

Festus Koech

AI Engineer: LLMs, Agentic Systems, RAG, MLOps, Full-Stack

[Gmail](#) | +254 701 183 935 | [GitHub](#) | [LinkedIn](#) | [Portfolio](#)



I build and ship production AI systems. Three years of daily engineering practice, six production deployments, and real users depending on systems I built from scratch. I work across the full stack: LLM integration, agentic pipelines, RAG, MLOps, and the infrastructure underneath.

Technical Skills

Area	Technologies
AI and Agents	LangChain, LangGraph, Anthropic Claude, Azure OpenAI, OpenAI, Hugging Face, Prompt Engineering, Human-in-the-loop, Agent governance, FastMCP, CrewAI
RAG and Search	Multi-source RAG pipelines, Qdrant, Weaviate, ChromaDB, Jina Embeddings, hybrid search, Crawl4AI, SearXNG, MCP
ML and Deep Learning	PyTorch, TensorFlow, LSTM, YOLOv11/12 fine-tuning, NLP, Behavioural ML
MLOps	Azure ML, MLflow, continuous retraining pipelines, CI/CD quality gates, model observability
Computer Vision	MediaPipe FaceMesh, OpenCV, YOLO object detection
Backend and APIs	FastAPI, Node.js, Express.js, REST APIs, WebSockets, Redis pub/sub, Celery, BullMQ, JWT/OAuth
Frontend	React, Next.js, TypeScript, TailwindCSS
Infrastructure	Kubernetes (AKS), Docker, NGINX, GitHub Actions, Azure Container Apps
Cloud	Microsoft Azure (AKS, ML, Blob Storage, Container Registry, Redis, PostgreSQL), IBM Cloud
Databases	PostgreSQL, MongoDB, Redis, Qdrant, Weaviate, ChromaDB
Languages	Python, TypeScript, JavaScript, Go, Kotlin, C/C++

Experience

Co-Founder and Lead AI Engineer Elimika AI

2025 - Present

- ❖ Built Elimika AI Tutor: a LangGraph agentic pipeline with real-time multi-source RAG (Crawl4AI, SearXNG, Qdrant) and a custom PyTorch LSTM model that adapts content delivery per student. 1,500+ active students, 100% uptime.

- ❖ Owned the full MLOps lifecycle: Azure ML, MLflow experiment tracking, automated model promotion, continuous retraining from production data.
- ❖ Provisioned Azure infrastructure from scratch: AKS, NGINX, Redis, Celery, PostgreSQL.
- ❖ Awarded Microsoft Azure, IBM Cloud, and Anthropic Claude credits.

Mid-Level Software Developer Martial Jenga Labs

November 2025 - Present

- ❖ Built an AI exam marking and supervision system using YOLOv11/12, MediaPipe FaceMesh, and a custom LSTM anomaly model. Handles 500 concurrent sessions on AKS.
- ❖ Engineered multi-worker WebSocket architecture via Redis pub/sub for real-time alerting at scale.
- ❖ Integrated dual LLM backends (Claude and OpenAI) via LangGraph across grading and tutoring layers.
- ❖ Built a continuous model retraining pipeline with Azure ML and MLflow, including automated quality-gate promotion.
- ❖ Progressed from intern to mid-level in under a year.

Peer Teacher, AI Engineering Bootcamp Martial Jenga Labs

September 2025 – February 2026

- ❖ Co-teach the AI Agents class covering LangChain, LangGraph, and production agentic system design.
- ❖ Mentor students through Python, ML, and real-world AI implementation.

Freelance Developer Events Zawadi

2025 - Present

- ❖ Built and maintain a full-stack social gifting platform: Node.js/Express, Next.js, PostgreSQL, BullMQ, Redis, WebSockets, Azure Blob Storage. 100% uptime.

Projects(titles are linked)

Elimika AI Tutor - Adaptive Agentic Learning Platform: LangGraph pipeline with real-time RAG (Crawl4AI, SearXNG, Qdrant) and a PyTorch LSTM tracking per-student learning patterns. Full MLOps on Azure. 1,500+ students, 100% uptime.

