Winston Karanja

AI Engineer: LLM agent systems in production

Nairobi, Kenya (UTC+3)
Open to fully remote roles worldwide, as employee or contractor. Full working-day overlap with EU and UK, mornings to 1pm ET.

winstonkaranja.me
ngigewinston@gmail.com
linkedin.com/in/winston-ngige
github.com/winstonkaranja

I ship agent systems that run in production. Recent work: a LangGraph natural-language-to-SQL agent answering executive BI questions on a national utility's Oracle warehouse; an enterprise analytics platform built on the rule that the model explains metrics and never computes them; and a hosted MCP server used from Claude and ChatGPT. Python-first, and I own the whole path from ingestion and agent design through evaluation, API and the React surface. Two years alongside distributed teams and clients.

4 subject areas	100+ hosts	94% mAP@0.5	13 tools
NL2SQL agent live on Oracle 23ai for a national utility	hand-deployed pilot turned into one repeatable rollout pipeline	YOLO11 model over 102 pest classes, 3,100+ downloads	hosted MCP server for ChatGPT, Claude and Claude Code

EXPERIENCE

Computech Limited · AI Engineer

Nov 2025 - Present

East African enterprise IT systems integrator. I am the AI engineering function, taking internal and client problems from scoping to production.

- **Oracle Agent (NL2SQL).** Built the LangGraph agent that turns executive questions into Oracle SQL across four subject areas (sales, finance, operations, customer) for a national electric utility, covering revenue, debt aging, outage MTTR and SLA compliance. Queued BI report requests became on-demand answers. Model and database agnostic by design, so swapping the LLM (Azure OpenAI, Huawei MaaS, Anthropic) or the database (Oracle, PostgreSQL, Supabase) is a configuration change, not a code change. Full LangSmith tracing, self-hosted Agent Protocol server instead of a managed one. LangGraph · SQLcl MCP · Oracle 23ai · LangSmith · Aegra · Azure OpenAI / Huawei MaaS
- **AnalyGeek, enterprise communication analytics.** Ingests Microsoft 365 metadata (mail, Teams, calls, meetings, presence) through Graph delta queries and webhooks into PostgreSQL, and computes performance metrics in deterministic SQL and Python. The LangGraph agent only reads pre-computed metrics through tools and explains them: it never sees message content and never computes a number. Privacy posture (metadata-only, RBAC, audit logging, retention) specified across 17 component specs. FastAPI · LangGraph · PostgreSQL · Microsoft Graph · React + Tailwind · Entra ID SSO · Docker
- **BidBuddy, tender-to-submission platform.** Scrapes six public procurement portals on a schedule, routes each tender through a weighted qualification scorecard, a commercial and financial-risk Go/No-Go form and quorum-based committee review, then exports one merged submission PDF. In daily use by the bids team; around 420 commits, co-built with two engineers. Next.js App Router · FastAPI · PostgreSQL · Entra ID SSO · Microsoft Teams JS · Docker
- **Legal Fiddle, contract review advisor.** A deep-agent LangGraph workflow returning structured plain-English business-risk memos, with the advisor-not-lawyer boundary enforced by a fixed output schema and a human review gate rather than by prompt wording. LangGraph · deepagents · FastAPI · Agent Chat UI
- **Infrastructure AI and observability rollout.** A scheduled agent inspects a VMware estate, correlates against Prometheus and Loki and proposes remediation into a Teams approval queue, so nothing acts unattended. Separately, turned a hand-deployed two-VM bank pilot into a discover, classify, generate-config, deploy and verify pipeline covering 100+ hosts with no shared asset register. LangGraph · pyVmomi · Grafana Alloy · Prometheus · Loki · Ansible · Kubernetes
- **Supporting data infrastructure.** Incremental Zoho CRM and Desk to PostgreSQL sync feeding Power BI; a Sage receivables service sending escalating reminders through a Microsoft Graph shared mailbox; a Swahili and Kenyan-language call transcription scaffold with Salesforce CTI.

LSL · Founder and AI Engineer

Jul 2024 - Present

Small AI consultancy I run alongside full-time work, part-time. Main product line is Denning, a legal-AI platform for Kenyan practice.

