OLUWABUSAYO OLUFUNKE AWOYOMI

Automation Engineer & NLP Researcher | Production Workflows • PyTorch & Hugging Face • GPU Fine-Tuning

Albany, NY ⋄ busayofunke@gmail.com ⋄ 518-690-5858 ⋄ linkedin.com/in/oluwabusayo-olufunke-awoyomi

PROFESSIONAL SUMMARY

Automation engineer and NLP researcher with 8+ years in data roles. On assignment at Regeneron through Insight Global, operates production data workflows with CI/CD, test harnesses, and runbooks, and works directly with scientists, QA/Compliance, and leadership to turn their requirements into tested, documented automation. As a volunteer NLP researcher with the Masakhane community, fine-tuned multilingual transformer models on GPUs using PyTorch and Hugging Face Transformers. Co-author of peer-reviewed NLP papers including AfroLM, CIRAL (ACM SIGIR 2024), and AfriCOMET (NAACL 2024); 212 Google Scholar citations. M.S. in Computer Science.

CORE COMPETENCIES

Production Workflow Operations (CI/CD, Test Harnesses, Runbooks) • Model Fine-Tuning (PyTorch, Hugging Face) • GPU Training Workflows • NLP/LLM Tooling • Technical Stakeholder Communication • Requirements Gathering & Solution Delivery • Data Quality Assurance & Testing • ETL Pipelines • Exception & Deviation Tracking • Cross-System Record Reconciliation • Data Lineage & Metadata Standards • Technical Documentation • SQL & Relational Database Design

PROFESSIONAL EXPERIENCE

Documentation Automation Engineer **Jan 2025 – Present**

Insight Global, LLC — assigned to Regeneron Pharmaceuticals, Inc. under a services agreement

- Operate production data workflows using test harnesses, CI/CD, and runbooks, testing changes before release and maintaining repeatable, documented processes.
- Partner with process development scientists, biostatistics, QA/Compliance, and leadership to refine requirements and deploy automation aligned with business and regulatory needs.
- Aggregate multi-source experimental data into standardized Drug Optimization Reports for executive review, using templated document generation with versioning, approvals, and provenance, supported by NLP/LLM tooling.
- Implement ETL pipelines that combine multi-source datasets, including cell-counting (microscopy/flow) and gene-mapping (omics/sequencing) data, with schema governance, version control, and automated validation.
- Build automated data quality checks for completeness, accuracy, and consistency, with tracking of changes, exceptions, and deviations to protect data integrity in a regulated setting.
- Enforce data lineage and metadata standards so that reports remain traceable to their underlying source data, with auditable output aligned to internal quality procedures.

NLP Researcher (Volunteer) **2022 – Present**

Masakhane open research community

- Fine-tuned multilingual transformer models for downstream NLP tasks using PyTorch and Hugging Face Transformers, running training jobs on a shared GPU cluster, hosted notebooks (Colab/Kaggle), and a personal GPU workstation.
- Co-authored AfroLM (SustaiNLP workshop, EMNLP 2022), a multilingual language model pretrained from scratch on 23 African languages using self-active learning; prepared training data, annotated evaluation data, and ran experiments. Trained on about 14x less data than existing baselines, AfroLM outperformed AfriBERTa, XLM-R-base, and mBERT on NER, text classification, and sentiment analysis.
- Annotated Yoruba data for CIRAL (ACM SIGIR 2024; co-author), a cross-lingual retrieval test collection with 1,600+ English queries, 2.5M+ passages in Hausa, Somali, Swahili, and Yoruba, and 30k relevance judgments by native speakers.

Data Administrator **Mar 2024 – Dec 2024**

Health Research, Inc., Albany, NY

- Identified records present in the CDC's Enhanced HIV/AIDS Reporting System (eHARS) but absent from the New York Electronic HIV Management System (NYEHMS), supporting data completeness and reconciliation between federal and state systems.
- Verified the accuracy of data used for the Bureau's core functions, applying quality assurance checks to sensitive public health surveillance data.
- Interpreted existing eHARS data import code to clarify how records were loaded into the system, and developed, tested, and updated its technical documentation. Tools: SQL, SAS.

