MUHAMMAD HAMZA GHAFRAN

AI Engineer
+92 3346138410 • hamzaghafran56@gmail.com Islamabad

Summary

Passionate AI and Machine Learning Engineer with hands-on experience in building intelligent systems, including agentic chatbots, monitoring pipelines, and ML deployments. Strong foundation in programming, data modeling, and algorithms, with a focus on applying advanced AI methodologies to solve real-world problems and drive innovation.

Experience

AI Mark Labs
Islamabad·Onsite

AI Engineer 01/2025 - Present

- Built and maintained MCP server build tools such as Citation Evaluation Tool and Analyze Tool, optimizing data processing and performance.
- Integrated Datadog for logging, monitoring, and performance insights across agents and internal tools, improving system observability and reliability.
- Configured Prometheus to track and visualize HTTP metrics, ensuring real-time performance monitoring and alerting.
- Developed an Agentic Chatbot using LangGraph, enabling multi-agent orchestration for intelligent and context-aware interactions.
- Implemented Arize Phoenix for monitoring, tracing, and logging of chatbot and ML pipeline performance, with self-deployment and integration on AWS infrastructure.
- Deployed and managed chatbot and monitoring pipelines on AWS, ensuring scalability, availability, and secure model serving

Riayatech
Rawalpindi·Onsite

Junior Machine Learning Engineer 07/2024 -12/2024

- Developed Time Series Model for Medicine Point-of-Sale forecasting.
- Evaluate and deployment of Large Language Model. Used Model for specific language transcription.
- Designed and implemented end-to-end ML pipelines using MLflow, Including model tracking, experimentation, and deployment.
- Set up MLflow server for efficient model management and performance monitoring in production environments.

Riayatech
Rawalpindi Onsite

Intern Machine Learning Engineer 05/2024 -07/2024

- Preprocessed data, engineered features, and analyzed trends through visualizations.
- Performed in-depth data analysis to identify patterns, trends, and relationships within the data.
- Created and selected relevant features to improve model performance.
- Built and refined time series models (ARIMA, SARIMA) and evaluated performance using MAE/RMSE.
- Visualized data and model results using tools like Matplotlib and Seaborn.

Education

Bahria University Islamabad, Pakistan
09/2020 - 06/2024

Bachelor of Science - BS, Computer Science

Projects

Medicine Point-of-Sale Forecasting — StatsModels
Developed ML system using Time Series Model for better medicine point-of-sales forecasting

- Designed and implemented time series technique.
- Data preprocessing, visualization, and data decomposition,
- Forecaster's toolbox (simple methods, Transformation and adjustment, Residual Interval, Prediction interval).
- Implemented time Series decomposition (Classical, X11, SEAT, STL decomposition).
- Implemented forecasting models (Regression model, Exponential smoothing, ARIMA/SARIMA).
- Model training & Evaluated model performance using metrics such as MAE, RMSE.
- Conducted cross-validation to ensure model robustness.
- Performed model benchmarking to compare different models and select the best-performing one.

Smart Room Designer (Stable Diffusion) — Large Language Model
Developed a web app using machine learning models and computer vision techniques for redesigning rooms.

- Utilized computer vision technique for edge detection.
- Utilized MSLD model and Stable Diffusion model pipeline for generating high-quality room redesigns.
- Fine-tuned the encoder using a line segmented dataset to enhance model accuracy and performance.
- Implemented a user-friendly front end using HTML, CSS, and JavaScript for an interactive user experience.
- Integrated the front end with Flask to create a seamless and responsive web application.

SmartDoc A — Retrieval-Augmented Generation Architecture
Developed a Information retrieval system using natural language techniques and for data retrieval for multiple schools.

- Designed and implemented a multi-tenant RAG system allowing each school to upload, retrieve, and manage documents (policies, events, admissions, classes).
- Used FastAPI to build scalable backend APIs for document handling, deletion, and retrieval.
- Leveraged OpenAI embeddings and Pinecone vector DB to enable semantic search across school-specific documents.
- Integrated LangChain with OpenAI models to generate context-aware responses tailored to each school's data.

Sales Bot – Retrieval-Augmented Generation Architecture
Worked on Moses Sales Bot with help of RAG to response client queries .

- Developed a Salesbot to handle sales queries using RAG, powered by LLaMA 3.2 and integrated via Groq API for low-latency inference.
- Generated domain-specific embeddings using Nomic and managed them with ChromaDB for efficient retrieval.
- Built a modular RAG pipeline using LangChain, streamlining embedding retrieval and dynamic response generation.
- Designed an interactive Streamlit UI for end-user engagement and conducted iterative testing to optimize accuracy and reliability.

Evaluation and Deployment Speech to Text Model (Whisper Large_v3) — LLM
Worked on Large Language Speech to Text Model (Whisper Large v3)

- Worked a multilingual speech-to-text model using Whisper large v3 to address transcription needs in English, Arabic, and Urdu.
- Collected and preprocessed audio data across the three languages for model evaluation.

- Evaluated model performance using Word Error Rate (WER) and Character Error Rate (CER) metrics across the three languages.
- Containerized the model using Docker for scalable deployment and smooth inference in production environments.

Skills

Language/Frameworks: Python, C++, Pandas, Stats Models, Scikit-learn, OpenCV, Matplotlib, Tensorflow, keras, Transformers, Hugging-face, Ollama Server, Langchain, OpenAI, Chromadb, Pinecone, Async

Machine Learning & AI: Classification, Regression, Clustering, Deep Learning, Neural Network, Convolutional Neural Network, Time Series Modeling, & Transfer learning, Vector Embeddings (OpenAI, Transformer, SBERT, Nomic), Image Generation(Stability AI, Controlnet),

Natural Language Processing and LLM: Natural Language Processing, Retrieval-Augmented Generation, Large Language Model (LLM), Finetuning Quantization, LORA, QLORA. Knowledge Graph

Tools/Technologies: Git, Github Actions, Docker, Jenkins, Continuous Integration and Continuous Delivery (CI/CD), Jupyter Notebook, Linux, Mlflow[piplines], Chatlint, Gradio, Streamlit, GCP, Azure

Database & Data Management: Postgres, Mongodb, SQLite, Vector Database (FAISS, Chroma, Pinecone, Qdrant, pgvector), Alambic, Redis

API Development: Fast Api, Flask Ap

Observability & Monitoring: Arize Phoenix, Langsmith, Langfuse, Datadog, Promethues, Kibana

Programming Knowledge: Data Structure & Algorithms (DSA), Object Oriented Programming.

Cloud Services: Aws(EC2, ECR, Sagemaker, S3, Lamda, Dynmodb, Cloud Watch, Secrets Manager)