ABUCHI OKEKE

Senior AI/ML Engineer · Data Engineer · LLM Systems · Data Platforms
San Francisco, USA • okekeag@gmail.com • +1 408-769-0346 • abuchiokeke.com • LinkedIn • Google Scholar

PROFESSIONAL SUMMARY

Senior AI/ML and data engineer with 10+ years shipping scalable systems and 4+ years productionizing enterprise GenAI at Fortune 50 scale. Owns end-to-end RAG, LLM evaluation pipelines, agentic workflows, and streaming data platforms. Shipped a RAG supply-chain forecasting engine that lifted accuracy by 30%, and an LLM eval harness that gates releases on faithfulness and contradiction scores. Ph.D. candidate researching memory architectures for long-running LLM agents.

CORE COMPETENCIES

LLM & GenAI	Production RAG, hybrid retrieval & reranking, prompt & retrieval tuning, function calling, agent workflows, multi-agent orchestration, guardrails, structured output, streaming.
Evaluation & Fine-tuning	Offline/online evals, golden sets, faithfulness & hallucination scoring, regression suites, LLM-as-judge, LoRA, PEFT, RLHF/DPO, distillation, quantization (GGUF, AWQ).
ML & MLOps	XGBoost, LightGBM, PyTorch, TensorFlow, scikit-learn, time-series forecasting, recommender systems, MLflow, model registry, drift monitoring, A/B testing.
Data & Streaming	Spark, Spark Structured Streaming, Kafka, Flink, Delta Lake, Iceberg, Databricks, Snowflake, BigQuery, Redshift, Airflow, dbt, CDC pipelines.
Backend & Systems	Python, Java, Scala, Go, TypeScript, FastAPI, Spring Boot, gRPC, GraphQL, event-driven microservices, Postgres, Redis, high-throughput services.
Cloud & DevOps	AWS, Azure, GCP, Docker, Kubernetes, Helm, Terraform, GitHub Actions, Jenkins, Datadog, Grafana, Prometheus, cost optimization.

PROFESSIONAL EXPERIENCE

Founder & Lead AI Engineer Ryko.AI

Remote | Feb 2025 Present

- Founded and shipped **Ryko.AI** an LLM-powered study companion with personalized study plans, an AI tutor, smart notes, automated quiz generation, and a global student community (ryko.ai).
- Designed the **multi-agent tutor architecture**: planner, retriever, solver, and critic agents coordinating over course material with tool/function calling, structured outputs, and guardrails.
- Built the **RAG layer** over textbooks, lecture notes, and user uploads hybrid retrieval (BM25 + dense embeddings), reranking, chunking strategies tuned per content type, and citation-grounded answers.
- Engineered an **LLM evaluation harness** with golden sets, faithfulness and on-syllabus scoring, hallucination detection, and regression suites that gate every model and prompt change before release.
- Implemented **personalized study plans** driven by spaced-repetition signals, mastery tracking, and adaptive difficulty; closed the loop with telemetry on student outcomes.
- Shipped **smart notes and automated quiz generation** (multiple-choice, short-answer, free-response) with grading rubrics, explanations, and difficulty calibration.
- Built and operate the backend in **Python / FastAPI** on managed cloud infrastructure: auth, billing, vector store, async job queues, observability, and per-tenant rate limiting.
- Delivered the **web and mobile clients** (React / React Native) including offline support, streaming responses, study-plan UI, and the student community feed.
- Integrated **Stripe-based payments and subscriptions**, trial flows, entitlement management, and revenue analytics.
- Run **growth, product analytics, and SEO** end-to-end funnel instrumentation, A/B tests on onboarding, retention cohorts, and content marketing and iterate the roadmap from those signals.
- Own the full stack end-to-end: product, design, ML, backend, web/mobile, payments, growth, and customer support.

Senior Software Engineer Walmart

Sunnyvale, USA | Jul 2021 Present

- Led design and production delivery of Python-based LLM applications RAG workflows, prompt and retrieval tuning loops, and agentic assistants that materially improved decision quality and reduced manual effort for thousands of internal users across merchandising, supply chain, and operations.
- Architected and shipped a RAG-powered supply-chain forecasting engine, grounding LLM responses in fresh telemetry and historical demand signals; lifted predictive accuracy by 30% and shortened planning cycle time for category teams.

