JOSHUA NWACHINEMERE

AI Engineer | LLM Applications, RAG, Multimodal AI & Evaluation
joshua0nwachinemere@gmail.com | +44 7344 403642
github.com/dk3yyyy | linkedin.com/in/joshua-nwachinemere | joshua-nwachinemere.pages.dev

PROFILE

AI Engineer building Python systems for LLM-based retrieval (RAG), multimodal inference and model evaluation. Built a local RAG application with validated citations and a 30-case retrieval benchmark, and a privacy-focused cross-platform desktop assistant routing across nine model providers. Twelve merged open-source pull requests, including the OpenAI Agents SDK and Pydantic AI Harness. Studying MSc Artificial Intelligence at Northumbria University.

CORE SKILLS

Languages: Python, SQL, JavaScript (Electron), Swift
LLM & Retrieval: RAG, semantic retrieval, BM25, ChromaDB, LangChain, Ollama, citation validation, structured outputs, provider routing, multimodal inference, voice AI; OpenAI, Anthropic, Gemini and Azure AI Foundry APIs, Ollama, Deepgram, whisper.cpp
ML & Evaluation: retrieval benchmarking, temporal (rolling-origin) evaluation, scikit-learn, XGBoost, pandas, feature engineering
Backend & Infrastructure: FastAPI, asyncio, REST APIs, n8n, webhooks, SQLAlchemy, Streamlit, PostgreSQL, SQLite, Redis, Docker, GitHub Actions, Linux, pytest

SELECTED AI & ML PROJECTS

Local Review Intelligence | Local RAG & Evaluation — Python, Ollama, ChromaDB, LangChain, Streamlit

- Built a local-first RAG application for review data: adaptive CSV mapping, content-addressed ChromaDB storage, Streamlit dashboard, installable CLI.
- Implemented validated citations, bounded answer repair and safe abstention; compared semantic retrieval with BM25 on a versioned 30-case evaluation set. github.com/dk3yyyy/local_AI_agent

Volyx Lens | Privacy-First Cross-Platform AI Assistant — JavaScript, Electron, Swift, Azure AI Foundry

- Built and released (v0.6.0, Apache 2.0) an Electron assistant for macOS and Windows, with experimental Linux support, combining selected screen context, separate microphone and meeting-audio transcription, and structured meeting notes.
- Routes requests directly to 9 response and 6 transcription providers, including local Ollama and whisper.cpp, with fast/smart model tiers, fallback, and bounded, relevance-ranked context (hash and visual deduplication; local OCR on macOS).
- Hardened with a sandboxed renderer, keychain-backed secrets and explicit data-sharing and cost confirmations; automated tests, secret scanning and cross-platform release CI with checksums and SBOMs. github.com/dk3yyyy/volyx-lens

Football Forecasting Lab | Temporal ML Evaluation Pipeline — Python, XGBoost, scikit-learn, FastAPI, Streamlit

- Built scrapers, temporal feature engineering, XGBoost outcome and Poisson goal models, a FastAPI service and a Streamlit dashboard.
- Rolling-origin evaluation on 1,140 Premier League matches: 53.77% outcome accuracy against a 56.70% bookmaker-implied benchmark; reported that the model did not outperform it. github.com/dk3yyyy/football_predictor

EXPERIENCE

AI Engineer, VolyxAI (independent AI product development) — Nov 2025 – Present

- Build applied AI systems integrating retrieval, multimodal inputs, voice workflows, model APIs and structured outputs.
- Develop Python services and n8n workflows with webhooks, validation, retries, logging, human approval and provider fallback.

Python Automation Developer, Freelance (independent paid client work) — Jan 2023 – Present

- Delivered Python tooling to client requirements, including streaming utilities for large text datasets (filtering, deduplication, splitting).
- Built Telegram integrations and async services (aiohttp, PostgreSQL, Redis, Docker) with explicit error handling, rate limiting and logging.
- Developed licensing and key-management components for proprietary Python tools; published reusable command-line utilities for large text datasets.

OPEN-SOURCE CONTRIBUTIONS (12 MERGED PULL REQUESTS, JUL–AUG 2026)

- **OpenAI Agents Python SDK** #3991: pre-response WebSocket server errors now follow the SDK retry policy, replay safety preserved.
- **Apache Arrow Rust** #10486: JSON reader now rejects nulls for non-nullable ListView children.
- **Pydantic AI Harness** #503, #520: Code Mode failures surfaced as model retries; fixed URL-safe base64 decoding of binary payloads.
- **Mellea** #1469, #1471: stopped Hugging Face generation workers on stream timeout; added OpenTelemetry tracing tests.
- **FastStream** #2961, #2976, #2984 + FastStream FastAPI #2: FastAPI 0.140 support, Redis consumer-group and AnyIO fixes.
- **Also merged:** Altair #4089, Calkit #1028.

EDUCATION & CERTIFICATIONS

MSc Artificial Intelligence, Northumbria University — Sept 2026 – 2027 (expected)

BTech Mathematics, Federal University of Technology, Owerri (FUTO), Nigeria — 2016 – 2021

Certifications: Scientific Computing with Python Developer Certification, freeCodeCamp (May 2026); Google AI Professional Certificate, Google / Coursera (Feb 2026); Model Context Protocol: Advanced Topics, Anthropic training (Mar 2026).