



MOAZ MUMTAZ

AI ENGINEER

moazmumtaz998@gmail.com | [03364515998](tel:03364515998)

Delivering production-ready AI and full-stack solutions!

PROFILE

Passionate AI Engineer with a deep-rooted love for technology, building real-world applications across web development and artificial intelligence. I've delivered impactful, production-grade solutions including real-time voice assistants, RAG-based systems, AI medical platforms, and full-stack SaaS products. My expertise spans Python, FastAPI, React, Next.js, and cloud-native architectures on AWS and Azure, with a strong focus on scalable system design. I bring strong problem-solving skills, a commitment to best engineering practices, and a drive to deliver reliable, high-quality, user-friendly applications.

LINKS

[LinkedIn](#), [Portfolio](#)

PROFESSIONAL EXPERIENCE

Lead AI Engineer - Full Time, Dev Entities

Jun 2025 — Present

Lahore - Onsite

- Built and currently scaling TwelveAM, a real-time, human-like voice assistant startup that handles live customer interactions (support, reception, FAQs) using streaming STT/TTS and LLMs, with a focus on low-latency responses and natural conversation flow.
- Built Byrddog, a B2B SaaS platform that automates electrical RFQ emailing, comparison, and tracking, reducing manual supplier communication and follow-ups for procurement teams.
- Designed the back-end architecture for customizable AI voice agents, allowing businesses to configure prompts, workflows, and responses based on their domain and use case.
- Leading development of AI solutions, including real-time voice assistants, RAG systems, and LLM-integrated apps using Python, FastAPI, and LangChain
- Architecting and deploying scalable data pipelines and AI infrastructure using Docker on Linux servers.
- Building and maintaining full-stack applications with React and Next.js integrated with AI backends.
- Overseeing end-to-end AI project delivery while actively contributing to code, architecture, and system design.
- Mentoring developers, conducting code reviews, and aligning technical strategy with product goals.
- Implementing custom web scraping tools for data collection, model training, and retrieval systems.

AI Engineer/Full Stack - Part Time, Dev Entities

Jun 2024 — Mar 2025

Lahore - Remote

- Developed advanced AI applications, including chatbots, real-time voice assistants, predictive systems, and RAG-based solutions, using Python, FastAPI, and LLMs to optimize performance and user interaction.
- Created web scrapers tailored for specific needs, including RAG-compatible data extraction, to support AI model training and enhance system accuracy.
- Managed Linux server environments with Docker, ensuring scalability and reliability for AI and scraping workloads.

- Designed and integrated responsive web applications using React and Next.js, connecting AI backends and RAG systems for seamless, high-quality user experiences.

AI Engineer - Full Time, Megasight.ai

Aug 2024 — Jul 2025

Lahore - Onsite

- Developed multiple LLM-powered applications across healthcare, legal, and energy domains, working hands-on with LLMs, RAG systems, computer vision, and big data pipelines.
- Built Times by DRH, an end-to-end medical transcription and summarization system that converts doctor–patient conversations into structured clinical notes using local Whisper STT, OpenAI LLMs, and AWS services within a Dockerized architecture.
- Designed and deployed Care Platform (Time+), a remote healthcare solution featuring an AI medical assistant that interacts with patients, captures health information, and generates structured summaries for doctors using OpenAI LLMs, Groq STT, OpenAI TTS, and cloud-native services.
- Implemented a RAG-based legal assistant (Jurisight) using LangChain and Qdrant, enabling semantic search and question answering over patent law documents parsed from PDFs.
- Performed data engineering and AI automation for renewable energy datasets (solar and wind) using Azure Databricks, transforming raw IoT data into analytics-ready pipelines to support energy insights and reporting.
- Developed and trained computer vision models for liveness detection and biometric verification, including spoof detection, 1:1 face matching, and fingerprint feature extraction using OpenCV, YOLO, ResNet-50, FaceNet, VGGFace, TensorFlow, and PyTorch.
- Owned the full ML lifecycle, including data collection, preprocessing, model training, evaluation, deployment, and iterative improvement in production environments.