Lead Programmer / Analyst **Jan 2024 – Mar 2024**

The State University of New York, Albany, NY

- Developed analytical models in PL/SQL supporting trend forecasting and resource allocation.
- Prepared reports and Tableau dashboards that gave university stakeholders actionable information.
- Worked with cross-functional teams to streamline data collection processes and improve data accuracy.

Graduate Assistant **Jan 2022 – Dec 2023**
The College of Saint Rose, Albany, NY

- Developed a solution integrating SQL FileStream with an SMB file share to meet data security and retrieval-efficiency requirements, including user authentication controls.
- Improved document retrieval performance and reduced routine user support requests through a simplified interface.

Customer Engineer (Internship) **Jun 2023 – Aug 2023**
Ricoh USA, Inc.

- Established a quality assurance checklist used during device setup, while re-imaging more than 300 devices with Acronis and Norton Ghost.
- Worked with cross-functional teams to keep technical documentation current and accessible.

Space Science and Technology Officer — Data Analyst **Aug 2018 – Dec 2021**
Defence Space Administration, Abuja, Nigeria

- Automated recurring reports using Python and ArcGIS, reducing the manual effort required each month.
- Performed data processing and analysis for departmental analytics projects and operational reporting.

Data Analyst — Admin Officer **Jan 2016 – Jul 2018**
GBIT Consulting Limited, Abuja, Nigeria

- Analyzed complex datasets using statistical analysis and data mining techniques.
- Communicated findings to decision-makers through written reports and visualizations.

TECHNICAL SKILLS

- **Machine Learning:** PyTorch, Hugging Face Transformers, vLLM, fine-tuning multilingual transformer models, GPU training (shared cluster, Colab/Kaggle, local GPU).
- **Data Quality & Engineering:** ETL, automated data validation, test harnesses, CI/CD, runbooks, Git/GitHub, data lineage and metadata standards, schema governance.
- **Analytics & Programming:** Python, SAS, Flask, statistical analysis, data mining, predictive modeling, report automation.
- **SQL & Databases:** SQL, PL/SQL, PostgreSQL, SQL Server, MySQL, relational database design; working familiarity with Oracle.
- **Visualization & Reporting:** Tableau, Power BI, Qlik Sense, ArcGIS Desktop.

SELECTED PROJECTS

LinguaCast — Public Health Data Analytics Prototype **Independent research**

- Analyzed 711 weeks of CDC FluView ILINet influenza-like illness data (2010–2024) to build forecasting models; the linear regression model reached an out-of-sample R^2 of 0.9927 on held-out weeks. Research prototype; not deployed or used for decision-making.

Viewer Feedback Database — Graduate Research Project **2022 – 2023**

- Designed a PostgreSQL database and built a Python (Flask) application that stores viewer comments on specific videos with video type, device type, IP address, and timestamp.

EDUCATION

Master of Science in Computer Science **Jan 2022 – Dec 2023**
The College of Saint Rose, Albany, NY

Bachelor of Science in Computer Science **Sept 2010 – Aug 2014**
Joseph Ayo Babalola University, Nigeria

TRAINING AND CERTIFICATIONS

SQL for Data Analytics and Data Science (Udemy) • Learning SQL Programming (LinkedIn Learning) • Getting Started with Big Data and the Hadoop Ecosystem (Udemy) • Data Analytics and Statistical Modelling in R (Udemy) • Foundations: Data, Data, Everywhere (Coursera) • Data Science for All (DS4A), Correlation One, 2022

RESEARCH AND PUBLICATIONS

- Co-author of peer-reviewed papers at NAACL 2024 (AfriMTE and AfriCOMET), ACM SIGIR 2024 (CIRAL), IJCNLP-AACL 2023, and the SustaiNLP workshop at EMNLP 2022 (AfroLM), and of journal articles on health data and AI; 212 Google Scholar citations (h-index 5) as of July 2026.
- Co-patentee on a Nigerian business method patent for an AI-driven drug repurposing system (2025); inventor on a U.S. provisional patent application for disease outbreak forecasting (filed 2026, not examined).
- Founder of the EquiHealth Institute, an independent research initiative on healthcare access (2025 – Present).