Saviour Henry

AI Engineer | Remote

johnsavilesh@gmail.com | +234 706 823 0510 | linkedin.com/in/saviour-henry-b35a7a166 | github.com/Ad163

PROFESSIONAL SUMMARY

AI Engineer with 5 years of production experience designing, implementing, and deploying LLM-powered systems, RAG pipelines, and NLP workflows for real-world applications. Proven track record of taking systems from POC to production with proper evaluation frameworks, monitoring, and hallucination mitigation. Built LegalRAG, a zero-hallucination retrieval system achieving 96.7% accuracy through rigorous evaluation methodology redesign. Expert in Python, vector databases, LangChain, LlamaIndex, and the full LLM orchestration stack. Published researcher in cross-domain ML generalization (Franklin Open, Elsevier, 2026). Brings strong client-facing communication skills and the ability to guide stakeholders from experimentation through scalable production deployment.

TECHNICAL SKILLS

LLM and RAG: RAG Pipeline Architecture, Multi-step Retrieval, Query Decomposition, Hallucination Mitigation, Citation-aware Responses, LangChain, LlamaIndex, Prompt Engineering, Function Calling, LLM Fine-Tuning (SFT, RLHF, LoRA, QLoRA), OpenAI API, Claude API, Llama, Gemini, vLLM, Ollama

Vector Databases: FAISS, Pinecone, Weaviate, ChromaDB, pgvector, Semantic Chunking, Embedding Model Selection, BM25 Hybrid Retrieval, Similarity Search Tuning

NLP and ML: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, BERT, DistilBERT, RoBERTa, Sentence Transformers, NER, Text Classification, Sentiment Analysis, Semantic Search, HDBSCAN, PCA, Ensemble Methods, Transfer Learning

Evaluation Frameworks: LLM evaluation methodology design, answer quality metrics (relevance, faithfulness, coverage), offline and human-in-the-loop evaluation pipelines, citation verification, zero-hallucination enforcement, MLflow, Weights and Biases

Production and MLOps: FastAPI, Docker, Kubernetes, CI/CD (GitHub Actions), structured logging, monitoring, error handling, model registries, A/B testing, canary deployments, model drift detection

Cloud: AWS (SageMaker, Lambda, S3, EC2, Polly, Rekognition), GCP (Vertex AI, BigQuery, Vision API), Azure

Data and Pipelines: ETL/ELT, Pandas, NumPy, Apache Airflow, PostgreSQL, MongoDB, Redis, Feature Engineering, Data Quality Monitoring

Languages: Python (primary), SQL, Bash

WORK EXPERIENCE

Turing | AI/ML Engineer

Jul 2025 – Present (Talent Pool, between contracts) | Remote, United States

- Designed and built automated LLM evaluation frameworks that systematically benchmark model performance across multiple quality dimensions, replacing manual spot checks with rigorous, repeatable measurement pipelines that catch failures before production deployment.
- Led Supervised Fine-Tuning (SFT) data pipeline design for enterprise LLMs, curating 500+ high-quality training examples that translated real-world task requirements into measurable model behaviour improvements.
- Executed RLHF workflows end-to-end, designed reward criteria, curated comparison data, and iterated on feedback loops that aligned model outputs with target quality standards and safety requirements.
- Built agentic evaluation pipelines for code repair tasks achieving 85% success rate on SWE-Bench across 200+ real-world GitHub issues, designed the failure-mode taxonomy and scoring rubrics that made systematic improvement possible.
- Collaborated with ML researchers to optimise reward models, reducing false positive bug detections by 40% through systematic reward model refinement.

Freelance | AI Engineer

Aug 2024 – May 2026 | Remote, United States

- LLM Swarm Engineering & Hardening (Project SwarmBench): Designed self-contained evaluation environments, custom verifiers, and task-hardening protocols for complex multi-agent frameworks using LLM orchestrators. Implemented strict budget control systems ($20/task optimization rules) and architectural patterns (Map-Reduce, Fan-Out/Synthesize) to validate multi-agent coordination capability gains over baseline single agents.
- Automated Target Testing: Developed regression test pipelines, deterministic targeted tests (run-tests-eval.sh), and reference golden solutions for industry-grade codebases. Engineered rigorous fail-to-pass ([f2p]) and pass-to-pass ([p2p]) validation checks deployed via isolated GitHub Actions workflows to capture exact model behavioral failures.
- MLOps & Optimization: Hardened Docker orchestration pipelines by implementing shallow cloning, slim base layers, and package purging protocols, directly solving runtime workspace constraint bottlenecks and accelerating remote runner container speeds.
- Agentic Workflows: Built multi-turn agentic conversation workflows with tool use, memory management, and orchestration logic, improving reasoning quality by 35% on held-out evaluation sets through Process-Supervised Fine-Tuning.
- RAG & Intelligence Systems: Deployed an automated research intelligence pipeline that aggregated heterogeneous sources, ran structured summarization with citation tracking, and delivered strategic weekly briefs, dropping manual research time to zero.
- Technical Authority: Hardened POC code into stable production environments with structured logging, explicit error handling, environment separation, and fallback configurations, translating trade-offs clearly to stakeholders.

