Muzammil Bin Sohail

Lahore, Pakistan
muzammilsohail1718@gmail.com LinkedIn GitHub Portfolio

Education

FAST - National University of Computer & Emerging Sciences (NUCES) **2022 – Present**
Bachelor of Science in Artificial Intelligence *Faisalabad, Pakistan*

- CGPA: 3.39 / 4.00 **(Ranked 1st in Cohort)**
- **Honors:** 4× Gold Medals, 1× Silver Medal, 3× Dean's List

Experience

Founding AI Engineer — Entracloud **Jul 2024 – Present**
Faisalabad, Pakistan

- Architected production-grade AI systems, wrapping ML models in FastAPI services for seamless cloud integration.
- Optimized inference infrastructure, reducing latency by 40% for real-time user applications.
- Led the translation of business requirements into technical AI specifications and scalable software solutions.

Teaching Assistant — FAST (NUCES) **Jan 2025 – Jan 2026**
Faisalabad, Pakistan

- Assisted in teaching *Deep Learning*, *Programming for AI*, and *Database Systems* to 150+ undergraduates.
- Mentored students on software engineering best practices, including API design, Dockerization, and Git workflows.

NLP Intern — Elevvo Pathways **Jul 2025 – Aug 2025**
Remote

- Developed an NLP text classification pipeline using Python and Scikit-learn to categorize unstructured text data.
- Implemented data preprocessing techniques (tokenization, lemmatization) to improve model inference accuracy.

AI Solutions Engineer — Moonbridge Systems **Nov 2023 – Jun 2024**
Remote / Contract

- **Delivered 200+ AI solutions** for global clients, serving as the lead engineer for high-volume agency accounts.
- Developed custom RAG pipelines and chatbot agents using LangChain and OpenAI, maintaining a 5-star success rate.
- Managed full lifecycle deployment of AI microservices, handling client requirements from concept to production.

Projects

Suspicious Behavior Detection System (3D CNNs & MIL)

- Architected a real-time surveillance system using **3D ResNet-18** and **Multiple Instance Learning (MIL)** to detect anomalies in the UCF-Crime dataset, achieving SOTA temporal localization accuracy.

End-to-End MLOps Pipeline (AWS, Airflow, Celery)

- Built an automated ML lifecycle platform using **DVC** for data versioning, **MLflow** for tracking, and **GitHub Actions** for CI/CD.
- Orchestrated ETL DAGs with **Apache Airflow** and deployed **FastAPI** services on **AWS EC2** with **Celery** workers.

Context-Aware Multimodal Recommender (RAG & Fine-Tuning)

- Engineered a hybrid RAG engine fusing user preferences with live **OpenWeatherMap API** data using **FAISS** vector search.
- Fine-tuned transformer models on synthetic data to serve dynamic lifestyle recommendations with millisecond latency.

Multimodal Cancer Diagnostic System (Explainable AI)

- Developed a fusion framework classifying cancer types (Clinical, Genomic, Imaging) using **SHAP** and **LIME** for interpretability.
- Implemented cross-modal attention mechanisms to correlate unstructured medical text with genomic sequences.

Technical Skills

Languages & Frameworks: Python, SQL, C++, FastAPI, Flask, LangChain, TensorFlow, PyTorch, Scikit-learn
AI & GenAI: OpenAI API, Hugging Face Transformers, LLMs, RAG, Fine-Tuning, Computer Vision, Vector Databases (FAISS/Pinecone)
MLOps & Cloud: AWS, Docker, GitHub Actions (CI/CD), MLflow, Apache Airflow, REST APIs, Celery, Redis
Tools & Platforms: Git, Linux, Postman, VSCode

Achievements & Certifications

AWS Cloud Solutions Architect Track: Comprehensive 5-course certification series covering 40+ AWS Services.

Reply AI Coding Contest: Secured **Top 29% Global Rank** in an algorithmic problem solving challenge.

Technical Certifications: Generative AI & LLMs (Duke Univ), Advanced SQL, Flask (Backend Development), Prompt Engineering.