- Designed and operate an end-to-end LLM evaluation pipeline offline and online evals, regression suites, golden-set benchmarking, faithfulness scoring, hallucination and contradiction detection that gates every release on quality deltas and surfaces drift early.
- Built hybrid retrieval and reranking stacks (BM25 + dense embeddings + cross-encoder rerankers) with chunking, metadata filtering, and tool/function-calling patterns to raise answer relevance across multiple business domains.
- Productionized classification, forecasting, and generative models with PyTorch, TensorFlow, and scikit-learn behind FastAPI on Kubernetes; established CI/CD, model registry, feature store, and observability standards adopted across the org.
- Implemented LoRA/PEFT fine-tuning and prompt-engineering workflows for domain-specific tasks; integrated GPT, Gemini, and Claude APIs alongside in-house models behind a unified inference gateway.
- Owned end-to-end ingestion and processing for petabyte-scale data pipelines supporting analytics and ML across streaming, batch, and backfill workflows for hundreds of internal consumers.
- Led the migration from batch to near-real-time ingestion using Spark Structured Streaming and Kafka, reducing data latency from hours to under 5 minutes and dramatically improving freshness for downstream dashboards and models.
- Built ETL and backfill frameworks in Spark, PySpark, Hive, and Airflow with idempotency, deduplication, and replay support cutting reprocessing effort by 30% and improving correctness under late-arriving data.
- Re-architected critical pipelines through partition strategy, bucketing, broadcast joins, and resource tuning improving throughput by 40% while lowering compute cost by 15%.
- Built and maintained governed data lake foundations on GCP and Azure with standardized schemas, access controls, and high-availability datasets; authored and tuned complex OLAP SQL / HiveQL queries that cut heavy report runtimes by 50%.
- Developed reusable feature-engineering pipelines for ML and LLM workflows, and integrated vector stores and embedding pipelines into the platform.
- Deployed and operated data and ML services on Kubernetes with health checks, autoscaling, and baseline observability; managed infrastructure as code (Terraform/Helm) to make deployments repeatable and reduce environment drift.
- Integrated dashboard data sources via internal REST APIs and semantic/caching layers to unify real-time telemetry with historical warehouse context and keep dashboards responsive over large datasets.
- Supported ML monitoring signals drift, anomaly scores, feature-contribution summaries for stakeholder interpretation, model auditability, and enterprise data-privacy compliance.
- Owned on-call for production data and ML pipelines, driving incident response, root-cause analysis, and permanent fixes; improved MTTR by 25%.
- Acted as technical lead in Agile delivery planning, sprint execution, retrospectives and partnered with product, analytics, and engineering teams to define SLAs and deliver roadmap initiatives.
- Mentored junior and mid-level engineers through design and code reviews, influenced technical hiring, and helped shape the ML platform roadmap.

Big Data Engineer Enhance IT

Atlanta, USA | Sep 2020 Jun 2021

- Designed and built real-time and batch data pipelines using PySpark, Spark Structured Streaming, Flume, Nifi, Kafka, Airflow, and Hive, supporting analytics and ML workloads at scale.
- Installed, configured, and operated end-to-end Apache Spark pipelines (Python and Scala) integrated with multiple RDBMS, HDFS, and Hive, and tuned them for throughput and reliability.
- Ingested structured, semi-structured, and unstructured data into HDFS using Spark RDDs, DataFrames, and Dataset APIs across file formats including AVRO, ORC, Parquet, CSV, and JSON.
- Imported and synchronized data from heterogeneous sources MySQL, PostgreSQL, SQL Server into HDFS, and exported curated datasets from Hive internal tables back into external RDBMSes for downstream consumers.
- Used Spark SQL and Hive (internal and external tables) for querying, transformations, and writing back to RDBMSes; enabled Hive Support inside Spark pipelines for unified analytics.
- Built large-scale web-crawling workflows with URL-queue management, rate-limiting, and retries; enforced governance with robots.txt compliance and content normalization, and integrated crawl output into Spark and Kafka pipelines.
- Collected, organized, and modeled data and trained supervised ML models reaching 92% accuracy; integrated visualization layers using Pandas, Matplotlib, and Power BI.
- Developed executive analytics dashboards in Power BI, Grafana, and Tableau, surfacing operational and ML signals for non-technical stakeholders.

