Syed Usman Bukhari

Telephone: 07462660889 | Email: usmanbukhari541@gmail.com | Portfolio: usmanbukhari.co.uk | Github : github.com/UsmanBuk

---

Results-driven Full Stack & DevOps Engineer with extensive experience in designing, optimising, and securing scalable backend systems and cloud infrastructure. Skilled in building robust APIs, automating deployments, and implementing CI/CD pipelines to enhance software delivery efficiency. Proven ability to improve system performance, migrate legacy applications to modern architectures, and enforce best security practices. Adept at leveraging technologies like Kubernetes, Terraform, and cloud platforms to drive operational excellence.

Key Skills / Tech Stack

Backend Development | API Design & Microservices | Cloud Architecture (AWS, Azure, GCP) | DevOps & CI/CD Pipelines | Infrastructure as Code (Terraform, Ansible) | Python (Django, Flask) | Node.js (Express.js) | C# | PostgreSQL, MongoDB, MySQL | DynamoDB | GraphQL, REST, OpenAPI | AWS API Gateway, Lambda, Step Functions | System Scalability & Security (JWT, OAuth2, RBAC, API Rate Limiting) | ETL Pipelines & Data Processing | Performance Optimization & Caching (Redis, CloudFront) | TDD & Continuous Security | Shopify (Liquid) | Containerization (Docker, Kubernetes) | (Git, BitBucket, Jira, Trello) | Apache airflow | Supabase | Low code tools | Lovable | Cursor | Claude Code | MCP | Linux | Pytest | PyTorch | Tensorflow | CI/CD | Playwright | Asynchronous (asyncio) | Langchain | Snowflake | FASTAPI | Outsystems | Powertools | Gradio | Hugging Face | Llamaindex | CosmoDB | KQL | Pyspark | NLP | Computer vision| MLops

---

Career Summary

May 2025 - Present **Gen AI Developer & Data Architecture Specialist -** NHS - South Yorkshire ICB

Greenfield Project

Production RAG System & LLM Orchestration

- Engineered production-grade **RAG system** using **Azure AI Agent SDK** with LangChain custom retrieval chains and LlamaIndex for input sanitization/enhancement, implementing automated critique loops and **multi-source grounding (Bing + proprietary data)**. Achieved 90% accuracy validation against NHS clinical guidelines with zero hallucination incidents in production
- **Built production Azure AI Agent with real-time web grounding via Bing Search API**, enabling dynamic knowledge retrieval beyond static datasets. Implemented function calling patterns and tool orchestration to blend real-time web data with proprietary NHS knowledge base, reducing outdated information issues by 95% and enabling responses to emerging health guidance within 24 hours of publication
- Developed **agentic LLM orchestration** pipelines with automated content enrichment (auto-tagging, extractive/abstractive summarization, ML-powered metadata augmentation), **integrating Azure AI Agent's tool-use capabilities for dynamic data validation**. Improved data completeness by 73% and search relevance by 61%
- Implemented advanced vector search capabilities using **Azure AI Search** with 768-dimensional embeddings (OpenAI Ada-003), to build **hybrid search architecture** (lexical + semantic + real-time web) that improved service discovery accuracy by 85% and reduced query resolution from 12 minutes to under 30 seconds

Full-Stack AI Application | Microsoft Framework Enhancement

- Forked and extensively customized Microsoft's sample-app-aoai-chatGPT framework with **Azure AI Agent integration**, multi-modal search, context-aware dialogue management, **Bing grounding for real-time information**, and custom knowledge grounding and **Reflexion agent pattern** for self correcting retrieval.
- **Stack**: Python, React (Fluent UI), TypeScript, Quart (async Flask), Azure CosmosDB (conversation history/memory), Azure CI/CD
- Deployed to production serving 3,000+ monthly active users with 94% query success rate and 4.8/5 satisfaction

MCP Server Integration | Patient-Facing AI Access

- Developed Model Context Protocol (MCP) server to expose RAG chatbot capabilities directly within the NHS South Yorkshire GP app, enabling patients to query local health services, commissioned care, and real-time guidance without leaving their existing care management interface
- Architected server to handle patient queries with full access to underlying RAG pipeline, vector search, and Bing grounding—delivering the same accuracy and real-time capabilities as the standalone application
- Extended AI assistant reach to South Yorkshire residents through native integration with existing NHS digital infrastructure

