted over 50 readers and users to the platform.
- Implemented application backends using Flask to integrate database APIS that enabled interaction with remote Postgres database servers hosted both locally and on the cloud.

Upwork, Nairobi, Kenya Feb 2021 – Sep 2022

Machine Learning Engineer - Freelance

- Delivered 33+ custom AI solutions leveraging AWS and Azure ML services, implementing semantic search and NLP systems for diverse client needs.
- Built production-ready GraphQL APIs with Express and React frontends, integrating LLM capabilities for intelligent data processing.
- Developed distributed data processing pipelines using Apache Spark that improved data ingestion speed by 65%.
- Created automated reporting systems in Tableau that saved clients 10+ hours weekly through intelligent data visualization

TEENS IN AI, London, UK.

Machine Learning Trainer Aug 2019 – Sep 2019

- Coached 5 teen girls out of 10 teams in Descriptive and Predictive Analytics using Azure Machine Learning visual tools.
- Guided approximately 10 students on using Azure Machine Learning to analyze access to affordable healthcare by low-income households in 5 countries across Sub Sahara Africa.

EDUCATION

KENYATTA UNIVERSITY, Nairobi, Kenya

Bachelor of Science, Biomedical Engineering

Relevant Coursework:

- Health Information Systems, Bioinformatics, Programming; Electronics; Database Systems.

Certifications:

- AWS Machine Learning Engineer Nanodegree; Udacity.
- Data Engineer Track; DataCamp.
- Full-Stack Developer Nanodegree; Udacity.
- Information Security Certification; FreeCodeCamp.
- IBM AI Analyst Mastery Award & Explorer Award (2019)
- Microsoft IoT & AI Certifications.

KEY PROJECTS

Pneumonia Detection & Monitoring in Infants

Huawei Global Innovator Award - **Johannesburg, South Africa - April 2022**

- Trained a neural network from scratch on 5,000+ rows of IoT sensor data to predict infant pneumonia risk, achieving 80% accuracy. Awarded at Huawei Global Innovator Summit, Johannesburg.

TAF: AI Caregiver Support Ecosystem [Link]

- Custom RAG architecture over a medical and legal knowledge base serving 1,000+ caregivers across Sub-Saharan Africa with 95% information accuracy. Demonstrates domain-specific LLM fine-tuning and production evaluation pipeline design.

CDIE Smart Lab: LLM-Powered SOP Automation [Link]

- LLM document processing system automating understanding and assessment of Lab SOPs for 10,000+ university users, achieving 75% faster lab preparation. Judge-style correctness evaluation at its core.

Kujia Jobs: Semantic Job Matching Engine [Link]

- Custom NLP semantic search engine matching 5,000+ job seekers to opportunities with 90% accuracy, handling 1,000+ daily queries on AWS Lambda with collaborative filtering.

ACTIVITIES

Google Developers Student Club, Nairobi, Kenya
Machine Learning Lead Sep 2020 – Aug 2021

- Led 3 Machine Learning competitions for over 40 students, increasing their Cloud Analytics skills in GCP and TensorFlow.