Python LangGraph LangChain PyTorch Qdrant FastAPI Azure ML MLflow Docker Kubernetes

Voicekit - Self-Hosted Voice Agent Infrastructure: Python package for building production voice agents without managed platform fees (\$0.15-0.25/min). Integrates Whisper STT, Chatterbox TTS with voice cloning, Silero VAD, and LLM routing into a single streaming pipeline. LLM tokens stream directly into TTS - audio chunks reach the caller before the LLM finishes responding. True end-to-end streaming. WebSocket gateway, Docker-deployable, fully configurable via YAML.

Python Whisper Chatterbox TTS Silero VAD Anthropic Claude OpenAI WebSockets Docker Redis FastAPI asyncio

Prepwise - AI Job Search Operating System: Four LangGraph agents (job search, document generation, interview prep, application tracker) sharing state over WebSockets. Ingests a CV, scores jobs, generates tailored resumes and cover letters, and runs conversational interview prep.

Python FastAPI LangGraph Qdrant Jina Embeddings Anthropic Claude PostgreSQL Redis Next.js Docker

Elimika 365 - AI Exam Marking System: Production AI marking pipeline combining YOLOv11/12 object detection, MediaPipe FaceMesh pose estimation, and a custom LSTM anomaly model with per-student calibration. Multi-worker WebSocket architecture via Redis. 500 concurrent sessions on AKS.

Python FastAPI YOLO PyTorch LangGraph Redis Azure ML MLflow Docker Kubernetes

Sieve - VoIP AI Call Interception System: Real-time agentic call handling system built on FreeSWITCH. Routes incoming calls to a live LangGraph agent (Whisper STT + Claude/OpenAI) that autonomously handles or escalates. Escalated calls pushed to a Kotlin Android app via FCM, connected via Linphone SIP - full human-in-the-loop telephony pipeline. Works anywhere a SIM card works with no VoIP dependency and no per-minute termination costs.

Python FastAPI LangGraph LiveKit Whisper Anthropic Claude FreeSWITCH VoIP/SIP Kotlin Redis Celery Docker

GitHub PR Review MCP Server: FastMCP server exposing GitHub pull request operations as LLM-callable tools; list open PRs, read diffs, post inline review comments, approve, and request changes. Single authenticated httpx client via lifespan context. Deployable via Docker with Streamable HTTP transport; tested end-to-end with MCP Inspector and Claude Desktop.

Python FastMCP httpx Docker

OpenSearch Engine - Self-Hosted Agent Search Infrastructure: Built for AI agents that fire dozens of search queries per task. Combines SearXNG metasearch with Spider-rs content extraction. Classifies query intent, routes to the best engines in parallel, and returns structured LLM-ready JSON with URLs, snippets, full markdown content, token counts, and relevance scores. Zero per-query cost.

Go Rust Python FastAPI SearXNG Docker Grpc ML AI Vector Embedding

Docker Container Manager MCP Server: FastMCP server wrapping the Docker SDK - gives any LLM client the ability to list, start, stop, inspect, and monitor Docker containers on the host. Accurate CPU and memory stats matching docker stats CLI output. Docker client initialised via FastMCP lifespan context.

Python FastMCP Docker SDK Docker

Education

BSc Computer Science Catholic University of East Africa (CUEA)

Advanced Data Science Bootcamp Martial Jenga Labs, Nairobi

Software Development Bootcamp Modcom Institute of IT, Nairobi

Certifications

- ❖ AI Agents Development - Anthropic (2026)
- ❖ Claude Code Skills - Anthropic (2026)
- ❖ Claude AI Skills - Anthropic (2026)
- ❖ LangGraph - DeepLearning.AI (2025)
- ❖ Deep Learning Specialization - Udemy (2025)
- ❖ Certificate in Data Science and AI - Martial Jenga Labs (2025)