Shahab Zaib

Artificial Intelligence Engineer

Islamabad, Pakistan, +923259593987, shahab111zeb@gmail.com, LinkedIn, GitHub

PROFESSIONAL SUMMARY

AI Engineer with experience designing and deploying ML/DL/NLP systems from data ingestion to production. Proven expertise in fine-tuning LLMs, building RAG pipelines, and deploying GenAI models under real-world constraints. Strong background in building scalable AI tools with a focus on operational reliability and data privacy.

SKILLS SUMMARY

- **AI Foundations:** Classical ML (supervised & unsupervised), deep learning, NLP, LLMs, GANs
- **Applied GenAI Systems:** RAG architectures (FAISS, Chroma), retrieval and reranking pipelines, LLM fine-tuning (QLoRA), prompt control and grounding, privacy-first and local deployment
- **Languages:** Python, C++, SQL
- **Forecasting & Climate:** ERA5 reanalysis, NetCDF processing, shapefile-based subsetting; ClimaX, LSTM, Prophet, ARIMA

System Architecture Snapshot:

End-to-end AI systems spanning data ingestion, model training, evaluation, and deployment, including classical ML, deep learning, and LLM-based components. Designed for deterministic behavior, low-latency inference, and reliable operation under constrained compute and data-sensitivity requirements.

PROFESSIONAL EXPERIENCE

AI Engineer National Disaster Management Authority **1/2025 – Present**

- Owned and operated production AI decision-support systems, including data ingestion, fine-tuning, inference pipelines, monitoring, and failure handling.
- Built and deployed a full LLM-driven advisory pipeline by fine-tuning LLaMA-2 (13B) on localized datasets, resulting in an 80% reduction in manual alert generation time and improved advisory relevance over baseline models.
- Built a BERT-based emergency detection system for social-media monitoring with ~15-minute response latency.
- Evaluated hybrid forecasting models (FB-Prophet + LSTM, MSE ≈ 1.5) and identified ClimaX spatial limitations for regional use.
- Coordinated with operational stakeholders to align model outputs with decision-making workflows.

Research Assistant: Hong Kong Polytechnic University **7/2023 – 02/2024**

- Conducted 80+ design experiments involving task-based product creation, quantitative questionnaires, and audio interviews to explore students' cognitive design abilities.
- Transcribed and analyzed interview data using early LLMs for thematic extraction and cross-validated results via students' annotated design instructions.
- Converted questionnaire data to structured CSV format and performed statistical analysis, including ANOVA, MANOVA, and regression, to identify key behavioral patterns.
- Created visualizations and extracted insights that directly contributed to two published papers on cognitive design and user behavior (co-author on one).

Research Assistant: IQRA National University **6/2021 – 6/2023**

- Optimized a 45K+ image dataset using perceptual hashing to remove redundant and noisy samples, resulting in a clean 27K-image dataset for plant disease classification.
- Trained and evaluated multiple CNN architectures (ResNet variants, DenseNet, VGG-16, and custom CNN) to detect plant diseases with high accuracy.
- Integrated OpenCV-based vision system with drone feed to enable real-time aerial disease detection via onboard inference.
- Developed a logistic regression-based recommendation model to prescribe targeted pesticide treatments alongside disease identification.

Freelancing **1/2023 – 3/2023**

- Delivered ML solutions for international clients, improving operational outcomes by 10–20%.
- Deployed CNN- and ANN-based models and used NLP for classification and sentiment analysis in production settings.

PROJECTS

Local RAG Platform — Privacy-First AI Knowledge System

- Designed and deployed a fully local, domain-agnostic RAG system for privacy-sensitive, compute-constrained environments used by a federal defense agency.
- Built the end-to-end retrieval pipeline, including document ingestion, vector indexing, retrieval, reranking, and local LLM inference.
- Implemented deterministic grounding using FAISS, custom reranking logic, and GGUF-based LLMs.
- Deployed as a GPU-backed, multi-user platform supporting ~100 concurrent users with low-latency inference and strict data locality enforcement.

LLM Fine-Tuning Pipeline for Government Policy Documents

- Built an automated pipeline to generate high-quality instruction–response datasets from mixed-format government policy documents.
- Supported ingestion from PDF, DOCX, TXT, MD, and HTML, with paragraph-aware chunking and overlap control.
- Generated QA pairs using OpenAI and Groq models, scored for structure, clarity, and information density.
- Applied multi-stage filtering: evidence grounding, truncation checks, heuristic validation, and LLM-based judging.
- Final datasets split into train.jsonl and dev.jsonl with leakage prevention; fine-tuning phase underway.

AI-Powered Clinic Automation Platform *(Freelance – US Client, In Progress)*

- Developed a full-stack LLM-powered system to automate patient intake, symptom collection, and doctor interaction.
- Built a dynamic form-filling interface where an LLM conducts patient interviews and auto-fills structured medical forms.
- Enabled real-time doctor review, symptom tracking, and LLM-assisted appointment scheduling.
- Integrated Django-based session management with two-way sync between patient and doctor portals.
- Currently expanding to include automated lab handoffs, billing, and HIPAA-compliant data handling.

Adaptive Learning Rate Algorithm for GAN – MS Research (Novel Research):

- Developed an adaptive learning-rate mechanism to stabilize training dynamics between discriminator and generator.
- Integrated and tested the method across multiple GAN architectures: DCGAN, WGAN, WGAN-GP, SNGAN, BigGAN, and StyleGAN.
- Improved training stability and output quality, validated through loss-gap analysis, Fréchet Inception Distance (FID), and Inception Score (IS).

Regional Temperature Forecasting Using ERA5 Reanalysis Data (Spatiotemporal AI)

- Built a forecasting pipeline using ERA5 hourly reanalysis data and station-level time-series modeling.
- Implemented shapefile-based spatial subsetting, year-wise NetCDF preprocessing, and memory-optimized data merging.
- Trained and evaluated localized models (LSTM, Prophet, ARIMA) with RMSE, MAE, and R^2 against ERA5 observations.
- Enabled regional scalability by decoupling spatial boundaries from model generation and training logic.

CERTIFICATIONS

Coursera – DeepLearning.AI

- Neural Networks & Deep Learning
- Supervised Learning: Regression & Classification
- Unsupervised Learning, Recommenders, Reinforcement Learning

EDUCATION

Malakand University **10/2024 - Present**
MPhil Computer Science (Generative AI)

IQRA National University **10/2018 – 9/2022**
BS Computer Science CGPA 3.45/4

RESEARCH PUBLICATIONS

1. **Enhanced deep learning architecture for rapid and accurate tomato plant disease diagnosis:**
   Google Scholar Link: *Published in AgriEngineering*, 2024

2. **Exploring designers' cognitive abilities in the concept product design phase through traditional and digitally-mediated design environments:**
   Google Scholar Link: *Published in Proceedings of the Design Society*, 2024

ACHIEVEMENTS

- **Research Grant Awardee**: $15K funding from University of Parthenope, Italy
- **Paper Accepted:** Proceedings of the Design Society, 2024. AgriEngineering
- **Bs Research Topper:** High distinction ranked in Final Year Thesis (BS)
- **Kaggle Competitions**: Top 500 (Titanic), Top 10% (NLP Disaster Tweets)