Talha Nadeem

ML Engineer | LLMs & Generative AI | Computer Vision | Time-Series Forecasting

✉ tnadeem7860@gmail.com ☎ +923204767652 Lahore, Pakistan

linkedin.com/in/talha-nadeem-709022186 github.com/talhanadeem7860

talhanadeem7860.github.io/

PROFILE

Results-driven Machine Learning Engineer with 5+ years of experience designing machine learning systems across research and production. I specialize in large language models (LLMs), transformers, RAG pipelines, and LangChain, and I have hands-on expertise in fine-tuning, quantization, and edge deployment. Skilled in building production-ready NLP and multimodal applications (chatbots, QA systems, semantic search, synthetic data generation) and integrating automated evaluation with CI/CD. Broad background in computer vision, time-series forecasting, biomedical AI, reinforcement learning, and deep unfolding networks, with proven ability to translate advanced research into scalable real-world solutions. Strong foundation in Python, PyTorch GPU, CUDA, and MLOps, with a track record of mentoring and delivering high-impact projects at the intersection of AI research and engineering.

CORE SKILLS

Languages & Tools: Python, C++, MATLAB, Git, Docker, Conda, Jupyter, CI/CD, FastAPI, Streamlit
ML Frameworks: PyTorch GPU, TensorFlow GPU, Hugging Face Transformers, PyTorch Geometric, RLlib
LLMs & GenAI: LangChain, LlamaIndex, RAG Pipelines, Prompt Engineering, PEFT/LoRA, Quantization (INT8), CLIP, Diffusion Models
Vector Databases & Search: FAISS, Pinecone, Weaviate
ML Techniques: Transformers, GNNs, Generative Models, Reinforcement Learning, Time-Series Forecasting, Deep Unfolding Networks, Kalman Filtering
GPU & Optimization: CUDA, ONNX Runtime, Model Compression, Edge AI Deployment
Domains: Generative AI, Biomedical AI, Computer Vision, Audio Processing, Spatiotemporal Forecasting

PROFESSIONAL EXPERIENCE

09/2019 – Present
Lahore, Pakistan

Lahore University of Management Sciences (LUMS)
Machine Learning Researcher

- Designed a GPU-accelerated deep unfolding-based image restoration pipeline in PyTorch, improving PSNR by 3 dB on real-world degraded images.
- Developed a Kalman filter-based state estimation with outlier rejection, improving accuracy by 20%.
- Built an RL-based forecasting model using RLlib & NumPy, reducing false positives in preterm birth prediction by 12%.
- Implemented AI-based iterative MRI reconstruction with CUDA acceleration for large datasets.
- Supervised graduate ML labs, mentored projects in time-series, control, and biomedical AI.

06/2018 – 05/2019
Islamabad, Pakistan

National University of Sciences and Technology (NUST)
Research Assistant

- Built real-time EEG-based 3D brain activity visualization using OpenVibe & LORETA.
- Designed artifact removal & online EEG streaming for accurate source localization.

07/2018 – 08/2018 Lahore, Pakistan	**Space & Upper Atmosphere Research Commission (SUPARCO)** Intern Research intern at Power &Control Systems Lab

SELECTED PROJECTS

Production-Ready RAG System with Automated Evaluation
Built a production-ready Retrieval-Augmented Generation (RAG) pipeline with CI/CD integration. Incorporated ”LLM-as-a-Judge” for automated evaluation of factual faithfulness and relevance, ensuring continuous reliability and preventing regressions on code changes.
Skills: LangChain, Hugging Face, FAISS, CI/CD, MLOps, Python, Evaluation Automation

Chat with Your Docs – Personal Knowledge Base Q&A System
Developed a document-grounded chatbot that allows users to upload PDFs and interact with them. Powered by RAG, the system ensures responses are strictly derived from the document content, with accurate citations for traceability.
Skills: LangChain, RAG, FAISS, PDF Parsing, Streamlit, Python

LLM-Verified Synthetic Data Generation
Created privacy-preserving synthetic datasets using generative models. Integrated LLM-based verification to assess the realism and fidelity of generated data while safeguarding sensitive information.
Skills: Generative Models, Transformers, Synthetic Data, Privacy-Preserving AI, LLM Verification

Real-Time Zero-Shot Voice Cloning
Built a generative audio application capable of cloning a voice from a short reference clip and synthesizing arbitrary text in real time. Showcased zero-shot learning for text-to-speech voice synthesis.
Skills: Generative AI, Speech Synthesis, Zero-Shot TTS, Real-Time Inference, PyTorch

Transformer QA Engine
Fine-tuned a DistilBERT model on the SQuAD2 dataset to create a context-aware reading comprehension engine. Integrated Hugging Face Transformers for model loading, inference, and deployment through a simple user interface.
Skills: Hugging Face, Transformers, NLP, DistilBERT, SQuAD2, Python, Streamlit

