Muhammad Faizan Raza

faizanraza766@gmail.com | (484).320.9679 | Dallas, TX | www.linkedin.com/in/faizanraza03 | www.github.com/faizannraza

Professional Summary: AI/ML Engineer skilled in Python, LLM orchestration, and scalable data systems; deployed production MCP AI workflows at Zscaler achieving 95%+ accuracy, 10x efficiency, and $1M+ annual savings.

EDUCATION

Pennsylvania State University ***Malvern, PA***
Master of Science in Data Analytics, GPA: 4.0 *Aug 2024–May 2026*
Relevant Coursework: Predictive Analytics, Deep Learning, NLP, Large Scale Databases, Data Visualization, Data Mining, Statistics

Lahore University of Management Sciences (LUMS) ***Lahore, Pakistan***
Bachelor of Science in Mathematics and Economics (double major) *Sep 2016–Dec 2020*

SKILLS

- **Languages:** Python (Pandas, NumPy, Scikit-learn, SciPy, PyTorch, TensorFlow, PySpark), JavaScript, SQL, C++, PHP, R, Linux
- **Analysis Techniques:** MCP, Machine Learning, Deep Learning, NLP, Gen AI, Statistical Modeling, Agentic AI, RAG, APIs, vLLMs, MLOps
- **Tools:** LangChain, Redis, DBT, Airflow, Spark, Hugging Face, Kafka, CI/CD, PostHog, AWS, Snowflake, Git, Tableau, Azure, GCP, Docker

PROFESSIONAL EXPERIENCE

AI/ML Software Engineering Intern, Zscaler – *San Jose, CA* *May 2025-Aug 2025*

- Architected and productionized an LLM-powered Model Context Protocol **(MCP) automation** system orchestrating Hadoop→DBT migrations, and pipeline deduplication across large-scale data infrastructure, achieving **95%+ task accuracy**.
- Shipped the system end-to-end using **LangGraph**, Redis, Postgres, and AWS, reducing runtime by **10×**, cutting **latency ~70%**, and driving **~$1M+** annual engineering cost savings across **100+ weekly** workflows in a **500B+/day** events environment.

Graduate Research Assistant (Data Science), Pennsylvania State University – *Malvern, PA* *Sep 2024-May 2026*

- Led end-to-end development of applied ML, NLP, and agentic AI systems analyzing **10M+ record**, multi-source datasets.
- Built and deployed **NLP** pipelines (scraping, embeddings, supervised ML) over 1M+ text records, achieving **~85–90%** classification accuracy on sentiment/risk labeling across energy, healthcare, and security domains.
- Developed and evaluated **predictive and causal** models, improving model accuracy metrics by **~19%** through **feature engineering**, PCA, and statistically rigorous experimentation.

Data Analytics Team Lead (GTM), Beam AI - *Berlin, Germany* *Feb 2023-Aug 2024*

- Built production-grade reporting analytics pipelines and executive dashboards using **Python, SQL, and ETL** workflows, cutting reporting turnaround by **40%** and enabling real-time, data-driven decision-making for leadership.
- Engineered a user recommendation system (collaborative filtering, Scikit-learn) to correct ranking inaccuracies, improving model precision by **30%** and increasing product **conversions by 12%.**
- Developed predictive ROI models using statistical learning (**Pandas, NumPy**) to forecast customer value, directly supporting 5 enterprise PoCs (**~€150K** total contract value) and accelerating sales close cycles.

Data Analyst (Operational Excellence, Daraz PK – *Lahore, Pakistan* *Jan 2021-Dec 2022*

- Led large-scale operations analytics across **50+ First Mile** logistics stations, analyzing 1M+ customer **feedback** records using SQL and Python to identify systemic bottlenecks, driving an **18% increase** in customer satisfaction nationwide.

PROJECTS

Multi-Agent Tornado Crisis Management & Relief Coordinator: Built a multi-agent AI system combining CNN radar models, LoRA-tuned NLP, and RAG over FEMA guidelines, achieving ~0.87 ROC-AUC for real time explainable disaster response (Github).

Scalable Political Finance Analytics Platform (U.S. Election Data): Architected a PySpark-based analytics pipeline for U.S. election donations, optimizing ETL throughput and benchmarking distributed systems to surface statewide donor insights.

Continual Learning for LLMs (STAR + FAR): Designed and evaluated continual learning pipelines for LLMs using STAR and FAR, reducing catastrophic forgetting while maintaining performance under streaming data updates at scale (Github).

RESEARCH PUBLICATIONS

- Unleashing the Potential of Large Language Models: A Blueprint for Real-Time, Enterprise-Ready Deployments. IEEE Computer / IEEE publication. DOI: https://doi.org/10.1109/mc.2026.3664470
- Enhancing Cybersecurity Readiness in Non-Profit Organizations Through Collaborative Research and Innovation: A Systematic Literature Review. https://doi.org/10.3390/computers14120539
- KANs Layer Integration: Benchmarking Deep Learning Architectures for Tornado Prediction. https://doi.org/10.3390/bdcc9120324
- SAGE: SLO-Aware Adaptive Retrieval for Production RAG Systems: Accepted in IEEE CoDIT 2026; develops an adaptive retrieval policy for latency-cost-quality tradeoffs in production RAG. https://www.codit2026.com/