- Orchestrated and scheduled all workflows in the Apache Spark pipeline using Airflow, authoring DAGs that decomposed jobs into modular, retryable tasks.

Graduate Research Assistant University of Kentucky

Lexington, USA | Jan 2019 Aug 2020

- Planned, modified, and executed research techniques, procedures, and experiments in the Biosystems and Agricultural Engineering lab, focused on applied ML for food-safety detection.
- Designed experiments, collected and curated datasets, performed exploratory analysis, and extracted insights from FTIR spectroscopy and related sensor data.
- Prototyped machine learning models for classification, regression, and prediction, then performed rigorous model evaluation, error analysis, and result visualization.
- Built end-to-end ETL data pipelines for research data processing ingestion, cleaning, feature extraction, and dataset versioning to support reproducible experiments.
- Deployed and integrated ML models using Pickle, Joblib, and MLOps workflows on AWS and Azure, exposing inference endpoints for downstream lab tooling.
- Prepared materials for reports and presentations, and authored / submitted peer-reviewed conference and journal publications on the research outcomes.
- Designed and developed **Glutini** (glutini-res.com) a smart, rapid analytical tool for authenticating gluten-free, grain-based foods as the core M.Sc. research project for the Biosystems and Agricultural Engineering Department.
- Collected and labeled training samples using the **ELISA technique** (RIDASCREEN Gliadin R7001 kit) sample preparation, supernatant extraction with Cocktail R7006, dilution, buffer washing, enzyme activation/deactivation, and absorbance reading to quantify gluten content for supervised learning.
- Built the end-to-end ML development pipeline for Glutini: sample data collection, pre-processing, supervised model prototyping, and evaluation for gluten classification and quantification on grain-based foods.
- Shipped the **Glutini mobile application** with an embedded **TensorFlow Lite** image-classification engine for on-device inference supporting new scans via an external scanning device, uploads from phone storage, on-device analysis, and result visualization.
- Collaborated with an advisory committee of professors across Biosystems Engineering and Electrical & Computer Engineering, and coordinated with Chemistry researchers on assay design, data labeling, and validation protocols.

Software Engineer Moniepoint Inc

Lagos, Nigeria | Mar 2018 Dec 2018

- Built and shipped **native Android applications in Java** for widely used fintech and mobile banking products, owning features end-to-end from spec to Play Store release.
- Designed and implemented responsive Android UIs in **XML** using ConstraintLayout, RelativeLayout, LinearLayout, Shape Drawables, Selectors, custom styles/themes, vector assets, and 9-patch resources strictly following Material Design guidelines across multiple screen sizes and densities.
- Developed reusable custom views and compound components (custom TextViews, keypads, OTP inputs, transaction list items, bottom sheets) to standardize the look and feel across banking apps and accelerate future feature delivery.
- Implemented core fintech flows onboarding/KYC, account opening, login with PIN/biometrics (Fingerprint/BiometricPrompt), fund transfers, bill payments, airtime/data top-up, card management, and transaction history wired to secure REST APIs over HTTPS.
- Integrated banking and payment backends using **Retrofit, OkHttp, Gson, and RxJava** with request signing, token-based auth, certificate pinning, encrypted SharedPreferences, and Android Keystore for sensitive credentials and session keys.
- Wrote production-grade Java covering the Android system display pipeline (drawing, rendering, compositing) and authored **JNI bridges** to native C/C++ code for performance-critical and security-sensitive operations.
- Tuned application performance and battery efficiency reduced ANRs and cold-start time, profiled with Android Studio Profiler and StrictMode, optimized RecyclerViews, image loading (Glide/Picasso), and background work via Services, AsyncTask, and WorkManager.
- Hardened the codebase against fintech-grade threats input validation, secure storage, anti-tampering checks, ProGuard/R8 obfuscation, root detection, and safe handling of card and PII data in line with **PCI-DSS-aware** practices.
- Implemented push notifications (Firebase Cloud Messaging) for transaction alerts and OTPs, deep links for promotional and in-app navigation, and in-app analytics/crash reporting (Firebase Analytics, Crashlytics).
- Built and maintained unit and instrumentation tests with **JUnit, Mockito, and Espresso**; set up Gradle build variants and flavors to manage staging/production environments and white-labeled bank builds from a shared codebase.

