Samuel Abolo

Agentic AI Engineer | Backend Systems | LLM Infrastructure
ikabolo59@gmail.com | +234 706 379 6022 | Lagos, Nigeria (Open to Relocation) |
linkedin.com/in/samuelabolo-24431a176

PROFILE

Senior AI Engineer with 5+ years shipping production LLM systems and AI agent backends. Currently at Boostr, architecting multi-agent orchestration pipelines on GCP using Google ADK, Vertex AI, and Gemini. Expert in production RAG with pgvector, structured outputs, evals, and cost-aware AI infrastructure. Track record: 99.5% reduction in deployment time and 360x query runtime improvement at TripAdvisor.

SKILLS

Agentic AI: Google ADK, multi-agent orchestration, tool-calling, structured outputs, memory management, LangChain, LangGraph

LLMs & Retrieval: GPT-4o, Gemini, LLaMA, RAG pipelines, pgvector, FAISS, embeddings, semantic search, prompt engineering

Backend & APIs: Python (FastAPI, AsyncIO, Pydantic), Go, Node.js, REST, gRPC, WebSockets, Kafka, Redis, MQTT

Cloud & MLOps: GCP (Vertex AI, Cloud Run), AWS, Kubernetes, Docker, Terraform, GitHub Actions, CI/CD automation

ML Infrastructure: Ray Serve, Seldon, Snowflake, Anyscale, Prometheus, Grafana, Loki, CloudWatch, distributed compute

EXPERIENCE

Boostr *Nov 2025 - Present*

Senior Software Engineer, Agentic AI Platform *New York, USA (Remote)*

- Architecting multi-agent orchestration pipelines (Google ADK, Vertex AI, Gemini) for automated AdOps decision-making on GCP; shipped to production within the first week on role
- Built answer routing logic with tool-calling, structured outputs, and confidence thresholds to determine when agents respond automatically vs. escalate to human review
- Designed production RAG system using pgvector (PostgreSQL) for semantic retrieval across large customer artifact stores, with full latency and cost monitoring
- Engineered scalable cloud-native ingestion and indexing pipelines for high-volume artifact processing on containerized GCP infrastructure
- Implemented end-to-end observability and cost-aware deployment patterns on Kubernetes; wrote evals to continuously validate agent behavior in production

TripAdvisor (via Zazmic Inc) *Oct 2024 - Nov 2025*

Software Engineer II, AI Platforms *USA (Remote)*

- Reduced AI service deployment time from 2 weeks to under 30 minutes (99.5% reduction) via full CI/CD automation and containerized Kubernetes infrastructure
- Migrated ML pipelines from Spark on EKS (Kubeflow) to Snowflake and Ray on Anyscale, cutting query runtime from 4 hours to 40 seconds (360x improvement)
- Reduced AI inference latency by 90%, cutting median execution time from 15 minutes to 62 seconds
- Built agent evaluation tooling and load-testing framework (Java/Spring) to stress-test AI microservices under high concurrency, turning production failures into durable fixes
- Implemented production ML observability using Prometheus, Grafana, Loki, and CloudWatch; served as DRI for on-call incident response across AI platform

NovaTrack *Sep 2025 - Nov 2025*

Software Engineer, Real-time Systems (Contract) *Lagos, Nigeria*

- Architected a high-throughput vehicle telematics platform in Go for a fleet of 10,000+ vehicles, processing live telemetry every 30 seconds per vehicle with strict SLA guarantees
- Built a real-time behavioral anomaly detection system flagging geofence violations, unauthorized route deviations, and reckless driving patterns, surfacing security alerts to live operational dashboards

- Designed event-driven ingestion pipelines using MQTT and WebSockets to handle continuous high-frequency sensor and GPS streams across the entire fleet
- Built distributed microservices with gRPC for low-latency inter-service communication across telemetry ingestion, storage (PostgreSQL + Redis), and real-time analytics layers
- Implemented predictive maintenance risk scoring on streaming sensor data, reducing unplanned vehicle downtime across the fleet

Credrails *Apr 2024 - Sep 2024*
Software Engineer, Reconciliation & Finance *Nairobi, Kenya (Remote)*

- Built an LLM-powered transaction intelligence system that analyzed unstructured transaction descriptions to detect discrepancies across accounts, identify anomalous patterns, and auto-generate classification rules at scale
- Engineered the full classification pipeline end-to-end: LLM inference for pattern extraction, structured output parsing, regex generation and validation, and deep integration into the core reconciliation engine
- Developed automated discrepancy detection logic that cross-referenced multi-source financial records to flag mismatches in real time, reducing manual reconciliation overhead
- Built client onboarding platform backend automating document verification and account setup, cutting onboarding time by 70% via workflow orchestration with Django and DRF

Quibble *Nov 2022 - Aug 2023*
Software Engineer, ML Engineering *Puerto Rico (Remote)*

- Built a TensorFlow-based pricing recommendation system forecasting optimal nightly rates across a 365-day horizon for short-term rental properties; deployed to production with full CI/CD and monitoring
- Developed computer vision pipeline for room-type classification and image attractiveness scoring using transfer learning on ImageNet architectures, directly influencing listing quality signals
- Designed and owned ETL pipelines for large-scale property data ingestion, cleaning, and feature engineering feeding both pricing and CV models
- Mentored a PhD-level junior engineer on model deployment, testing, and production performance tuning, bridging the gap between research and engineering

SELECTED PROJECTS

Agentic RAG System (ExcelMind) End-to-end retrieval pipeline using NLP models, vector search, and structured output generation for contextual exam explanation at scale. Designed data pipelines for training, evaluation, and deployment of NLP-driven exam pattern prediction models.

Zero-to-One AI Deployment Automation (TripAdvisor) Built an internal CLI tool that automated the full lifecycle of deploying new AI inference services on EKS and AWS from scratch. Given a service spec, the tool auto-generates production-grade Terraform and Helm charts, provisions infrastructure, and raises a fully-reviewed PR — cutting deployment time from 2 weeks to under 30 minutes (99.5% reduction) and eliminating error-prone manual ops for every new model rollout.

Real-time Fleet Intelligence Platform (NovaTrack) High-throughput vehicle telematics platform in Go with live behavioral anomaly detection (geofence violations, unauthorized routes, reckless driving), MQTT event ingestion, and predictive maintenance scoring on streaming sensor data.

EDUCATION

B.Sc. Software Engineering, Babcock University *2021 - 2024*
Thesis: ML-Based Predictive Model for Colorectal Cancer Patient Survival