SIMBAREMUTEURO CHIN'OMBE
131-07 160th Street, Queens, NY 11434 | +1 929 602 5125 | Email | Portfolio | LinkedIn | GitHub

EDUCATION

YESHIVA UNIVERSITY — New York, NY
Master of Science in Artificial Intelligence — August 2024- May 2026
Coursework: Numerical Methods ,Computational Stats, Machine Learning, Predictive Models, Data Acquisition, Artificial Intelligence, Reinforcement Learning, Deep Learning, Natural Language Processing, Financial Time Series

HARARE INSTITUTE OF TECHNOLOGY — Harare, Zimbabwe
Bachelor of Technology Honors Degree in Information Security and Assurance — August 2015- June 2019

PROFESSIONAL EXPERIENCE

HUMANITARIANS AI — Mycroft Project — Boston, MA (Remote)
AI Engineer — August 2026 – Present

- Sole engineer developing an end-to-end LLM inference gateway with a six-task routing policy and three model tiers, integrating output validation, single-retry escalation, and per-attempt cost and latency logging.
- Built a frozen benchmark harness and evaluation tooling to compare model quality, cost, and latency across repeated live runs, with evaluation underway against predefined cost and quality thresholds.
- Developed 185 automated tests covering the gateway, routing policies, model adapters, validation, and evaluation tooling.

TENNESSEE PERSONAL ASSISTANCE — Tennessee, USA
Software Engineer Intern — June 2025- August 2025

- Developed and tested .NET Core backend services for enterprise HR and payroll systems, working with data-intensive transactional workflows.
- Designed and optimized SQL Server queries and stored procedures, improving data retrieval latency by 15% for payroll and attendance analytics.
- Performed exploratory data analysis on employee attendance and payroll datasets to identify anomalies and trends, supporting operational decision-making.
- Built automated reports and visualizations using Tableau, translating raw datasets into actionable insights.
- Followed strict data integrity, security, and compliance practices when handling sensitive financial and personal data.

ZIM PROFESSIONALS — Harare, Zimbabwe
Software Developer — January 2023 – May 2024

- Increased application performance by 20% by optimizing .NET Core APIs and streamlining database queries using SQL Server.
- Collaborated with cross-functional teams to identify bottlenecks in the development pipeline, leading to a 15% increase in overall project efficiency.
- Designed and implemented modular Angular components, improving front-end maintainability and reducing development time by 30%.
- Utilized Power BI for generating insightful reports and data visualization.
- Implemented caching strategies using Redis, reducing database load and improving response times for high-traffic scenarios.
- Developed and integrated secure authentication and authorization mechanisms using OAuth 2.0 and JWT, enhancing application security and user access control.
- Refactored legacy .NET applications to .NET Core, improving performance, maintainability, and compatibility with modern cloud architectures.

ESCROW GROUP — Harare, Zimbabwe
Software Developer — October 2019 – December 2022

- Developed and maintained capital markets applications supporting IPO workflows and stock-exchange data integration using C# and .NET.
- Designed and optimized SQL-based data pipelines for financial transactions and market-related datasets, ensuring accuracy, consistency, and auditability.
- Built reporting and analytics tools using Crystal Reports and DevExpress to support market monitoring and operational decision-making.
- Integrated internal and external APIs to enable secure, low-latency data exchange across financial systems.
- Collaborated with business stakeholders to translate regulatory and market requirements into robust software solutions.

UPWORK
Freelance Software Developer — June 2022 – August 2024

- Delivered data-driven software solutions for international clients, with an emphasis on backend performance and system reliability.
- Built scalable applications using Python, .NET Core, and SQL, supporting data processing and analytics workflows.
- Integrated third-party APIs and payment systems, ensuring transactional accuracy and fault tolerance.
- Managed end-to-end development lifecycles, balancing technical rigor with real-world constraints.

SKILLS

- Programming Languages: Python, C++, R, C# ,Javascript
- Quantitative & Mathematical Skills: Numerical Methods, Computational Statistics, Probability, Time Series Analysis, Predictive Modeling
- Systems & Performance: Performance Optimization, Low-Latency Systems, Caching (Redis), API Optimization
- Machine Learning Frameworks & Tools: PyTorch ,OpenCV, Tensorflow, XGBoost, Scikit-learn, Databricks, Spark
- Frameworks & Tools: Azure, .NET Core, Blazor, Angular.js , Nx Toolkit
- Quality Assurance Tools: Selenium , NUnit
- API & Integration: REST APIs, FastAPI, .NET Core, SOAP, OpenAPI/Swagger, API Integration, Webhooks
- Mobile Development: .NET MAUI
- Databases: MS SQL Server, AWS S3 ,Neo4j
- Reporting Tools: Crystal Reports , Tableau, PowerBI ,Dev Express
- Authentication & Security: OAuth 2.0, JWT, RBAC, API Security
- Cloud & Infrastructure: Azure, Docker, Kubernetes

PROJECTS

Autonomous Excel Operations Agent — New York, NY
Python, TypeScript, FastAPI, React, Office.js, MCP, MySQL, Anthropic API

- Designed and built a multi-agent AI system (orchestrator, data agent, report agent) that converts natural-language requests into live operations on an open Excel workbook , native tables, PivotTables, and charts rendered in real time through an Office.js add-in with an in-Excel chat interface.
- Developed two MCP tool servers enforcing safety structurally rather than behaviorally: a SQL gate (read-only database user, single-SELECT validation, row caps, timeouts) and an Excel gate with user-approval workflow, overwrite protection, and an operation ledger powering inverse-operation undo.
- Engineered the reliability layer: typed execution plans, per-step validation including ground-truth cross-checks of written aggregates, and a failure taxonomy mapped to a static recovery policy (retry / replan / escalate) , with every decision persisted to an append-only event log enabling full run replay.
- Built an automated eval harness with a headless workbook simulator and deliberate failure injection across 15 tasks, measuring autonomous completion and recovery rates with programmatic, ground-truth graders.

LLM Inference Platform & Performance Benchmarking
Python, vLLM, Docker, Kubernetes (k3s), Prometheus, Grafana, NVIDIA RTX 4080

- Built and containerized a vLLM serving platform for Qwen2.5-3B-Instruct, deploying on Kubernetes with GPU resource allocation, health probes, persistent model storage, and inference monitoring.
- Benchmarked 1,812 requests across 43 configurations, measuring time to first token, inter-token latency, throughput, and KV-cache capacity under varying prompt lengths and concurrency.
- Measured peak aggregate throughput of 837.6 tokens/second at concurrency 32—a 13.6× increase over single-request throughput—and quantified the accompanying per-user latency trade-offs.
- Diagnosed Kubernetes service-routing failures, GPU memory constraints, and container shutdown issues; reduced pod termination time from over 60 seconds to 1.35 seconds by correcting signal handling.

Nx Video Solutions Hackathon | HackerEarth – New York, NY
Python, OpenCV, TensorFlow, Real-Time Data Processing

- Led a team in developing an AI-driven analytics system for real-time monitoring of venues, applying computer vision and deep learning to quantify crowd engagement
- Processed and modeled a dataset of 20,000+ images to generate emotion, age, and crowd density metrics, providing high-frequency real-time insights
- Optimized video pipeline with Nx Toolkit to handle high-throughput streams, demonstrating system-level performance engineering skills
- Delivered a dashboard with quantitative metrics on audience behavior, enabling data-informed event decisions