Trainee Data Scientist - Part Time , Mikrostar Tech

Apr 2023 — Jun 2024

Lahore

- Worked as a Data Scientist on AI solutions involving IoT sensor data and computer vision, contributing to model development, experimentation, and deployment.
- Developed a posture classification model to detect sitting vs. standing using time-series data from accelerometer and pressure sensors embedded in smart socks.

Assistant lecturer, Riphah International University

Oct 2023 — Jun 2024

Lahore

- Taught Programming Fundamentals and DSA to 3 Batches.
- Organized and supervised lab sessions, ensuring a productive learning environment.
- Conducted tutorial sessions to reinforce key concepts and facilitate student learning.
- Emphasized the development of analytical skills in students through practical exercises and projects

EDUCATION

Riphah International University, BSCS

Oct 2020 — Jun 2024

Lahore

Address: 13 km Raiwind Rd, Chamru Pur, 54000 Lahore (Pakistan) | Website: riphah.edu.pk/home-lahore

FEATURED PROJECTS

TwelveAM - ReactJS, NodeJS, Python, PostgreSQL, AWS, Groq ,Dockers

- Building a startup focused on a real-time, human-like AI voice assistant for business use cases.
- Developing natural, low-latency conversational experiences to handle customer queries effectively.
- Enabling businesses to create and customize AI assistants for roles like customer support and reception.
- Designing a scalable platform adaptable to multiple industries and operational needs.

Link: <https://twelveam.io>

Byrddog Solutions - ReactJS, Node Express, Tailwind CSS, Python, PostgreSQL, Redis ,Cloudfare R2

- Developed a B2B SaaS platform to automate the emailing workflow for electrical RFQs.
- Enabled users to send, track, and compare multiple RFQs across suppliers in a centralized system.
- Built a responsive and intuitive frontend using React.js.
- Implemented scalable backend services using Python and Node.js.
- Optimized system performance and reliability to support business-scale operations.

Link: <https://www.byrddogsolutions.com>

Times By DRH - Python, Local Whisper, OpenAI, Docker, AWS (EC2, Lambda, S3, Cloudwatch), PosgreSQL

- Deployed an end-to-end machine learning pipeline to generate structured medical notes from doctor–patient conversations.
- Integrated local Whisper for speech-to-text and OpenAI LLMs for medical note generation.
- Designed a scalable, event-driven architecture using AWS services (EC2, S3, Lambda, SQS, CloudWatch).
- Containerized the entire system using Docker for consistent deployment and portability.
- Ensured reliability, performance, and scalability for real-world clinical workflows.

Link: <https://time.drh.ai>

TECHNICAL SKILLS

AI & Machine Learning	React 18	Tailwind CSS	Azure ML, Azure Databricks
Langchain	LLM & Agent Systems	Data & MLOps	Retrieval Augmented
Responsive & Accessible UI	Generative AI	Next.js	Generation(RAG)
Docker	Conversational AI Agents	LLM Evaluation, Prompt	WebSockets
Authentication &	Async & Concurrent Backends	Versioning & Cost-Aware	Systems & Architecture
Authorization	Applied Machine Learning	Inference	TypeScript
PostgreSQL	JWT & OAuth2	Python	RAG & Knowledge Systems
Express.js	LLM's & SLM's	ReactJs, NextJs & NodeJS	Node.js
AWS (SageMaker, EC2, S3,	Database Schema Design	FastAPI	Speech & Multimodal AI
Lambda, SQS, CloudWatch)		Python, FastAPI, & Django	
Webhooks			

ACCOMPLISHMENTS

- [25/04/2024] Riphah International University - **Riphah Acknowledgement Award 2024**
- [08/07/2025] Riphah International University - **Riphah Acknowledgement Award 2025**