Muhammad Ahsan Shakeel

Lahore, Pakistan · ahsanshakeel13@gmail.com · +92 348 8691752 · linkedin.com/in/m-ahsan-

Education

Lahore University of Management Sciences (LUMS), M.Sc. Artificial Intelligence *Aug 2024 to Jun 2026 (Expected)*

Thesis: **Failure-Mode-Guided Directional Feedback (FMGDF) for Inference-Efficient LLM Reasoning.** *Supervisor:* Dr. Naveed-ul-Hassan

University of Management and Technology (UMT), B.Sc. Computer Science *Oct 2019 to Dec 2023*

Publications

[1] **M. A. Shakeel**, N. U. Hassan. "Directional Feedback Guided by Failure Modes for Inference-Efficient LLM Reasoning in Telecom Mathematics." *Manuscript in preparation.*

[2] Z. Hussain*, **M. A. Shakeel***, N. U. Hassan, H. Chen, C. Yuen. "Intent-Driven LLM-Based Two-Level Control in AI-Native 6G Networks." *IEEE PIMRC 2026.*

[3] **M. A. Shakeel** et al. "Noise Resilience of Quantum Support Vector Machines: A Systematic Evaluation of Feature Map Architectures for NISQ Deployment." *Submitted to IEEE Proceedings.*

Research Experience

Graduate Research Assistant, LLM Reasoning and Inference Efficiency, *LUMS EE Dept.* *Sept 2024 to Present*

- Designed FMGDF, a three-pass agentic inference pipeline (critic LLM, failure-mode taxonomy, diversity prefix) to maximise reasoning accuracy under joint call-count and token-budget constraints; packaged as a FastAPI microservice with async endpoints and OpenAPI schema for downstream integration.
- Benchmarked AWS Bedrock-hosted models (Claude 3 Sonnet, Llama 3.1-70B) alongside self-hosted OSS models as interchangeable backends; abstracted provider differences behind a unified Python interface so model swaps required only a config change.
- Tracked all experiments with MLflow (metrics, hyperparameters, model artefacts) and built an automated evaluation harness that runs ablation studies and pushes results to the tracking server, cutting manual reporting time by 70%.
- Quantified 40 to 60 point accuracy variation from prompt sensitivity; proposed a KG-based hallucination mitigation framework and documented findings as structured technical reports following MLOps best practices.

Research Contributor, LLM-Driven 6G Network Control (PIMRC), *LUMS EE Dept.* *2025*

- Co-designed an intent-driven two-level control framework for 6G; LLM upper layer refines Lyapunov parameters via structured feedback, validated across GPT-4.5, GPT-OSS-120B, and Llama-3.3-70B.
- Integrated the LLM control layer with the simulation environment via a REST microservice, enabling backbone model swaps with no changes to downstream components; wrote integration tests covering all API contracts.

Research Contributor, Quantum Machine Learning, *LUMS Physics Dept.* *2024 to 2025*

- Ran 52 experiments on four QSVM feature map architectures under three noise types; Amplitude-inspired encoding maintained 100% accuracy up to noise level 0.10, a 10x gain over the next-best approach; tracked full experiment lifecycle with version-controlled configs and MLflow.

Professional Experience

Junior Lecturer, CS and IT, *University of Lahore* *Jun 2024 to Present*

- Teach Programming Fundamentals (Python, C++) and OOP to 300+ undergraduates per semester; design assessments, supervise labs, and serve on the university conference committee.
- Designed a hands-on lab where students fine-tune a small language model, wrap it in a FastAPI endpoint, and deploy to a cloud instance, bridging theory directly to production MLOps practices.
- Introduced an applied AI ethics module covering bias auditing, fairness metrics, and responsible deployment of generative models, attended by 150+ students per cohort.