- Owned release engineering for signed Android APKs versioning, keystore management, staged rollouts on the Google Play Console, post-release crash triage, and hotfix delivery.
- Collaborated daily with product, design, QA, and backend teams in an **Agile/Scrum** workflow participated in code reviews, mentored junior Android engineers, and contributed to architectural decisions (MVP/MVVM, modularization).
- Shipped features into widely used financial mobile banking apps including **Sterling OnePay, Sterling AltPay, and Unify by Unity Bank**, reaching hundreds of thousands of end users across Nigeria.

Web Developer Belle Naturals NG

Lagos, Nigeria | Jan 2015 Feb 2018

- Designed, built, and operated a full e-commerce platform on WooCommerce / WordPress (bellenaturels.com) product catalog, variants, inventory, checkout, shipping zones, tax rules, and order management.
- Integrated payment gateways (Paystack, Flutterwave, PayPal) and reconciliation flows; implemented coupon, discount, and abandoned-cart recovery logic.
- Developed and customized WordPress themes and plugins in PHP, including custom post types, REST endpoints, and admin dashboards tailored to the merchandising workflow.
- Built companion web applications in Python (Django and Flask) following an MVC architecture for inventory, CRM, marketing automation, and internal reporting.
- Modeled and tuned the data layer in MySQL/PostgreSQL, wrote efficient queries and migrations, and exposed REST APIs consumed by the storefront and back-office tools.
- Implemented responsive front-ends using HTML5, CSS3, JavaScript, jQuery, and Bootstrap; ensured cross-browser compatibility and mobile-first experiences.
- Owned on-page and technical SEO schema markup, sitemaps, canonical tags, page-speed optimization, image compression, and Core Web Vitals driving organic traffic and conversion lift.
- Set up analytics, conversion tracking, and pixel integrations (Google Analytics, Search Console, Facebook Pixel) and used the data to guide UX and merchandising decisions.
- Ran digital marketing and growth: email campaigns (Mailchimp), content marketing, social commerce, and paid acquisition on Meta and Google.
- Hardened the site for security and uptime SSL, automated backups, malware scanning, CDN/caching and managed Linux-based hosting, DNS, and deployments.
- Provided customer support, processed orders, coordinated shipping, and iterated the storefront based on user feedback and sales data.

SELECTED PROJECTS

- **RAG-Powered Supply-Chain Forecasting** Grounded LLM responses in fresh supply-chain telemetry; +30% predictive accuracy. (Python, FastAPI, PyTorch, LangChain, Kafka, Spark, Kubernetes, GCP)
- **Production LLM Evaluation Pipeline** Offline + online harness with golden sets, faithfulness and contradiction scoring, and release gating.
- **Summarization vs RAG for Long-Term LLM Memory** Ph.D. dissertation comparing memory architectures for long-running conversational agents.
- **Glutini Research** FTIR spectroscopy + ML to detect and quantify gluten contamination in grain-based foods (glutini-res.com).

EDUCATION

Ph.D. in Information Technology (AI / LLM specialization) University of the Cumberlands
Dissertation: Summarization vs Retrieval-Augmented Generation for long-term LLM memory.

M.Sc. in Biosystems Engineering Machine Learning specialization University of Kentucky

B.Sc. in Environmental Engineering University of Ibadan, Nigeria

PUBLICATIONS & CERTIFICATIONS

- Adedeji, A. A., Okeke, A., & Rady, A. M. (2023). Utilization of FTIR and machine learning for evaluating gluten-free bread contaminated with wheat flour. *Sustainability*, 15(11), 8742.
- Okeke, A. G. (2020). Fourier Transform Infrared Spectroscopy Coupled with Machine Learning Approaches for Detection and Quantification of Gluten Contaminations in Grain-Based Foods.
- Certification: Microsoft Azure AI Engineer Associate.