

KATLEGO MOKOTO

CURRICULUM VITAE · AI ENGINEER / FULL-STACK .NET DEVELOPER

MENS MERCATUS

PROFILE

I BUILD AGENTIC WORKFLOWS AND FULL-STACK SYSTEMS A TEAMMATE CAN REVIEW, DEPLOY, AND TRUST

Career pipelines, multimodal assistants, and research desks — with a human decision on every irreversible step. I work across architecture, implementation, and deployment, and I understand both ML application behavior and the production engineering required to make systems secure, observable, and maintainable.

PHONE 065 842 4274
EMAIL KATLEGO.MOKOTO.KM@GMAIL.COM
WEB MENS MERCATUS.ONLINE
LINKEDIN LINKEDIN.COM/IN/KATLEGO-MOKOTO1759
LOCATION JOHANNESBURG, SA · ON-SITE / HYBRID / REMOTE

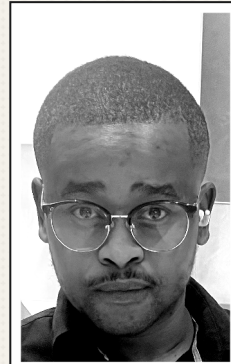


FIG. 1 — KATLEGO MOKOTO, AI SYSTEMS ARCHITECT.

CORE ARCHITECTURE PHILOSOPHY

AGENTIC WORKFLOWS

01 INTENT User objective.	02 CONTEXT Scoped data only.	03 ROUTER Model / fallback.	04 TOOLS APIs & sandboxes.	05 HUMAN GATE Irreversible acts.	06 RECORD Latency, audit, log.
---	--	---	--	--	--

ENGINEERING EXPERIENCE

MAR 2026 - PRESENT

AI SYSTEMS ARCHITECT & FOUNDER · PROJECT ANTIGRAVITY

MAR 2026 - PRESENT
JOHANNESBURG, SA

LOCAL-FIRST AGENT TOOLING · APPLE SILICON M4

- Engineered local-first agent tooling on Apple Silicon M4 running **Ollama** and local quantized LLMs.
- Architected agentic workflows using **Agent Zero** and **Cursor** for code synthesis, tool orchestration, and sandbox tests.
- Conducted distributed model training and fine-tuning experiments across **Nebius** cloud infrastructure.

FULL-STACK, .NET & AI SYSTEMS ENGINEER · MENS MERCATUS

ONGOING
PRODUCTION DESK

PRODUCTION SYSTEMS · SHIPPED AND HOSTED

- Developed **Mercatus Assist**: ASP.NET Core 8 service routing Gemini calls through an OpenAI-compatible interface with exponential backoff retries, systemd unit supervision, and Cloudflare caching.
- Engineered **THINK** workspace in React 19 featuring server-side token proxying and real-time live-voice WebSocket relay.
- Architected **JobNet** career automation pipeline, **Forex Mercatus Codex** market terminal, and **SoundX** Web Audio engine.

OPERATIONS & OPTIMIZATION SYSTEMS · EARLIER SYSTEMS WORK

PRIOR
FLEET & INDUSTRIAL

PHYSICAL SYSTEMS · ROUTING AND FAIL-SAFES

- Designed routing and load allocation heuristics utilizing **Google OR-Tools** under strict operational constraints.
- Managed diesel and heavy mechanical systems operations, applying rigorous systems thinking to physical fail-safes.

01 · CAREER PIPELINE

HUMAN APPROVAL REQUIRED

A daily career pipeline running a 20:00 SAST scrape, sentence ranker, and one-role-per-company rule. A note is dispatched only when the posting prints an employer address, that mailbox verifies, and the letter cites the role. Otherwise the send is held for human confirmation.

ROLE: DESIGNED AND OPERATED THE PIPELINE
STACK: AGENTS · GMAIL · MACOS AUTOMATION · CV TAILORING

03 · QUANTITATIVE RESEARCH DESK

SESSION-SCOPED AND BOUNDED

A hosted terminal for charts, multi-market views, and a session-scoped paper bot. Compact footprint (~87 MB on disk), quotes refreshing near every 60 seconds, up to 24 concurrent sessions with a 2-hour idle TTL, and a public health JSON status endpoint.

ROLE: BUILT, HOSTED, AND BOUNDED THE DESK
STACK: PYTHON · HIGHCHARTS · YFINANCE · UBUNTU VPS

02 · MULTIMODAL ASSISTANT

LIVE, WITH A LABELLED FALLBACK

A .NET Core 8 service and React workspace calling Gemini from the server, keeping the API key off the browser. Provider-swappable chat completions, exponential backoff on 429/5xx, per-IP caps, systemd daemonization, and a labelled demo mode when upstream is degraded.

ROLE: BUILT THE API, THE PROXY, AND THE WORKSPACE
STACK: C# · ASP.NET CORE 8 · REACT 19 · TYPESCRIPT · GEMINI

04 · IN-BROWSER AUDIO STUDIO

CLIENT-SIDE EXECUTION

A guided four-step browser workstation: Load Demo, Play, Shape Master, and Generate. Audio processing executes entirely client-side with zero data egress and no required authentication. It listens to the audio buffer, estimates tempo from onset intervals, and requests user confirmation prior to application.

ROLE: BUILT THE BROWSER DAW AND THE AI PRODUCER
STACK: WEB AUDIO API · TYPESCRIPT · DSP · AI PRODUCER

TECHNICAL EXPERTISE & PRODUCTION STACK

VERIFIED IN PRODUCTION

AI & AGENTS

LLM orchestration, RAG, LangChain, ChromaDB, Ollama, Gemini API, Agent Zero, tool-using agents with human approval gates.

LANGUAGES

C#, Python, TypeScript, JavaScript, SQL, Bash.

BACKEND & CLOUD

ASP.NET Core 8, FastAPI, Node.js, WebSockets, Redis, Docker, systemd, nginx, Cloudflare, Ubuntu VPS, Nebius.

FRONTEND

React 19, Next.js, Vite, TypeScript, Tailwind CSS, Flutter, Web Audio studio DSP interfaces.

DATA & MODELING

PostgreSQL, Pandas, NumPy, yfinance, OR-Tools (combinatorial optimization), regime research models.

```
KATLEGO@MERCATUS - ZSH

% tool-plaza sweep --concurrency 47 --json-rpc
handshake ..... ok · protocol v2024-11-05 negotiated
concurrent ..... 1.7s → 7.4× improvement over sequential
% soundx master --measure ebu-r128
output ..... -14.0 LUFS · written and re-measured
% retrieval index --model-aware
score guard ..... same-model vectors only · no mixed-index ranking
```

EDUCATION & CREDENTIALS

Self-Taught Engineer — Dedicated deep study across Distributed Systems, .NET Architecture, Agentic AI Workflows, and Applied Quantitative Computing.

Available for AI engineering, full-stack .NET, and product engineering teams. On-site, hybrid, or remote — Johannesburg, Pretoria, Cape Town, Durban, and remote teams.