Language Model Optimization for Edge Deployment
Applied post-training INT8 quantization to Transformer-based language models to reduce memory footprint and improve inference speed on edge devices. Benchmarked model performance pre- and post-optimization to assess trade-offs in accuracy and latency.
Skills: Transformers, Quantization, INT8 Optimization, Model Compression, ONNX, Edge AI

Semantic Image Search using CLIP
Built a semantic search engine leveraging OpenAI’s CLIP model to retrieve relevant images based on natural language queries. Aligned vision and language embeddings to enable zero-shot semantic search from a local image dataset.
Skills: CLIP, OpenAI, Vision-Language Models, Zero-Shot Learning, PyTorch, Semantic Search

Wake Word Detector
Designed a real-time audio wake-word detector using CNNs trained on MFCC features to identify the command "go" from microphone input. Deployed live inference pipeline using PyAudio and custom preprocessing.
Skills: Python, CNN, MFCC, PyAudio, Real-Time Inference, Audio Signal Processing

Molecular Property Prediction with GNNs
Developed a graph neural network (GNN) model to predict molecular properties based on structural graph data. Applied techniques such as message passing and node embedding to learn chemical representations for AI-assisted drug discovery.
Skills: PyTorch Geometric, GNNs, Molecular Graphs, Chemistry ML, Node Embedding, Regression

Retail Sales Forecasting
Built a demand forecasting model using SARIMA and k-NN regression to predict item-level sales across multiple stores. The pipeline included preprocessing of time-series data, seasonal trend analysis, and performance evaluation, aiding inventory planning decisions.

Skills: Python, Time-Series Forecasting, SARIMA, k-NN, Data Preprocessing, Pandas, Matplotlib

EDUCATION

09/2021 – Present
Lahore, Pakistan

PhD Electrical Engineering
Lahore University of Management Sciences (LUMS)

- 3.18 GPA
- Completed Coursework: Robot Motion Planning, Applied Probability, Remote Sensing of the Environment, Information Theory & ML, Smart Grid Systems
- Dissertation proposal: From Optimization to Learning: Adapting Model-Based Methods for Designing Learning Algorithms

09/2019 – 05/2021
Lahore, Pakistan

MSc Electrical Engineering
Lahore University of Management Sciences (LUMS)

- 3.63 GPA
- Completed Coursework: Stochastic Systems, Linear System Theory, Machine Learning, Advanced Digital Signal Processing, Convex Optimization, Multiagent Systems, Deep Learning, Digital Control Systems
- Thesis Title: Generalized Norm Estimator Based on Observer Principle for Robust State Estimation

09/2015 – 05/2019
Islamabad, Pakistan

BSc Electrical Engineering
National University of Sciences and Technology (NUST)

- 3.35 GPA
- Selected Coursework: Calculus, Linear Algebra & ODEs, Applied Physics, Linear Circuits Analysis, Electrical Network Analysis, Complex Variables and Transforms, Probability and Statistics, Signals and Systems, Electromagnetic Field Theory, Electrical Machines, Communication Systems, Digital Signal Processing, Microwave Engineering, Digital Image Processing
- Research Project :Real Time 3D Brain Visualization Depicting Source-localized Activity

CERTIFICATIONS

- AI for Medical Prognosis (2025) - DeepLearning.AI
- Medical Image Processing (2025) - Mathworks
- Machine Learning (2020) - Stanford University
- AI for Medical Diagnosis (2025) - Deep Learning.AI
- Introduction to Neural Networks & Pytorch (2023) - IBM
- Python for Everybody (2020) - University of Michigan
- Generative AI for Everyone (2025) - Deep Learning.AI
- Image Denoising using Autoencoders & Keras (2023) - Deprecated Guided Projects

PUBLICATIONS

- T. Nadeem, K.Ali, and M. Tahir **"NIR-EKF: Normalized Innovation Ratio based EKF for Robust State Estimation,"** *IEEE Sensors Letter*
- T. Nadeem, and M. Tahir, **"DUOV PCA: Deep Unfolded Orthogonal V ariational PCA Network for Image Denoising,"** *Under Review to IEEE Signal Processing Letters.*
- T. Nadeem, and M. Tahir, **"Multi-Degradation Image Restoration Network Based on Deep Unfolding Neural Network"** *Under Preparation for Submission to IEEE Transaction on Image Processing*
- T. Nadeem, and M. Tahir, **"Hybrid Reinforcement Guided Deep Unfolded Estimation for Preterm Birth Prediction under Influence"** *Under Preparation for Submission to IEEE Transaction on Biomedical Engineering*

LANGUAGES

English	●●●●○	Urdu	●●●●●