- **Denning AI.** Rebuilt the legal assistant as two LangGraph chat graphs sharing a deterministic router (frontend action, keyword rules, then a structured-output intent classifier), after a single ReAct agent was found answering legal questions from training data without retrieving. Research follows a fixed retrieval plan over curated Kenyan cases, a statute and doctrine wiki, a BM25 judgment archive and web search last, and citation guards drop citations that cannot exist. Case writes go to a Go backend over typed gRPC. LangGraph · LangChain · Pydantic · gRPC · LangGraph Cloud · LangSmith · SQLite FTS5
- **First-generation platform.** A FastMCP server for legal document processing and semantic search over Pinecone on AWS EC2, fed by a checkpointed scraper on AWS ECS that ingested Kenyan case law and statutes with Celery and Amazon SQS. FastMCP · Pinecone · Celery + SQS · AWS ECS · AWS EC2

Web3 Clubs · AI Consultant

Aug 2025 - Oct 2025

- Fraud and anomaly-detection workflows over on-chain transaction data, advising stakeholders on detection thresholds and reporting cadence. Built n8n agents integrating the WhatsApp Business API with Google Calendar.

Kenya Civil Aviation Authority · Engineering Attachment

Jun 2023 - Aug 2023

- Built a Python pipeline extracting and sorting flight-record logs off Linux servers, cutting the billing office backlog by 30% and surfacing usage patterns for operations staff.

SELECTED PROJECTS

Hedge Fund: a 13-seat AI investment committee

github.com/winstonkaranja/hedge-fund

A personal Nairobi Securities Exchange fund run as a committee rather than a chatbot. Twelve LangGraph specialists analyse one ticker in parallel, each strictly from its own lens; a Chief Risk Officer attacks the strongest thesis adversarially; a Portfolio Manager chairs and weighs argument quality over vote count. A deterministic risk gate sizes the position and enforces hard limits, and nothing executes without approval. Python · LangGraph · LangSmith · backtesting and calibration harness

Mint: hosted email and calendar MCP server

github.com/winstonkaranja/mint

One URL, pick Google or Microsoft at sign-in. Thirteen tools do priority scoring, email and calendar cross-referencing, conflict detection and travel-aware planning server-side, so the assistant receives 3 to 10 actionable items instead of 200 raw emails. Most tools accept shared and delegated mailboxes, and it runs in Claude, ChatGPT and Claude Code. Python · FastMCP · OAuth proxy · Azure Container Apps · Google and Microsoft Graph APIs

Agent harness: one LangGraph graph, any business domain, configured by versioned packs

A configuration-driven engine where one compiled graph runs both agent-directed and step-driven execution, and each business process is a domain pack of YAML, prompts, skills and tools rather than engine code. Built at Computech. LangGraph · LangChain · deepagents · LangSmith · Pydantic · PostgreSQL

SKILLS

Agents and LLM. LangGraph, LangChain, deepagents, LangSmith (tracing, evaluations, prompt hub), MCP and FastMCP, Agent Protocol (Aegra), RAG and semantic search (Pinecone), structured output, tool calling, human-in-the-loop gates, Azure OpenAI, Anthropic, Gemini, Huawei MaaS.

Languages and data. Python, TypeScript and JavaScript, SQL, PostgreSQL, Oracle 23ai, NL2SQL, Alembic, pandas, NumPy, Prior exposure to Go, Rust and Solidity.

Platform and infrastructure. FastAPI, Next.js App Router, React, Docker, Celery and Amazon SQS, AWS (ECS, S3, Lambda, ECR, EC2), Azure (Entra ID SSO, Container Apps, Blob Storage), Supabase, Grafana, Alloy, Prometheus, Loki, Kubernetes, Ansible.

Machine learning and other. PyTorch, Ultralytics YOLO, Microsoft Graph, geospatial processing (GDAL, PDAL, rasterio, GeoPandas), Selenium, BeautifulSoup, n8n.

EDUCATION

B.Sc. Telecommunications and Information Engineering

Sep 2020 - May 2025

Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya

MAUI: Agronomics Recommendation System. Final-year project, source: github.com/winstonkaranja/mars. Trained a YOLO11s model on IP102 (102 pest classes, 896 px, 100 epochs on an A100) to 94% mAP@0.5 and 92% precision on its validation split, then wrapped it in a LangGraph agent that combines detections, NDVI from multispectral imagery and Open-Meteo evapotranspiration into specific irrigation or pest advice. Published on Hugging Face with 3,100+ downloads (underdogquality/yolo11s-pest-detection). Python · Ultralytics YOLO · PyTorch · LangGraph · Anthropic Claude · AWS S3

Carbon-Sync: Project Jumanji. Source: github.com/winstonkaranja/project-jumanji. Computes NDVI and crop biomass from multispectral aerial imagery and converts it to carbon credit volume in tonnes of CO2 equivalent, the measurement step that normally keeps small farms out of carbon markets. A LangGraph graph checks the crop and the farmland before any credits are computed. Python · LangGraph · tifffile · NDVI and biomass estimation · AWS S3

CERTIFICATIONS

LangChain Academy: Introduction to LangChain · Introduction to LangGraph · Agent Observability and Evaluations · Ambient Agents with LangGraph · Deep Agents with LangGraph · Deep Research with LangGraph