Aman Kumar Sahu

+91 9664866438 ⬨ Github ⬨ Linkedin ⬨ sahuaman454@gmail.com

I am a passionate Data Scientist with over 3+ years of experience building ML and GenAI solutions, with core expertise in LLM pipelines, GraphRAG, Advanced RAGs, Text-to-SQL, Finetuning LLMs and knowledge graph-driven automation for real-world enterprise use cases.

TECHNIAL STRENGTHS

Languages	Python, SQL, C, C++, Scala (Beginner)
Technologies	Machine Learning, LLM, AI and Deep Learning, Gen-AI, ChatGPT, RAG, GraphRAG, Vector DB, Retrievers
Framework/Libraries	Pandas, NumPy, Scikit, TensorFlow, Keras, Tableau, Langchain, LangGraph, Langsmith, Dash Leaflet, Plotly Dash
Vector Store/VectorDB	FAISS, ChromaDB , Neo4j, Pinecone

WORK EXPERIENCE

KPMG Global Services- *Data Scientist Consultant - Gen AI (Full time)* ***Sept 2025 – Present***

- **Contract Data Extraction (RAG-based):** Built a RAG pipeline using **Azure OpenAI + Azure AI Search** to extract **20 key fields** from **2,400+ unstructured contract documents**. Created a **135-record annotated dataset**, upsampled low-frequency fields, and iteratively refined prompt-based extraction logic. Delivered high-accuracy extraction models for **5 complex fields**, optimized for recall and accuracy.
- ***Technology Stack: Python, NumPy, NLTK, Scikit-learn, Seaborn, Plotly and Matplotlib, Gen-AI, Langraph, Langchain, Prompt Engineering, vector stores, RAG, Azure OpenAI, Azure AI Search.***

Arcadis - *Data Scientist Associate Consultant (Full time)* ***June 2023 – Aug 2025***

- **Tap Card Digitization:** Led the end-to-end digitization of **10.6K+ Tap Cards**, building a template-classification pipeline and training **Azure Custom Document Extraction models** to extract key fields (e.g., SVC pipe size, date), delivering structured outputs for QC review.
- **GraphRAG using Neo4j:** Automated structured field extraction from unstructured legal PDFs by combining AzureOCR with LangChain's LLMGraphTransformer to build a Neo4j-based knowledge graph, enabling Cypher querying and LLM-powered Q&A over graph data.
- **Text to SQL (Phase 2):** Enhanced the system by introducing LangChain's **SQLToolkit** with **ReAct** Agent and migrating the architecture using LangGraph, enabling multi-turn query handling and Human in the Loop.
- **Text to SQL (Phase 1):** Developed a Text to SQL system using Streamlit and OpenAI – **GPT – 4o**, reducing data retrieval times by 50% and enhancing accessibility for non-technical users.
- **Retrieval-Augmented Generation (RAG) System:** Deployed a RAG system using LangChain, FAISS, **Google Flan T5** and Streamlit, enhancing data retrieval accuracy by 40% and completing the project within a month.
- **NRW Leak Detection:** Developed a GNN-based (**ChebNET Model**) leak detection system for WDNs, **achieving a 87% accuracy rate** in leak localization and reducing **water loss by 15%.**
- **IT Ticket Analysis:** Implemented advanced NLP techniques (keyBERT, Gensim**, Latent Dirichlet Allocation (LDA)**, GPT models) for efficient IT support, reducing ticket resolution time by **30% and improving service quality**.
- **Dynamic Spatial Analytics Web App:** Created a spatial data web app using Plotly Dash, Python, PostgresSQL and ArcGIS, improving spatial data utilization by **35%** and aiding decision-making processes.
- ***Technology Stack: Python, NumPy, NLTK, Scikit-learn, TensorFlow, Keras, Seaborn, Plotly and Matplotlib, Gen-AI, Langraph, Langchain, Prompt Engineering, vector stores, RAG, llama, MS Fabric and PowerBI, Gradio, KNIME, Plotly Dash, Dash Leaflet***

GEP Worldwide - *Data Engineer / Data Scientist (Full time)* ***June 2022 – June ,2023***

- Enhanced data quality and efficiency, **reducing data errors by 30%** and **increasing processing speed by 20%.**
- Conducted EDA, improved model **performance by 15%,** and optimized model accuracy.
- ***Technology Stack: Azure Data Bricks, Scala/Python, Spark, Tableau, NumPy, Scikit, Pandas and Matplotlib.***

Myraa Technologies - *Machine Learning Engineer (Internship)* ***July 2021 – June 2022***

- Spearheaded a resume ranking project, **increasing candidate selection efficiency by 40%** using TF-IDF, CBOW, Word2Vec, and GloVe embeddings, **enhancing accuracy by 30%**.
- ***Technology Stack: Python, NLTK, SpaCy, NumPy, Pandas and Matplotlib. Python, TF-IDF Embeddings, CBOW (Continuous Bag of Words), Word Embeddings (e.g. Word2Vec, GloVe), Document Embeddings (e.g., Doc2Vec)***

EDUCATION

K.J Somaiya Institute of Engineering and Information Technology, Mumbai

- B. Tech, Information Technology - *2018 - 2022* ***CGPI: 9.04***

Kendriya Vidyalaya Bhandup, Mumbai

- Higher Secondary Certificate - *May 2018* ***81.80 %***
- Secondary School Certificate - *May 2016* ***91.20 %***

PROJECTS

Nano Satellite Project: BeliefSat-0 - *Command and Telemetry* **July 2019 - June 2022**

- Developed BeliefSat-0, a 2p-PocketQube nanosatellite, as part of K.J.S.I.T Satellite Group.
- Promoted to Senior Command and Telemetry Developer for leadership and contributions.
- Successfully launched BeliefSat-0 aboard **ISRO's** PSLV-C58 XPoSat on January 1, 2024.
- Enhanced team efficiency **approximately by 20%** and **reduced development time by 10%.**

SOCIAL ACTIVITIES AND ACHIEVEMENTS

- **IITM BS Degree –** Completed Foundation (FY) **with 8.00 CGPA** currently pursuing **Diploma in DS with an 8.12 CGPA.**
- **Arcadis** - Awarded with **2 Spot Bonuses** in a span of **four months** (**Nov 2024** and **Feb 2025)**
- **Organizer | Somaiya Space Conclave** **February 2020**