

STEFAN O. JUMA

APPLIED AI & FULL-STACK ENGINEER

AI agents • RAG • ML systems • production full-stack engineering

NAIROBI, KENYA
stephenjuma500@gmail.com
+254 707 322 993
linkedin.com/in/stefan-juma-a13489150
github.com/steph500
selfawaretech.com

8

YEARS ENGINEERING

10,000+

USERS SUPPORTED

20+

PRODUCTION FEATURES

40%

FASTER BUILDS

CORE AI STACK

APPLIED AI

RAG • embeddings • pgvector • LLM agents • structured outputs • semantic evidence planning • tool/action routing • model routing • evals

AGENTIC SYSTEMS

Department skills/capabilities • approval-gated actions • source grounding • commitments/conflicts • time-aware prioritization • deterministic policy

ML FOUNDATIONS

TensorFlow • Keras • CNNs/RNNs • EMNIST/MNIST • Adam optimizer • Google Colab • pandas • NumPy • Prophet

FULL-STACK

TypeScript 6 • Angular 22 • React 19 • Next.js 16 • React Native / Expo • Node.js • Python / FastAPI • Java / Spring Boot

DATA & PLATFORM

PostgreSQL • Supabase / RLS • REST / GraphQL • WebSockets • Redis • Docker • Kubernetes • AWS • GCP • CI/CD

HOW I BUILD

- Ground AI behavior in evidence, permissions and deterministic policy.
- Own systems end to end: architecture, APIs, data, UI, testing and deployment.
- Ship in small verified increments with strong security and regression coverage.
- Work effectively across distributed teams; mentor, review code and unblock delivery.

PORTFOLIO

selfawaretech.com

PROFILE

Applied AI and full-stack engineer with 8 years of experience shipping production web, mobile, backend, data and AI-enabled systems. I build practical AI products that combine retrieval, structured model outputs, tool use and secure software boundaries rather than treating the model as the whole system. My current flagship work is Tasko, a multi-tenant AI operations platform with permission-scoped RAG, department assistants, persistent context, skills/capability routing, exact approvals and audited provider actions. I also bring hands-on ML foundations with TensorFlow/Keras, neural networks, EMNIST/MNIST and evaluation workflows.

EXPERIENCE

Cell Collective / University of Nebraska-Lincoln | Full-Stack Software Engineer

Jan 2023 - May 2026 | Remote | Distributed US research team

- Delivered 20+ production features for a scientific modeling platform supporting 10,000+ researchers and users across data-heavy React/TypeScript interfaces, Node/Python services and API workflows.
- Extended kinetic-model and PBPK-related simulation workflows, scientific data handling and interactive result analysis while improving shared monorepo tooling enough to cut build times by about 40%.
- Owned performance profiling, production debugging, Cypress testing, code review, CI/CD and staging/production releases in a complex research software environment.

SelfAwareTech | Founder & Independent Software Engineer

Feb 2021 - Present | Remote | Part-time / concurrent product studio

- Built Tasko, a live multi-tenant AI operations platform spanning Marketing, Sales, Support, Operations, Finance and People/HR with grounded department assistants and permission-scoped RAG.
- Implemented private document ingestion, semantic chunking, pgvector retrieval, citation-first answers, structured AI generation, skills/capability policies, tool/action routing, persistent context and exact approval/execution boundaries.
- Designed advanced AI intelligence around semantic evidence planning, cross-source commitments/conflicts, time-aware priorities, stable prepared-work identities, model routing and policy-gated external actions.
- Also shipped Nyumba Services, Stok and Haulio across React Native/Expo, Angular/Ionic, Next.js/NestJS, PostgreSQL/PostGIS, Supabase, analytics and Android production releases.

EnigmaScore | Senior Software Developer

Feb 2019 - Nov 2022 | Hybrid | Nairobi, Kenya

- Led full-stack development of an enterprise recruitment and psychometric platform using Angular, TypeScript, Java/Spring Boot and PostgreSQL, from UML/data modeling through APIs, scoring logic, analytics and deployment.
- Redesigned candidate-analysis and assessment workflows around real customer needs and helped onboard early customers including KFC and Rise & Learn.
- Delivered remote client work using React, Node.js and GPT-3-based interview/question scoring while mentoring, reviewing code and owning AWS releases and production debugging.

TASK0

FLAGSHIP AI OPERATIONS PLATFORM

Secure multi-tenant AI employee system for small and growing companies

GROUNDING INTELLIGENCE

Private knowledge ingestion • semantic chunking • pgvector retrieval • permission-scoped RAG • citation-first answers • structured output validation • semantic evidence planning

AGENT ARCHITECTURE

Department-specific skills/capabilities • tool routing • persistent context • commitments/conflicts • dynamic priorities • stable prepared-work identities

SAFETY & EXECUTION

Tenant/RLS isolation • source grounding • policy gates • exact approvals • provider action boundaries • idempotent execution • audit trails • regression/security suites

INTEGRATIONS & RUNTIME

Google Workspace • Slack • provider/OAuth framework • local/cloud model routing • Ollama/OpenAI-compatible runtime • Angular 22 • TypeScript 6 • Supabase/PostgreSQL

ML MODEL TRAINING & EVALUATION

TensorFlow / Keras / EMNIST

- Trained and evaluated neural-network models in Google Colab using TensorFlow and Keras with EMNIST/MNIST datasets.
- Worked with CNN/RNN architectures, Adam optimization, preprocessing, training loops, tuning and evaluation workflows.
- Used Python, pandas and NumPy for data preparation and analysis; also worked with Prophet for forecasting.

CELL COLLECTIVE / IMMUNE DIGITAL TWIN

Scientific software / modeling / data

- Built production scientific interfaces and services used by 10,000+ users, including kinetic-model and PBPK-related workflows.
- Combined React/TypeScript visualization, API-driven analysis, simulation results and scientific data interpretation.
- Improved monorepo build performance by about 40% while owning testing, debugging and release quality.

AI ARCHITECTURE HIGHLIGHTS

GROUNDING RAG

Private ingestion, semantic chunking, pgvector retrieval, citations, freshness and permission scope before similarity search.

AGENTIC SKILLS

Semantic tool/evidence planning, department capabilities, structured outputs, persistent context and approval-gated provider actions.

EVALS & SAFETY

Schema validation, deterministic policy, conflict handling, source grounding, tenant isolation, auditability and AI-boundary regression tests.

ADDITIONAL PRODUCTION SYSTEMS

NYUMBA SERVICES

React Native / Expo marketplace

1,000+ Play downloads • 120+ providers in first month • matching, scheduling, messaging, ratings, admin and analytics.

STOK + BI

Inventory + forecasting

Offline-first inventory and business dashboards across Angular/Ionic, Next.js, PostgreSQL, forecasting, testing and web/mobile delivery.

HAULIO

Freight marketplace

Route/capacity matching, pricing, quotes/bookings and a modern Angular/Ionic/NestJS/PostgreSQL/PostGIS architecture.

SELECTED IMPACT

AI ENGINEERING

Built grounded RAG and structured assistant workflows with explicit approval and security boundaries.

PRODUCT SCALE

Supported 10,000+ scientific users; shipped consumer products reaching 1,000+ downloads and 120+ providers.

ENGINEERING LEADERSHIP

8 years owning architecture, delivery, code review, mentoring, deployment and production debugging.

EDUCATION

Bachelor of Science in Computer Science

Meru University of Science and Technology | Kenya

LINKS

linkedin.com/in/stefan-juma-a13489150
github.com/steph500 | selfawaretech.com