VICTOR ADEDEJI

AI SOFTWARE ENGINEER | LLM INFRASTRUCTURE & BACKEND | APPLIED ML SYSTEMS

Website | Lagos, Nigeria | +234 706 133 7286 | Victor.adedeji.t@gmail.com | LinkedIn | GitHub | Google Scholar

PROFESSIONAL SUMMARY

Product-focused AI Software Engineer (Backend) with 4 years building end-to-end intelligent systems from high-performance LLM and ML backend infrastructure to user-facing AI products. Specialise in shipping production-grade generative-AI capabilities at scale: real-time LLM streaming, agentic orchestration with function calling and MCP tooling, retrieval-augmented generation, and resilient async service integration. Proven record of translating complex ML research into measurable business value, including a 95% reduction in streaming event load, 60% lower CPU utilisation, and sub-100ms cached response latency across systems serving 1,000+ concurrent sessions. Adept at bridging research, engineering, and product teams to deliver compliant, fault-tolerant AI systems that scale.

Technical Domains: Backend & Microservices, LLM Infrastructure, Agentic Systems & MCP, Retrieval-Augmented Generation , Multimodal Deep Learning, MLOps / CI-CD, Cloud-Native Deployment, Model Fine-Tuning, AI Guardrails & Compliance

PROFESSIONAL EXPERIENCE

AI Research Fellow Apr 2026 – Present

Fatima Institute for Global AI Research | San Francisco, CA (Remote)

Research fellow advancing multimodal deep-learning architectures, focused on aligning high-dimensional vision and sequential data for large-scale pattern-recognition and detection tasks.

Multimodal AI Research & Architecture

- Co-authored research on auditable, agentic AI workflows for spatial gene expression prediction, accepted for presentation at the NeurIPS 2026 AgenticLS Workshop.
- Collaborated on evaluating the robustness and generalization of pathology foundation models,accepted to the Main Track of GenAI4Health 2026.
- Architected and trained multimodal deep-learning systems that ingest, process, and align disparate high-dimensional datasets (computer vision + sequential data), establishing a unified representation layer for downstream detection tasks.
- Engineered self-supervised data pipelines that extract features from large unlabelled corpora, lifting predictive accuracy on region-detection and pattern-recognition benchmarks while removing dependence on costly manual labelling.

Technical Environment: *Python, Modal, PyTorch, Transformers, Self-Supervised Learning, Computer Vision, NumPy, Distributed Training*

Backend Engineer – AI Stack Oct 2025 – Mar 2026

My Ladder Africa | Accra, Ghana (Remote)

Owned the AI backend for an LLM-driven fintech product, delivering real-time conversational infrastructure and an agentic payment-orchestration engine under strict latency and compliance constraints.

LLM Streaming Infrastructure

- Architected real-time LLM streaming infrastructure (SSE) for 1,000+ concurrent sessions, cutting SSE event volume by 95%, CPU usage by 60%, memory footprint by 40%, and network requests by 90% while achieving sub-100ms cached response latency.
- Designed a dual-mode LLM response architecture (instant fallback + contextual reasoning) with intelligent caching, eliminating 8–12s of latency in default operations and materially improving perceived responsiveness.

Agentic Payment Orchestration

- Built a production-grade, LLM-powered payment orchestration system using OpenAI function calling and 14 MCP tools, enabling end-to-end AI-driven goal-to-payment flows (KYC, OTP, card checkout, direct debit) that accelerated setup completion and lifted 24-hour deposit conversion.
- Engineered AI payment guardrails with human-in-the-loop confirmation, sensitive-data masking, and structured audit logging to guarantee compliance-safe transaction execution.
- Delivered resilient async financial integrations (httpx with retry/backoff) across 7+ payment endpoints, with unified error handling and automatic recovery that minimised failed-transaction fallout.

Cross-Functional Delivery

- Partnered with frontend and product teams to define performant API contracts, embedding complex LLM streaming and agentic logic into the client experience without compromising stability or latency.

Technical Environment: *Python, FastAPI, OpenAI API, MCP, SSE Streaming, Redis, httpx, Docker, AWS*

Key Achievements: 95% reduction in streaming events · 60% lower CPU · sub-100ms cached latency · increased 24-hour deposit conversion · compliance-safe AI transaction execution

AI Backend Engineer Jan 2025 – Oct 2025

Revamp Insights | Lagos, Nigeria (Remote)

Lead AI backend engineer productionising generative-AI features and autonomous agents for enterprise applications.

Applied LLM & Agent Engineering

- Developed and deployed scalable AI agents using Transformers and LangChain, integrated into FastAPI services and shipped to production enterprise AI workloads.
- Designed backend systems for LLM-powered features that improved end-to-end performance and enabled the production rollout of generative-AI capabilities.
- Built and maintained high-performance APIs underpinning AI-driven workflows and automation, raising throughput for downstream consumers.

