

ABDUL MOIZ

AI/ML Engineer

moiz44386@gmail.com | [GitHub](#) | [LinkedIn](#) | [Portfolio](#) | [Medium](#)

AI/ML Engineer who builds, trains, and integrates machine learning and AI systems into production applications end to end, from data cleaning and model training to REST API integration and deployment. Comfortable across classical ML, deep learning (PyTorch), and LLM based systems (RAG, embeddings, vector databases), with hands on experience shipping ML features into real products and working independently.

TECHNICAL SKILLS

Languages & Core: Python, SQL, REST APIs, Git/GitHub, Testing & Debugging

ML Fundamentals & Data: Data Cleaning & Preprocessing, Feature Engineering, Exploratory Data Analysis (EDA), Statistical Analysis, Model Training & Evaluation, A/B Testing, model evaluation, fine-tuning

Deep Learning & Computer Vision: PyTorch, TensorFlow, CNNs, U-Net, YOLO, OpenCV

NLP & LLMs: LangChain, LangGraph, Hugging Face Transformers, OpenAI API, Anthropic Claude, RAG, Embeddings, Prompt Engineering

Data & Vector Stores: PostgreSQL, MongoDB, Pinecone, FAISS, Weaviate, Chroma, Firebase/Firestore

Cloud & MLOps: Docker, Kubernetes, CI/CD, MLflow, AWS, GCP, Azure

Full-Stack & API Integration: FastAPI, Flask, Node.js, React, Next.js

PROFESSIONAL EXPERIENCE

Kinetiq · Sep 2025 – Oct 2026 · Remote

AI/ML Team Lead · Jul 2026 – Oct 2026

- Led a remote, cross-functional team of 5 AI/ML engineers, setting GenAI/ML direction and mentoring on model design and deployment, shipping 4 production ML/LLM features in 4 months.

AI/ML Engineer · Sep 2025 – Jun 2026

- Built ML and LLM models end to end, from data cleaning and feature engineering to FastAPI endpoints on SQL and vector databases, improving model accuracy by 18% across 4 production use cases.
- Implemented post-deployment monitoring of accuracy, latency and output quality, iterating on data and models to raise accuracy by 10% and cut response latency by 32%.

Artificial Intelligence Intern | Digifloat · Apr 2026 – Jul 2026 · Hybrid

- Developed real-time object detection and tracking with YOLO, ROS2 and Gazebo, taking image data from preprocessing to deployed models running at 28 FPS with 0.86 $mAP@0.5$.

AI Researcher | Prosilient Systems Inc. · Jul 2025 – Sep 2025 · Remote

- Engineered automated web-scraping and data pipelines that cleaned and structured 120K unstructured records, cutting duplicates and noise by 30% for downstream model training.

AI Developer | Center for Advanced Research in Engineering · Jul 2024 – Sep 2024 · On-site

- Built RAG and NLP document-intelligence systems with LangChain, LangGraph and vector databases, served through FastAPI/Flask to a React front end, answering queries over 2,000 documents at 87% accuracy.
-

PROJECTS

StockPredictor – Daily Equity Forecasting | LightGBM, FinBERT, FastAPI, PostgreSQL, GitHub Actions

- Built a daily forecasting pipeline on 10 years of free market data and tested 24 models with purged walk-forward evaluation; the live move-size model beat the HAR baseline in 6/6 years at 80.6% band coverage.

HandoffCheck – Product-to-Engineering Handoff Workspace | React, Supabase, PostgreSQL, Gemini, App Script, Tool Calling

- Built a requirements workspace where AI review flags gaps with source-quoted evidence and two leads approve one exact revision, generating implementation briefs and QA scenarios across web, Windows and 3 integrations.

ContentForge AI – Brand-Aware Content Platform | Next.js, TypeScript, Agentic AI, Tool Calling

- Built a brand-guided generation workspace with repurposing, saved drafts and Google Calendar-synced planning, producing 8 content formats with up to 3 distinct variations per brief in one reusable brand voice.

Synapse – Multi-Agent Document Intelligence | PostgreSQL/pgvector, LLM Eval, Finetuning, OCR, Multi-Agent Orchestration

- Built a multi-tenant RAG platform with GPU ingestion, hybrid retrieval and knowledge graphs, orchestrating grading and verification agents to reach 83.9% document hit@5 and 80% judged-correct answers.

AutoBG – AI Car Background Replacement | BiRefNet, SDXL, FastAPI, GCP, Segmentation

- Built a car-matting and studio-compositing pipeline with reflection inpainting and AI backgrounds, served on an NVIDIA L4 at ~0.7s per template, ~13s per reflection and ~18s per background.

EDUCATION

B.S. in Artificial Intelligence | Sir Syed CASE Institute of Technology

Jun 2022 - Jun 2026

- CGPA: 3.91/4.0. Majors in NLP and Computer Vision. Thesis: "A Scalable Multi-Tenant RAG System with GPU-Optimized Retrieval."