Start Innovation Hub Nigeria | **Machine Learning Engineer**

Mar 2024 – Present | Uyo, Nigeria

- Architected end-to-end NLP pipelines integrating DistilBERT for text analysis and YOLOv5 for computer vision, automating processing of 50K+ documents and images monthly with 92% classification accuracy.
- Optimised real-time inference pipelines for production deployment, reducing latency from 200ms to 60ms (70% improvement) on edge devices through systematic profiling and targeted optimisation.
- Designed and deployed ML models to production with FastAPI on AWS (Terraform-provisioned), reducing deployment cycles from 4 days to 8 hours through CI/CD automation.
- Mentored 15+ data professionals on LLM workflows, RAG pipeline design, and production deployment best practices, growing team capacity and accelerating delivery speed by 50%.
- Recognised with Staff of the Year Award twice (2024 and 2025), awarded NGN 1,000,000 for measurable contributions to engineering output and organisational performance.

Visis Startup | **Machine Learning Engineer**

Sep 2024 – Dec 2024 | Remote, Nigeria

- Built production NLP and computer vision pipeline for an accessibility platform serving 3,000+ visually impaired users, integrating AWS Polly, Rekognition, and Google Vision API for OCR and TTS functionality.
- Processed 10K+ documents at 99.2% OCR accuracy with automated AWS Lambda workflows, reducing per-request latency from 5.0s to 1.2s (76% improvement) through systematic pipeline optimisation.
- Hardened POC code into a production system with proper error handling, retry logic, monitoring, and A/B testing infrastructure, confirming 40% improvement in user satisfaction before full deployment.

Presprint Digital | **Machine Learning Engineer**

Mar 2022 – Oct 2022 | Calabar, Nigeria

- Developed and fine-tuned ML models for predictive analytics and anomaly detection achieving 98% accuracy through hyperparameter optimisation and systematic feature selection.
- Streamlined ML deployment workflows with Docker and CI/CD pipelines, reducing deployment time by 40% and establishing MLOps best practices for the team.

OPEN SOURCE PROJECTS

LegalRAG System - Zero-Hallucination RAG for Regulated Domains | github.com/Ad163/LegalRAG-System

- Architected a production RAG system for legal documents achieving 96.7% accuracy (29/30 questions) with 100% citation backing and zero hallucinations. Built with hybrid BM25 + vector retrieval, pgvector, FastAPI, Redis rate limiting, and JWT authentication, designed for domains where accuracy is not optional.
- The critical engineering insight: the system started at 29% accuracy. Rather than chasing retrieval algorithms, I rebuilt the evaluation framework from scratch, a 30-question suite covering simple facts, complex reasoning, multi-document synthesis, and edge cases, with citation verification and zero-hallucination tolerance. Accuracy jumped to 96.7%. Measurement quality drives production ML quality.
- Shipped with production-grade observability: structured logging, Redis-enabled rate limiting, JWT auth, input validation, cost tracking, and comprehensive test coverage. Public repository: github.com/Ad163/LegalRAG-System

InsightFlow - Communication Intelligence Platform | github.com/Ad163/InsightFlow

- ML-powered platform that ingests emails and meeting transcripts, runs HDBSCAN clustering and sentence-transformer embeddings to detect emerging topics and stalled decisions, and delivers automated weekly strategic briefs. Built with FastAPI, PostgreSQL + pgvector, Docker, and n8n - AI-optional with deterministic rules-first pipeline and 28-day baseline comparison.

RESEARCH AND PUBLICATIONS

Ensemble Transfer Learning for Cross-Dataset Phishing Detection

Franklin Open (Elsevier), 2026 | DOI: 10.1016/j.fraope.2026.100579 | Open Access

- Developed the Cross-Domain Ensemble Probability Fusion (CDEPF) framework for NLP-based phishing detection across heterogeneous datasets with zero feature overlap, using PCA-based feature harmonisation and information-theoretic weighted fusion to achieve 94.4% cross-dataset accuracy, a 64.3% relative improvement over the 57.4% baseline.

EDUCATION

University of Uyo | *Nov 2019 – Dec 2024*

Bachelor of Engineering, Computer Engineering

RECOGNITION

- **Staff of the Year Award x2 (2024 and 2025), Start Innovation Hub Nigeria:** Recognised twice for outstanding engineering contributions. Awarded NGN 1,000,000 in 2025.
- **Young Africa Innovates Programme, Top 30 Finalist out of 9,000+ applicants (Mastercard Foundation and UNDP, 2024):** Selected for EnoExplore Global Limited. 0.3% selection rate.
- **African Impact Challenge (AIC) Cohort 6 Builder Track (2025):** Selected from 12,101+ completed applications.
- DeepTech Ready Programme Cohort 2, Spotlight Mentor, Data Science Nigeria (Google.org and 3MTT, 2025).
- **1st Runner-Up, USPF Hackathon for Agriculture 2024:** Built AgriPro computer vision model deployed on GCP.