

Abdul Karim

abdulkarim25446@gmail.com | +923141870320 | Islamabad, Pakistan | [LinkedIn](#) | [GitHub](#)

WORK EXPERIENCE

AIREC - Machine Learning Engineer (LLMs / NLP)

June 2023 - PRESENT

- Designed and deployed domain-specific conversational AI systems by building high-quality knowledge bases from multimodal data sources.
- Engineered scalable data pipelines for automated data ingestion, transcription, and preprocessing, improving data collection efficiency.
- Transformed unstructured data into instruction-tuning datasets (JSON/QA pairs), enhancing model alignment and downstream performance.
- Fine-tuned and evaluated multiple LLMs, including LLaMA 3 (7B, 70B), Qwen 3 (30B), and Mixtral (70B), improving response quality across tasks.
- Built and optimized Retrieval-Augmented Generation (RAG) systems, improving factual accuracy and reducing hallucinations.
- Implemented advanced prompt engineering techniques, increasing response relevance without retraining.
- Conducted cost-performance optimization, transitioning from fine-tuning to inference-time optimization (DSPy), reducing operational costs.
- Developed modular LLM pipelines using DSPy, enabling structured, self-improving prompt workflows and rapid experimentation.
- Designed and deployed agentic LLM systems using Microsoft Azure AI Foundry, enabling dynamic multi-turn interactions, tool integration, and adaptive reasoning beyond static pipelines.
- Built backend systems using Python, FastAPI/REST APIs, integrating end-to-end flows from data ingestion → retrieval → LLM inference → user response.
- Migrated infrastructure from Firebase to MongoDB/PostgreSQL, improving scalability, query performance, and production readiness (~30% performance gain).
- Collaborated across frontend and backend to deliver real-time, production-grade AI applications, improving user experience and task completion rates.

Datalytics AI - Machine Learning Engineer

Jan 2023 - May 2023

- Designed and implemented an inference pipeline for a content-based recommendation system, to suggest personalized posts to users based on their interests and preferences. The pipeline involved loading the trained model, preprocessing the input data, and executing the model to generate predictions for new user data.
- Proficient in Python, Pandas, NumPy, and Scikit-learn. Strong problem-solving and communication skills, with the ability to work effectively in a team.

CERTIFICATIONS, SKILLS & INTERESTS

- **Certifications:** Data Science & Machine Learning (Dice Analytics); Generative AI with Large Language Models (Coursera); Data Analytics (Brainnest).
- **Skills:** Microsoft Foundry (Azure AI Foundry), LLM Applications, Agentic AI, Retrieval-Augmented Generation (RAG), Prompt Engineering, Tool Calling / Function Calling, Fine-tuning, Natural Language Processing (NLP), LangChain, LangGraph, Transformers, Hugging Face, PyTorch, TensorFlow, Model Evaluation, Embeddings, Vector Databases, Semantic Search, LLM Deployment & Inference, FastAPI, Python, Machine Learning, Deep Learning, NoSQL, SQL, Amazon SageMaker AI, AWS, Paperspace, Runpod.
- **Projects:** FLAN-T5: PEFT & PPO for Non-Toxic Summaries; Retrieval-Augmented Generation (RAG) Chatbot using OpenAI API; EEG-Based Emotion Detection using Deep Learning (Python); QLoRA-Enhanced LLaMA 2 7B: Chatbot; Multi-Agent User Simulation using AutoGen with Custom LLMs; n8n Workflow for RAG Pipeline: Google Drive Ingestion, Chunking, and Vector Database Storage.

EDUCATION

Iqra University - Bachelor of Science in Computer Science BS(CS)

May, 2021