William Miller Juma

Senior Software Engineer · jumamiller@yahoo.com · +254748730956 · https://www.millerjuma.co.ke

PROFESSIONAL SUMMARY

Senior Software Engineer with 8+ years of experience across Microsoft, high-growth startups, and multinational teams, specializing in performance engineering and distributed systems at scale. Led teams of 10+ engineers and delivered results including 70-80% infrastructure cost reductions, 85% fraud reduction, and 99.9% system uptime. Currently pursuing an MSc in Computer Science while building distributed LLM inference systems achieving 2.5x throughput gains on multi-GPU clusters.

EDUCATION

University of East London London, United Kingdom
MSc. Computer Science

Jomo Kenyatta University Nairobi, Kenya
BSc. Computer Technology
Thesis: Elliptic Curve Cryptography Based Secure Mobile Banking Application with Integrated Distress PIN (Dec 2022)

WORK EXPERIENCE

Microsoft Redmond, United States
Software Engineer *Sep 2024 - Present*

- Built an automated regression-triaging platform that analyzes benchmarking failures and posts diagnostic comments, enabling engineers to resolve issues in under two minutes and cutting manual triage time by 98%.
- Architected a centralized benchmarking data platform consolidating millions of performance records across multiple labs, reducing query and data-access time by 20% for 5+ engineering teams.
- Engineered Azure Data Factory pipelines to ingest benchmarking data from Azure Blob Storage, clean and transform records, and load databases for downstream performance analysis, maintaining 99.99% pipeline uptime.
- Led development of a benchmarking toolkit supporting 16 industry-standard benchmarks, establishing repeatable performance evaluation across Windows systems and reducing evaluation time by 80%.
- Expanded access to benchmarking data by developing an MCP server that enabled 20+ engineering teams to query performance datasets through AI-compatible interfaces, cutting data lookup time by nearly 40%.
- Integrated automated performance-regression detection into Windows build pipelines as release-blocking quality gates, reducing post-release bugs by 10%.
- Drove data-driven Windows optimization: analyzed device telemetry to pinpoint performance-critical OS components, then applied Sample-based Profile-Guided Optimization (SPGO) to optimize code paths shipping to 1B+ devices.
- Built a centralized Azure cost-monitoring platform adopted by 10+ engineering teams, improving cost visibility and identifying 40% in savings opportunities.
- Benchmarking SME across CPU, GPU, AI/ML, and browser workloads (MLPerf, Procyon AI, Geekbench AI) on Windows and Linux; designed repeatable test methodology and used WPR/WPA trace analysis to root-cause up to 15% Windows–Linux gaps in Windows internals, guiding platform fixes.
- Developed a Windows trace-escalation platform that lets engineering teams collect diagnostic traces from devices worldwide, reducing investigation time by 30%.
- Implemented remote benchmark scheduling so engineers can run suites on remote devices without owning hardware, saving 20 hours per week across the team.

Griffin Global Technologies Atlanta, Georgia, USA
Senior Software Engineer *Jan 2024 - Sep 2024*

- Architected and delivered an internal Resource Utilization Management System in .NET for employee-to-project allocation and organization-wide resource planning.
- Replaced a licensed planning stack with the internal platform, owning rollout and driving about $100K in annual cost savings after the SAGE Intacct cutover.
- Modeled staffing, project assignment, and utilization in a single system so planning stopped living in disconnected tools and spreadsheets.

- Owned end-to-end delivery of the platform: requirements with ops and finance, technical design, implementation, and production release.
- Ran design and code reviews, enforced standards, and kept the codebase maintainable so the team could ship on time without silent quality debt.
- Mentored interns and junior engineers on .NET design and delivery; their work landed in production allocation and planning features.

Solutech Limited Nairobi, Kenya
Senior Software Engineer *Mar 2023 - Jan 2024*

- Moved core sales automation from per-client Digital Ocean VMs to a multi-tenant GKE platform, cutting infrastructure cost 70-80% with consolidation and autoscaling.
- Upgraded production Laravel from 5 to 10 with zero downtime, closed security gaps, improved performance ~40%, and left a maintainable codebase.
- Published shared Node.js packages on NPM for auth, utilities, and UI, cutting duplication ~60% and speeding delivery ~50%.
- Launched a real-time ticketing portal for 200+ clients, replacing WhatsApp support and cutting resolution time ~65%.
- Owned production reliability with customer service, shipped targeted fixes, and cut recurring bugs ~50%.

ViewTech Limited (SasaPay) Nairobi, Kenya
Software Engineer *Oct 2022 - Mar 2023*

- Built the AML engine that automated onboarding, KYC, and risk scoring, cutting manual review time over 70% while keeping the compliance bar.
- Architected and launched the Developer Portal from scratch - API docs, sandbox, and integration tools - so third-party developers could integrate without a custom handoff.
- Delivered end-to-end CRM integrations that synced customers, transactions, and communications across systems and cut data errors at source.
- Served as the technical escalation point for developers and partners, unblocking integrations and production issues on live payment traffic.
- Kept high-volume payment operations stable by owning incident response and hardening the paths partners and internal teams depended on.

