Muhammad Mursaleen

AI & ML Engineer

+92 307 770 6100 | mursaleen.sengr@gmail.com | LinkedIn | Github | Islamabad, Pakistan

Professional Summary

AI & ML Engineer with 1.5 years of experience building production-grade Generative AI and Agentic systems. Specialized in architecting scalable RAG pipelines and fine-tuning LLMs for niche domains, with expertise in bridging the gap between ML models and real-time users by deploying FastAPI backends on AWS and DigitalOcean using event-driven architectures.

Experience

AI/ML Engineer — **March 2025 – Present**

DiveDeepAI — Islamabad, Pakistan

- Designed and implemented scalable AI & backend pipelines serving real-time users, deployed on AWS and DigitalOcean with high availability and fault tolerance
- Collaborated with cross-functional teams including product managers, designers, and DevOps engineers to deliver innovative AI solutions on schedule
- Successfully led project initiatives from conception to deployment, ensuring alignment with business objectives and technical best practices

Jr AI Engineer — **August 2024 – February 2025**

Softoo — Islamabad, Pakistan

- Developed RAG-based applications using LangChain, LLMs (GPT, Claude, LLama), and vector databases (Qdrant, FAISS) for context-aware information retrieval systems
- Built Chat-with-Database proof-of-concepts using embedding models (OpenAI, Sentence-Transformers) to enhance data-driven decision-making for enterprise clients
- Implemented semantic search and question-answering systems that improved information retrieval accuracy by 35% compared to traditional keyword-based approaches
- Optimized vector database queries and embedding pipelines, reducing latency by 40% and improving user experience in production environments

Projects

eQRA-AI: Islamic AI Assistant Mobile Application — 2024

- Architected and deployed complete FastAPI backend for Flutter-based Islamic mobile app, serving dual RAG-powered chatbots: Fatwa Research Assistant and Situational Duas Recommendation system using Qdrant vector database
- Engineered real-time Voice Tasbeeh feature integrating Voice Activity Detection (VAD), WebSocket streaming, and Whisper speech-to-text for hands-free Islamic prayer counting and transcription
- Built production infrastructure on DigitalOcean with Qdrant Cloud for vector search, migrated from AWS EC2/S3 architecture while maintaining AWS SES for email delivery and 99.9% uptime
- Designed event-driven push notification pipeline using AWS EventBridge, Lambda, and SNS integrated with Firebase Cloud Messaging (FCM) for real-time mobile alerts
- Implemented asynchronous file upload processing using Redis and Celery task queue for handling document ingestion and heavy computational workloads at scale
- **Tech Stack:** Python, FastAPI, LangChain, Qdrant, Redis, Celery, WebSocket, Whisper, AWS (SES, Lambda, SNS, EventBridge), Firebase, Docker, DigitalOcean

Fine-Tuning Whisper for Quranic Arabic Transcription — 2024

- Fine-tuned OpenAI's Whisper model on custom Quranic Tarteel dataset (10,000+ audio samples) to improve transcription accuracy for Classical Arabic
- Enhanced Harakat (diacritical marks) recognition by 90%, enabling precise pronunciation feedback for Quran reciters and learners

- Leveraged PyTorch and Hugging Face Transformers for distributed training, optimizing model to detect subtle phonetic variations in Quranic recitation
- Achieved Word Error Rate (WER) reduction from 30% to 40% on test set, significantly outperforming baseline Whisper model
- **Tech Stack:** Python, PyTorch, Hugging Face, Whisper

AI-Powered Database Chat & Visualization Agent 2024

- Built intelligent chatbot enabling natural language queries on relational databases, democratizing data access for non-technical stakeholders
- Trained Vanna.ai model on PostgreSQL schemas to convert natural language into optimized SQL queries with 92% accuracy
- Integrated Autogen multi-agent framework for automated CSV analysis and exploratory data analysis (EDA) visualization through conversational prompts
- Automated analytics workflow reducing time-to-insight by 60%, enabling business users to generate reports without SQL knowledge
- **Tech Stack:** Python, Vanna.ai, Autogen, PostgreSQL, Pandas, Matplotlib, Streamlit

Skills

Programming Languages: Python, C++, SQL, Julia

AI & ML Frameworks: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, OpenAI API, LangChain, LangGraph, Autogen, Haystack

Data Science Libraries: NumPy, Pandas, Matplotlib, Seaborn, NLTK, SpaCy

Databases: PostgreSQL, SQL Server, MySQL, Vector Databases (Qdrant, FAISS, PgVector), MongoDB

Cloud & DevOps: AWS (EC2, S3, Lambda, EventBridge, SNS), DigitalOcean, Docker, Linux, CI/CD

Backend & APIs: FastAPI, Flask, Django, REST APIs, Redis, Celery, WebSockets

Tools & Frameworks: Git, GitHub, GitLab, Streamlit, Postman, Jupyter, VS Code, Prompt Engineering

Education

Bachelor of Science in Software Engineering; CGPA: 3.70/4.00 **Oct. 2020 – June 2024**

The Islamia University of Bahawalpur, Punjab, Pakistan

- **Relevant Coursework:** Databases, Object-Oriented Programming, Software Engineering, Data Structures and Algorithms, Artificial Intelligence, Machine Learning, Cloud Computing, Computer Networks
- **Academic Achievements:** Maintained consistently high academic performance throughout degree program

Certificates & Learning

Intermediate Python DataCamp, 2024

- Covered advanced Python techniques, data manipulation with Pandas, and visualization with Matplotlib

Introduction to Large Language Models Google Cloud, 2024

- Mastered LLM fundamentals, model training, fine-tuning, prompt engineering, and transformer architectures

Leadership & Extracurricular

Volunteer Tech Destination of Pakistan Digital Conference

- Organized and hosted training sessions for 200+ students, focusing on emerging AI/ML technologies and practical implementations
- Collaborated with industry experts to create interactive workshops, improving attendees' hands-on understanding of GenAI and LLMs