Muhammad Haroon

ranaharoong14@gmail.com — (312) 409 3711 — Linkedin — Github

Experience

Facility for Rare Isotope Beams, MSU *East Lansing, MI*
Machine Learning Engineer *Sep 2023 – Aug 2024*

- Reduced cavity maintenance cost by **30%** at a $700M nuclear facility by developing and deploying a deep learning model for automated surface defect classification.
- Increased defect classification performance to **F1-score of 0.68** by implementing a model based on deep convolutional neural network and `transformer encoders`, enhancing critical defect detection.
- Established comprehensive data validation systems to ensure training dataset integrity, implementing automated checks for image quality, proper labeling, and compliance with facility data standards.

Ledavio *Remote*
Machine Learning Engineer *Dec 2021 – Nov 2022*

- Developed a natural language-based image search tool using BERT model, empowering designers to efficiently find sentiment-aligned images for mood board creation—boosting workflow efficiency by **25%**.
- Deployed a scalable data ingestion, cleaning and AI model inference pipeline on `AWS Cloud`, ensuring seamless image search performance.

Upwork *Remote*
Machine Learning Engineer *Jun 2020 – Dec 2021*

- Fine-tuned multiple object-tracking models for Tennis, achieving 0.51 mean average precision using CNNs on video data for player, ball, court and scorebox detection.
- Created and deployed a Flask API that uses PySpark RDD operations for scalable, distributed data processing across a multi-node cluster.
- Built automated ETL pipelines to ingest, preprocess, and partition large volumes of tennis video data, implementing frame extraction, resolution standardization, and metadata tagging for model training datasets.

Noerric Technologies *Islamabad, PK*
Co-founder and Machine Learning Engineer *Sep 2017 – Mar 2020*

- Developed an AI system for No-ball detection in Cricket, achieving **95% accuracy** for MVP.
- Designed ETL workflows to process and transform massive point cloud datasets from depth sensors and stereo cameras, implementing data compression, noise filtering, and coordinate system transformations for precise spatial analysis.
- Won a **$120K grant** from IBM GEP for innovative sports AI research.

Skills

Programming Languages	`Python (Primary)`, `R`, `MySQL`, `C++`
Generative AI & LLMs	`LLM Fine-tuning & Deployment (BERT, GPTs, Llama, Stable Diffusion)`, `Retrieval Augmented Generation (RAG)`, `Prompt engineering`
Machine Learning	`HuggingFace`, `TensorFlow`, `PyTorch`, `XGBoost`, `Pandas`, `NumPy`, `Scikit-learn`, `AWS (SageMaker, Glue)`, `NLTK`, `Spacy`, `OpenCV`
MLOps	`Docker`, `Git`, `CI/CD (GitHub Actions)`, `PySpark`, `GCP`, `Linux`, `TensorBoard`, `Matplotlib`, `MLflow`, `Pytest`, `Flask`, `Weights & Biases`
Agentic AI	`AWS (Bedrock)`, `LangChain`, `CrewAI`, `Pinecone`

Education

Michigan State University *East Lansing, MI*
Master of Data Science (GPA:4) *Aug 2023 – April 2025*
Relevant Coursework: Deep Learning, NLP, Foundations in AI, Big Data Analysis.

National University of Sciences & Technology *Islamabad, PK*
Bachelor of Electrical Engineering *Sep 2015 – Jun 2019*

Projects

GenAI - Automated Code Documentation *Aug 2024 – Dec 2024*

- Developed an automated code documentation solution using the `CodeT5 LLM` to enhance code readability, maintainability, and collaboration.
- Fine-tuned the model on the Google MBPP dataset to achieve high performance **(BLEU score of 0.46, ROUGE-L of 0.70)** and optimized runtime, processing 14.76 samples per second.
- Performed literature review of code-to-text models and datasets to select optimal model and establish benchmarks.

Enhanced Text Summarization using LLMs *Aug 2024 – Dec 2024*

- Fine-tuned T5 LLM model on the WikiHow dataset, improving text summarization performance across diverse topics.

– Engineered an efficient preprocessing pipeline, optimizing tokenization and sequence length for improved data representation and model performance.
– Delivered actionable insights with **ROUGE-1: 0.22 and METEOR: 0.20**, showcasing the model's performance for real-world applications.

Sentiment Classification of Amazon Reviews with LLMs *Aug 2023 – Dec 2023*

– Created a customer feedback analysis tool by transfer learning of transformer-based LLMs (`BERT` and `GPT-2`) for sentiment classification.
– Achieved a high-performance **F1-score of 0.85** with OpenAI GPT-2 and 0.80 with BERT, enabling actionable insights from customer reviews.

GenAI Board Game Design Tool *Jan 2023 – Mar 2023*

– Developed a prototype tool for generating board game designs by fine-tuning the `Stable Diffusion` model.
– Created an automated data collection tool which collects pertinent game board design images along with textual description from web.

GenAI - Automated Mobile UI Design *Jun 2020 – Feb 2021*

– Fine-tuned generative models for mobile UI designs using the RICO dataset, leveraging convolutional graph neural networks and self-attention-based graph transformers.
– Created unidirectional graph representations for UI designs. Explored various generative modelling approaches and data formats, including bidirectional graphs, adjacency matrices, and segmentation masks, to enhance representational learning.