

ERIC PETER WAIRAGALA

Machine Learning / Research Engineer · Inference & Production ML Systems

Kampala, Uganda — Open to Remote & Relocation | kigaye.ericpeter@gmail.com | linkedin.com/in/eric-peter-wairagala | ericpeter.github.io/io
| github.com/EricPeter | @ericpeterwepson

SUMMARY

NLP research engineer who bridges model training and production deployment — from published research on low-resource African languages to scaling a real-time multilingual ASR pipeline **100x** (8 → 800 concurrent live-streaming connections) at P99 latency under 500ms, using TensorRT-LLM, ONNX, and NVIDIA Triton across T4/A10/L4 GPUs. The system now serves **35M+ minutes** of production transcription and translation. Comfortable owning the full lifecycle: model training and fine-tuning, evaluation, deployment, and **production monitoring for drift and safety**. Practices TDD and ships through CI/CD pipelines (Concourse, ArgoCD, GitHub Actions). Strong collaborator across research and engineering teams — co-author on **10+ peer-reviewed NLP publications** with the Masakhane community, including MasakhaNER and InkubaLM. Growing focus on production model monitoring, safety evaluation, and drift detection, alongside hands-on use of AI coding agents to accelerate development. M.Sc. Computer Science (AI & Data Science, completed 2025).

TECHNICAL SKILLS

- **ML & research:** PyTorch 2.x, TensorFlow, HuggingFace Transformers, PEFT/LoRA/QLoRA/SFT, transformer architectures, fairness evaluation, multilingual & low-resource NLP, RAG, Whisper/ASR
- **Inference & serving:** TensorRT-LLM, ONNX Runtime, NVIDIA Triton Inference Server, continuous batching, KV cache optimization, FP8/INT8 quantization, model parallelism, torch.compile
- **Evaluation & experimentation:** MLflow, Weights & Biases, latency/quality regression testing, TTFT/throughput/GPU-utilization tracking, benchmark design
- **Production monitoring & safety:** model drift detection, safety/quality evaluation for deployed models, Prometheus/Grafana dashboards, latency & regression alerting
- **Engineering practices:** test-driven development (TDD), CI/CD with Concourse and ArgoCD, GitOps deployments, code review, cross-functional collaboration between research and engineering teams
- **AI-assisted engineering:** Claude Code, GitHub Copilot, Cursor — agentic coding workflows for rapid prototyping, code review, and test generation
- **Systems & APIs:** WebSocket/SSE low-latency streaming, gRPC, FastAPI, high-concurrency async Python, Prometheus/Grafana, canary deployments, zero-downtime rollouts
- **Infra:** AWS (EC2, S3, SageMaker), GCP (Vertex AI), Azure; NVIDIA T4/A10/L4/A100; Docker, Kubernetes, Helm; Concourse CI, ArgoCD, GitHub Actions, Jenkins
- **Languages & data:** Python (systems-level & async), SQL, Bash; PostgreSQL, MySQL, MongoDB, Elasticsearch

EXPERIENCE

Research Engineer — *Lelapa AI, Johannesburg, South Africa (Remote)*

2024 – Present

- **Scaled inference capacity 100x:** redesigned the real-time multilingual transcription/translation stack with continuous batching and KV cache optimization, taking live-streaming capacity from 8 to 800 concurrent connections while sustaining P99 end-to-end latency under 500ms on T4/A10/L4 GPUs.
- **Cut inference cost and latency:** optimized Whisper ASR with TensorRT-LLM (FP8/INT8 quantization) and ONNX export, served via NVIDIA Triton — now powering 35M+ minutes of production transcription and translation.
- **Built EvalKit:** an internal evaluation framework with automated smoke tests and performance benchmarks (TTFT, throughput, GPU utilization), wired into Concourse CI to catch inference regressions before deployment.
- **Monitored models in production:** tracked drift and quality/safety signals on deployed models, closing the loop between training-time evaluation and live behaviour.
- **Shipped production streaming APIs:** designed high-concurrency WebSocket/REST services in FastAPI, containerized for low-latency audio delivery to production clients.
- **Operated multi-cloud GPU infrastructure:** across AWS, GCP, and Azure using Concourse and ArgoCD for GitOps-style CI/CD, with canary deployments and zero-downtime model rollouts, maintaining production SLOs under live traffic with TDD.
- **Worked closely with cross-functional teams:** partnered with research, product, and infrastructure engineers to bridge the gap between model training and deployment.
- **Built a repeatable sandbox-to-production promotion pipeline:** designed a self-service sandbox tool that lets researchers train and validate models against evaluation gates — concurrency/performance and quality — with models automatically promoted to staging once they pass, replacing ad-hoc handoffs with a fast, repeatable path from research to production.

Research Engineer — *Makerere University AI & Data Science Lab, Kampala, Uganda*

2022 – 2024

- Fine-tuned multilingual transformer models (LoRA/SFT) on low-resource African-language datasets, collaborating closely with the Masakhane research community across multiple countries and institutions.
- Applied fairness evaluation frameworks to reduce bias in NLP models for under-represented languages.
- Built end-to-end experiment pipelines (PyTorch/TensorFlow) with MLflow tracking; shipped an AI-powered job-screening tool (scikit-learn, PostgreSQL).

Software Engineer — *Uganda Cancer Institute, Kampala, Uganda*

2022 – 2023

- Built a Flutter/Firebase cancer-screening system supporting patient data collection and clinical decision workflows across multiple hospital sites.

Data Engineer — EzyAgric, Kampala, Uganda

2021 – 2022

- Built ETL pipelines (Apache Spark, AWS Redshift) for large-scale agricultural datasets powering data-driven decisions across East African farming operations.

EDUCATION

M.Sc. in Computer Science — Artificial Intelligence & Data Science

Makerere University, Kampala, Uganda | 2022 – 2025 | Completed

B.Sc. (Hons) in Software Engineering

Makerere University, Kampala, Uganda | 2016 – 2021

SELECTED PUBLICATIONS

Co-author on 10+ peer-reviewed NLP papers with the Masakhane research community. Full list: scholar.google.com/citations?user=al9kLwwAAAAJ

- **Gender Bias Evaluation in Luganda-English Machine Translation** — EP Wairagala et al., *AMTA* 2022 (First author)
- **Building Luganda Machine Translation Models for the Medical Domain** — EP Wairagala, J Nakatumba-Nabende, R Nakasi (First author)
- **InkubaLM: A Small Language Model for Low-Resource African Languages** — AL Tonja, ..., EP Wairagala, et al., *arXiv* 2024
- **MasakhaNER: Named Entity Recognition for African Languages** — DI Adelani et al., *Transactions of the ACL* 2021 (244 citations)
- **MasakhaNER 2.0: Africa-Centric Transfer Learning for NER** — DI Adelani et al., *EMNLP* 2022 (105 citations)
- **A Few Thousand Translations Go a Long Way!** — DI Adelani et al., *NAACL* 2022 (109 citations)
- **Afri-MCQA: Multimodal Cultural Question Answering for African Languages** — AL Tonja et al., *ACL* 2026
- **Amplify Initiative: Building a Localized Data Platform for Globalized AI** — QM Rashid et al., *ACM FAccT* 2026