Azure MLOps & Data Infrastructure

- Designed event-driven data pipeline on **Azure Functions** with serverless microservices, queue triggers, and automatic failover, achieving **99.7% uptime** while processing 15,000+ daily records with sub-5-second latency
- Architected fault-tolerant web scraping infrastructure using **Crawl4AI** (Python) with custom middleware, extracting structured data from 50+ regional websites supporting 1.4M population across South Yorkshire
- Standardized disparate healthcare datasets into AI-optimized JSON schemas with semantic validation, enabling **API integration** with NHS systems and real-time federation across 12 regional health authorities

Technical Skills

- **AI/ML:** Azure AI Agent, **Agent Orchestration**, LangChain, LlamaIndex, RAG, **Bing Grounding**, Azure OpenAI (GPT-4o mini), **Azure AI Search**, **Function Calling**, vector search, embeddings, conditional GANs, NLP, prompt engineering, **Tool-use Pattern,MCP (Model Context Protocol).**
- **Development**: Python, TypeScript, React, Quart(Async Flask), REST APIs, event-driven architecture
- **Cloud/DevOps**: Azure Functions, CosmosDB, Azure CI/CD, serverless architecture, microservices, Azure Durable Functions
- **Data**: Crawl4AI, ETL pipelines, vector databases, hybrid search, data standardization

---

June 2024 – April 2025 - Contractor **Sage Intacct full Stack developer & Solutions Architect** – TSG/Dayta (Sage Partner)

- Designed, implemented, and optimized end-to-end **CI/CD pipelines using AWS CodeBuild, CodePipeline**, and API Gateway, achieving streamlined deployments and significantly enhanced software delivery efficiency.
- Architected and maintained serverless applications leveraging **AWS Lambda** and Step Functions to orchestrate sophisticated, event-driven workflows, improving automation and reliability.
- Integrated security best practices across development and deployment processes by enforcing **TDD, continuous security practices, authentication, encryption standards, and comprehensive API security.**
- Built and deployed a two-way integration application hosted using **Docker** and **Kubernetes**, managed CI/CD processes via **GitHub and Jenkins**, ensuring robust automation and streamlined continuous delivery.
- Developed and automated internal tools with **Flask and React on AWS Amplify,** significantly enhancing developer productivity, client support capabilities, and workflow automation.

Jan 2022 – Feb 2024-Contractor **Full Stack Engineer** – Citadel Health

- Developed **RESTful APIs** using **Django REST Framework (DRF) and OpenAPI** to support high-volume LIMS workflows.
- Designed cloud-native solutions leveraging AWS services (**EC2, S3, RDS, Lambda, API Gateway, IAM**) to deliver robust and scalable LIMS applications.
- Developed comprehensive API documentation using **Swagger and OpenAP**I, facilitating improved developer collaboration and efficiency.

Jan 2019 – Jan 2022 **Software/Data Engineer** – Chalkys.com (Shopify)

- **Built an AI-driven recommendation algorithm using AWS SageMaker and Databricks**, leveraging ML models to analyze user behavior patterns. Utilized **Databricks' collaborative notebooks** for exploratory data analysis and feature engineering, processing millions of customer interaction events. Deployed models enabling dynamic recommendations such as "Recently Viewed", "You May Also Like" and "People usually buy with", increasing upsell conversions by 25%.
- **Implemented an image recognition system using AWS SageMaker** and ML techniques to automatically generate descriptive alt-text and tag product images, significantly enhancing SEO
- **Designed and orchestrated ETL pipelines using Apache Airflow and Databricks**, automating competitor price analysis & product data scraping workflows. Utilized **Databricks SQL** for real-time analytics and reporting dashboards, processing structured and semi-structured data with scheduled DAGs for reliable data processing. Implemented comprehensive monitoring using **Grafana** to track pipeline performance, with Databricks handling large-scale data transformations and aggregations.
- Modernized product catalog content generation using **LangChain** + **Tavily** + **GPT-4o**, replacing legacy Google Search-based pipeline. Built a **RAG (Retrieval-Augmented Generation) system** that automatically researches products via web search and generates SEO-optimized descriptions. Implemented **LangSmith** for tracing, debugging, and API cost monitoring. Deployed with **SQLite caching** to minimize redundant API calls. System processes millions of SKUs with automated Shopify **GraphQL** integration, reducing manual content creation by 90%

Education and Professional Training

AWS Cloud Practitioner Certification & Solutions Architect Certification – eCareers
Data Engineering Program – Ai Core
BSc Engineering Mathematics – University of Bristol
Portfolio: usmanbukhari.co.uk
Live demos, open-source contributions, further contract roles and extended project history