AI and Web Projects Associate, *Phoenux Design (Remote)* *Jul 2024 to Feb 2026*

- Built multi-turn agentic workflows using LangChain and AWS Bedrock Agents with tools for CRM lookup, live web search, and document Q&A via Bedrock Knowledge Bases; containerised with Docker, orchestrated with Kubernetes on AWS EC2, and load-balanced behind an ELB, reducing p95 latency by 35%.
- Invoked Claude 3 and Amazon Titan via the AWS Bedrock boto3 SDK for production inference; compared foundation models on cost, latency, and task accuracy to select the right model per client use case.
- Implemented DVC for dataset and model artefact versioning alongside MLflow for experiment tracking; every deployed model was fully reproducible from a single config file, enabling clean rollbacks during production incidents.
- Set up a GitHub Actions CI/CD pipeline to lint, test, and redeploy updated model configs and prompt templates on every merge to main, reducing manual deployment effort from hours to under 10 minutes.
- Evaluated 12 prompt and feature engineering variants through structured A/B experiments; surfaced performance degradation signals 48 hours earlier than manual monitoring, directly shaping sprint priorities in collaboration with developers and product analysts.

AI Research Intern, *Bugsfree Solutions* **Web Designer**, *Phoenux Design (Remote)* *2022 to 2024*

- Fine-tuned a PyTorch text classification model and served it via a FastAPI microservice integrated into a client SaaS platform; containerised the full ML stack with Docker and wrote OpenAPI documentation for frontend handoff.
- Applied end-to-end feature engineering pipelines (Scikit-learn and Pandas) on 500k+ row tabular datasets; improved F1 score by 18% through SMOTE rebalancing, recursive feature elimination, and cross-validated hyperparameter search.
- Performed systematic model validation including stratified k-fold CV, calibration curves, and bias audits across demographic slices; packaged findings into a responsible AI compliance report delivered to the client.
- Wrote Apache Airflow DAGs to schedule nightly data refresh and automated retraining jobs, keeping production models current with zero manual intervention; used DVC to version datasets between retraining runs.
- Delivered responsive UI design with HTML5, CSS3, Bootstrap, and Webflow (Phoenux); maintained structured technical documentation for all AI pipeline components (Bugsfree).

Selected Projects

LLM Agent Orchestration	Multi-agent workflow with LangChain tool use; agents call REST APIs for real-time data retrieval, execute code, and return structured summaries, deployed as a microservice on AWS.
Melanoma Detection	Binary classifier on HAM10000 (10k images) using Dask distributed computing; 92.81% accuracy; model exported to ONNX and served via a REST endpoint.
Bankruptcy Prediction	Compared LR, RF, XGBoost, GBM with SMOTE rebalancing on imbalanced financial data; full experiment lifecycle tracked with MLflow.
Real Estate Voice Agent	End-to-end agentic assistant with RAG over FAISS and Chroma for property search and legal Q&A; integrated with a FastAPI backend.
Big Data Pipeline	ETL on HDFS and Hive for batch processing across a multi-node Hadoop cluster; extended with an Apache Spark transformation layer for analytics.

Technical Skills

Generative AI and Agents	LLM fine-tuning, RAG, LangChain, AWS Bedrock (Claude 3, Titan, Llama 3.1), Bedrock Agents, Bedrock Knowledge Bases, tool-use agents, FAISS, Chroma
ML and AI Frameworks	PyTorch, TensorFlow, Scikit-learn, XGBoost, Qiskit; CNNs, Transformers, ONNX
MLOps and Deployment	MLflow, DVC, Apache Airflow, FastAPI, Docker, Kubernetes, CI/CD (GitHub Actions), AWS (EC2, ELB, Bedrock, SageMaker), REST microservices
Data and Distributed	Pandas, NumPy, Dask, Hadoop/HDFS, Hive, Apache Spark, feature engineering, ETL, SMOTE
Programming	Python, Java, C++, SQL, HTML5/CSS3, Dart, LaTeX
Research and Practices	Experimental design, model validation, bias auditing, responsible AI, statistical analysis, technical writing