Kinetic Technology Limited Nairobi, Kenya
Software Engineer *Jan 2022 - Jun 2022*

- Led the KISEB and EBK exam platforms from build through production so candidates could register, sit, and get results online.
- Cut board admin overhead ~70% and result turnaround ~60% by moving those flows off manual processes.
- Integrated Laravel with Dynamics 365 Business Central so exam and operations data posted into accounting without double entry.
- Implemented CRM flows for the boards to track candidates, payments, and case history in one place.
- Built ticketing so board staff could log, assign, and close candidate and operations issues instead of handling them ad hoc.

Sanifu AI Nairobi, Kenya
Software Engineer *Aug 2018 - Dec 2021*

- Led ML for automated onboarding, KYC, and fraud detection; cut fraudulent accounts ~85% and shortened onboarding from 3 days to 30 minutes.
- Built USSD loan origination on mobile-money rails for thousands of daily applications, including for non-smartphone users in rural Kenya.
- Built and scaled the Ajira Digital job portal to peak loads above 100,000 concurrent users for a national youth-employment program.
- Built data pipelines to move operational data between systems, supporting lending workflows and downstream analytics.

- Delivered real-time loan dashboards and approval analytics, turning operational data into actionable insights and reducing manual processing time by ~70%.
- Productionized the risk models into the loan workflow so verification and fraud checks ran in the origination path, not as a side notebook.

AI PROJECTS

CoRAG - Co-trained Retriever-Generator RAG github.com/jumamiller/CORAG
PyTorch, Hugging Face Transformers, PEFT, TRL

- Jointly fine-tuned a RAG reranker (bge-reranker-v2-m3, LoRA) and generator (Llama-3.1-8B-Instruct, 4-bit QLoRA) in alternating rounds, using generator answers as the reranker's training signal.
- Trained the generator with GRPO reinforcement learning and a custom answer-match reward; evaluated on the PopQA open-domain QA benchmark.

Multilingual Customer Support Chatbot github.com/jumamiller/customer-support-chatbot
Mistral-7B, LoRA, FAISS, LangChain

- Built a RAG support agent for East African businesses that answers English, Kiswahili, and Sheng queries from company FAQs, using multilingual-e5-large embeddings and FAISS retrieval.
- LoRA fine-tuned Mistral-7B-Instruct for a brand-aligned support tone, running fully on local hardware.

RAG From Scratch github.com/jumamiller/RAG
Azure OpenAI, FAISS, LangChain, LangGraph, MLflow

- Implemented a framework-free RAG pipeline (loading, chunking, embedding, FAISS vector search, grounded generation with source citations), then rebuilt it in LangChain to compare approaches.

Multi-Agent Deliberation Engine github.com/jumamiller/my-agent-playground
Python, Azure OpenAI, Prompt Engineering

- Designed a Proposer/Critic/Summarizer agent system that turns vague feature requests into structured decision documents through multi-round debate with an explicit termination contract.

SKILLS & COMPETENCIES

Data Engineering: ETL Pipelines, Data Ingestion, Data Cleaning & Transformation, SQL, Python, Pandas, Azure Data Factory, Azure Blob Storage, Azure Data Lake Storage, Azure Databricks, Apache Spark / PySpark, Azure Synapse Analytics, Microsoft Fabric, Apache Airflow, Apache Kafka, Azure Event Hubs, Delta Lake, PostgreSQL, SQL Server / Azure SQL, MySQL, Snowflake, BigQuery, Amazon Redshift, Power BI

Systems & Infrastructure: Performance Engineering, Windows Internals, Windows Performance Recorder (WPR), Windows Performance Analyzer (WPA), Debugging, Kubernetes, Docker, Terraform

AI Systems & Machine Learning: Supervised & Unsupervised Learning, Feature Engineering, Model Evaluation, Scikit-learn, Pandas, NumPy, Deep Learning, Neural Networks, Transformers, PyTorch, TensorFlow, LLMs, Fine-tuning, Quantization, Flash Attention, DeepSpeed, Hugging Face Transformers, Reinforcement Learning (RL), RLHF, LangChain, LlamaIndex, Vector Databases, RAG, Agents, Tool Use, MLOps, Experiment Tracking, Model Serving, LLM Evaluation & Benchmarking

Programming Languages: C, C++, Rust, Go, C#, Python, Java, JavaScript/TypeScript, SQL, PHP, React, Angular, Laravel, Node.js, Vue JS, Tailwind, Bootstrap

Cloud & Backend: Azure, Google Kubernetes Engine (GKE), PostgreSQL, Redis, gRPC, RESTful APIs, Microservices, .NET

PUBLICATIONS

Miller Juma. *Elliptic Curve Cryptography Based Secure Mobile Banking Application with Integrated Distress PIN.* BSc Thesis, Jomo Kenyatta University of Agriculture and Technology, December 2022. DOI: 10.13140/RG.2.2.13504.90888 - Advisor: Steven Kung'u

URL: https://www.researchgate.net/publication/412700535_...