Technical Environment: *Python, FastAPI, LangChain, Transformers, Vector Databases, RAG, Docker, CI/CD, GCP*

AI Engineer (Consultant) Oct 2024 – Oct 2025

My Health Integral | Cambridge, United Kingdom (Remote)

Consulting AI engineer delivering a multimodal clinical-decision-support assistant for a healthcare platform.

Multimodal Clinical AI

- Collaborated with frontend and backend stakeholders to architect and integrate an end-to-end multimodal LLM-powered chatbot, delivering a scalable solution for real-time symptom analysis, drug interaction and clinical decision support.

Technical Environment: *Python (FastAPI), LLM APIs, Multimodal Models, OCR, RAG, LangGraph, AWS*

Backend Software Engineer Jun 2023 – Dec 2024

Revamp Insights | Lagos, Nigeria (Remote)

Backend engineer building the microservices foundation for a generative-AI platform.

Distributed Backend & Messaging

- Architected and deployed microservices-based backend systems for a generative-AI platform, improving scalability and maintainability across services.
- Designed asynchronous messaging with RabbitMQ to enable reliable, decoupled inter-service communication.
- Optimised backend performance, improving system efficiency and reducing operational overhead.

Technical Environment: *Python, RabbitMQ, Microservices, PostgreSQL, Docker, Kubernetes*

Machine Learning Software Engineer (Intern) May 2021 – Oct 2021

Favourite Doctor | Akure, Nigeria

Applied ML Prototyping

- Reduced operational response time by 30% by implementing an AI-powered chatbot solution.
- Developed a symptoms-checker API in Python and FastAPI, improving diagnostic accuracy and user experience.

Technical Environment: *Python, FastAPI, NLP, Machine Learning*

Machine Learning Intern Jul 2020 – Oct 2020

Rural Farmers Hub (RFH) | FCT Abuja, Nigeria

Data & Computer Vision

- Built a Python data-wrangling automation script with the lead engineer, improving data-processing efficiency.
- Annotated multi-spectral imagery for computer-vision models, enhancing model accuracy and training efficiency.
- Implemented a research paper on soil-nutrient mapping, contributing to improved agricultural insights.

Technical Environment: *Python, Computer Vision, Multi-Spectral Imagery, Pandas, NumPy*

TECHNICAL EXPERTISE

LLM & Generative AI: LLM APIs (OpenAI, DeepSeek, Gemini, Qwen, Claude) , Agentic Systems & Function Calling , Model Context Protocol (MCP), Retrieval-Augmented Generation (RAG) , Real-Time Streaming (SSE) , AI Guardrails & Human-in-the-Loop , Model Fine-Tuning

Machine & Deep Learning: PyTorch ,Transformers , Scikit-learn , NLTK , LangChain , LangGraph , Multimodal & Self-Supervised Learning , Model Optimisation , Pattern Recognition

Backend & APIs: Scalable RESTful APIs & Microservices, FastAPI, Async Integrations (httpx, retry/backoff) , RabbitMQ , Redis , Vector Databases (Milvus, Chroma, Qdrant), Webscraping

Frontend: React(Familiar), javascript/Typescript

Programming: Python , SQL , Golang , C , Kotlin , Shell, Java (Spring Boot - Familiar)

MLOps, Cloud & DevOps: AWS , Azure , Docker , Kubernetes , CI/CD , GitHub Actions , ZenML , Git

Data & Analytics: Statistical Analysis , Pandas , NumPy , Power BI , Excel , LaTeX

PUBLICATIONS

Alim Akram, **Adedeji, V. T., et al. (2026).** "An Auditable Agent for Adaptive Tissue-Expert Selection in Spatial Gene Expression Prediction." *Accepted to the Agentic Life Sciences (AgenticLS) Workshop at NeurIPS 2026*. Openreview Link

Saeed, H. A., **Victor, A. T.**, et al. "Integrating Large Language Models in Robotics to Empower Autonomous Agents with Natural Language Comprehension Capabilities: A Comprehensive Review." International Journal of Future Engineering and Innovation (IJFEI), 2(3), 103–112.

EDUCATION & CERTIFICATIONS

B.Tech, Mathematical Sciences (Industrial Mathematics) **2017 – 2023**
Federal University of Technology, Akure (FUTA), Nigeria — Magna cum laude

Thesis: Pattern-Formation Process with the Gray–Scott Reaction-Diffusion Model.

Most Creative Student Award (2019); Dean's List (2021).

Certifications: KCNA – Kubernetes & Cloud Native Associate, ALX/Holberton Software Engineering (Backend specialisation), Udacity Computer Vision Nanodegree, Microsoft Professional Program for Data Science, DeepLearning.AI: Neural Networks & Deep Learning; Improving Deep Neural Networks (Hyperparameter Tuning, Regularisation & Optimisation)