Taimour Abdul Karim

Lahore, Pakistan | +923261127700 | taimour.a.karim@gmail.com
Github | LinkedIn | Portfolio

Summary

Data Scientist and AI Engineer with 3+ years of experience building production machine learning and generative AI systems. Specializes in Large Language Model (LLM) evaluation and safety, Retrieval-Augmented Generation (RAG), and AI agent systems. Leads platform development end to end, owning architecture, delivery, and operations from data pipelines to deployment.

Technical Skills

Gen AI / LLM Engineering: Python, C/C++, SQL, Pandas, Scikit-Learn, PyTorch, Keras, TensorFlow, Huggingface, AWS Bedrock, OpenAI APIs, LangChain, LangGraph, FAISS, ChromaDB, Pinecone, llamacpp
Infrastructure: FastAPI, Docker, AWS, TimescaleDB, Railway, Vercel, GitHub Actions, Redis, MQTT

Work Experience

Datality, London (Remote) **Jan 2024 – Present**

AI/ML Engineer

- Led the architecture, design, and development of Bryge.io, a multi-tenant industrial IoT analytics platform that infers unknown sensor schemas at runtime and answers questions in plain English over live MQTT streams.
- Engineered a streaming LLM agent on AWS Bedrock with 15 schema agnostic tools and adaptive reasoning budgets.

Stack: FastAPI, React + TypeScript, TimescaleDB, AWS, Redis, MQTT, pgvector

Qult Technologies, Lahore **Aug 2023 – Dec 2024**

AI/ML Engineer

- Fine-tuned Llama 3.1 8B for text classification, improving accuracy by 20% over the base model on Natural Language Understanding (NLU) tasks.
- Automated end-to-end machine learning pipelines leveraging AWS services and OpenAI tools to streamline workflows, ensuring scalability, efficiency, and quick deployment of ML solutions.

Stack: Python, LLMs, AWS (SageMaker, Lambda), Docker, TensorFlow

BornGreat, California (Remote) **Sep 2022 - Feb 2023**

Data Scientist

- Delivered sentiment analysis over social media streams using NLP and LLMs, for product decisions.
- Developed a FastAPI inference service with GitHub Actions CI/CD, deploying automatically on every merge.

Stack: Selenium, NLP, LLMs, FastAPI, GitHub Actions, Docker, TensorFlow

Project Work

- **Clinical LLM Bias Audit:** Built a reproducible fairness audit framework that measures whether a clinical LLM changes its care recommendation when only the patient's geography or name is perturbed. Implemented Wilcoxon, BCa bootstrap, and Cohen's, ran 300 cases per model across 3 providers, and reported a calibrated null result with confidence intervals.
- **SEC RAG Analyst:** Built a production RAG assistant over SEC 10-K filings using four stage retrieval: BM25 and dense embeddings in FAISS, fused with Reciprocal Rank Fusion and reranked by a cross-encoder, then answered by Claude with inline citations.
- **Credit Default MLOps**: Built an end-to-end MLOps pipeline around an XGBoost credit-default classifier: DVC versioned data, MLflow tracking and registry, an automated CI quality gate that exits non-zero and blocks any model below the ROC-AUC, PR-AUC, and Brier floors, PSI drift detection, and FastAPI serving instrumented for Prometheus and Grafana.
- **Uplift Targeting Engine**: Built a causal uplift engine that predicts the incremental effect of an intervention per user. Implemented S, T, X, and R meta-learners from scratch over XGBoost, evaluated with Qini curves and policy value, and cross-checked the R-learner against Microsoft econml at 0.96 rank correlation. On the 64k-row Hillstrom experiment the policy beat random targeting at a fixed 30% budget.

- **LLM Evaluation and Safety Suite**: Built five evaluation tools: a red-teaming framework (520 AdvBench prompts, 40+ categories, before/after hardening), a hallucination benchmark by topic on TruthfulQA (817 questions), an LLM-as-judge system calibrated against blind human labels (800 scores), CI prompt-regression gates (55-case golden suite), and a RAGAS-based RAG evaluation framework.
- **everytongue and site2bot:** Shipped two installable open-source tools. everytongue fine-tunes a neural translator for any language from a spreadsheet using an NLLB-200 low-resource recipe, demonstrated on Q'eqchi' from 656 pairs. site2bot turns any website into a fully offline RAG chatbot in about 600 lines of Python, no API keys and no vector database.

Education

MSc. in Artificial Intelligence | Lahore University of Management Sciences (LUMS)

BSc. in Data Science | National University of Computing and Emerging Sciences

Awards and Certificates

- Earned 2,500+ GitHub stars and 500+ forks across open-source repositories
- 2nd Position, Genesis Hackathon Dubai
- Data Manipulation, DataCamp
- Joining Data with Pandas, DataCamp