

Khuzaima Hassan

AI/ML Engineer | Data Science Graduate | Python Developer

khuzaimahassan52@gmail.com +92 321 3937589 Karachi, Pakistan [linkedin.com/in/khuzaima-hassan-2004mkh](https://www.linkedin.com/in/khuzaima-hassan-2004mkh)
github.com/KhuzaimaHassan khuzaimahassan.me

PROFESSIONAL SUMMARY

AI/ML Engineer and Data Science graduate with hands-on experience building machine learning, Generative AI, agentic, and MLOps solutions. Proficient in Python, SQL, FastAPI, cloud platforms (AWS, GCP), and LLM integration (Gemini, Groq, Qwen, Llama) for developing production-ready AI applications, including RAG pipelines, LangGraph agents, and voice AI systems. Experienced in end-to-end ML pipelines, model serving, and scalable AI systems through industry internships and real-world projects.

TECHNICAL SKILLS

Programming: Python, SQL, TypeScript.

Machine Learning: Classification, Regression, Deep Learning, LSTM, XGBoost, LightGBM, SHAP, Feature Engineering, Model Evaluation, Cross-Validation, Model Monitoring, Drift Detection, MLOps.

Generative AI & LLMs: Gemini API, Groq API, LangChain, LangGraph, MCP, Qwen, Llama, RAG, FAISS, pgvector, Vector Embeddings, Sentence Transformers, Prompt Engineering, Voice AI (Vapi), GenAI App Development.

Frameworks & Tools: FastAPI, Next.js, Streamlit, MLflow, Feast, Redpanda, Prefect, Evidently, Git, GitHub Actions, Hopsworks.

Cloud & Deployment: AWS (EC2, IAM), GCP (Compute Engine, GPU Instances), Render, Vercel, Docker.

Data & BI: PostgreSQL, Supabase, Redis, MySQL, Power BI, Pandas, NumPy, Matplotlib.

PROFESSIONAL EXPERIENCE

10Pearls | Data Science Intern

Dec 2025 – Feb 2026

- Built and deployed an end-to-end MLOps pipeline for 3-day AQI forecasting (32 features, 15 Optuna-tuned models), automating hourly ingestion, daily retraining, and dashboard deployment.
- Automated data pipelines and model deployment workflows utilizing GitHub Actions.
- Trained and evaluated regression and classification models to forecast AQI levels, applying cross-validation to improve generalization.

Plus W | Japanese Language & IT Internship

Apr 2025 – Sep 2025

- Contributed to backend development by implementing database constraints and migration scripts that improved data integrity and consistency across environments.
- Provisioned and connected to AWS EC2 instances via Ubuntu, gaining practical exposure to AWS services and cloud infrastructure.

SELECTED PROJECTS

Quasar – Full-Stack AI Developer Workspace: [GitHub](#)

- Built a full-stack AI workspace (Next.js, FastAPI, pgvector) with multi-model streaming chat and PDF/DOCX RAG with citations.
- Implemented hybrid vector + BM25 retrieval, long-term memory, and LangGraph agents with MCP tools (GitHub, filesystem).
- Encrypted user API keys with AES-256-GCM, added a RAGAS evaluation suite, and deployed on Vercel and Render.

Logistics Demand & ETA Forecasting Platform: [GitHub](#)

- Built a live MLOps platform with Redpanda streaming, Feast/Redis feature store, LightGBM, MLflow, and Prefect retraining.
- Delivered sub-15ms p95 FastAPI serving and a LangGraph Ops Copilot with read-only tools and prompt-injection guardrails.
- Processed 3.06M taxi trips; added Evidently drift monitoring and a promotion gate (models beat baselines by ~39%).

Voice Patient Registration – Voice AI Intake Agent: [GitHub](#)

- Built a phone-based voice intake agent with Vapi (Soniox STT, GPT-5 Mini), tool calling, and a secret-verified FastAPI webhook.
- Stored records with async SQLAlchemy/Alembic on Supabase Postgres, Pydantic validation, 11/11 tests; deployed on Render.

Job Finder AI – Full-Stack AI Job-Matching Platform: [GitHub](#)

- Built a Next.js/FastAPI platform with Gemini 2.5 Flash for resume parsing, ATS scoring, skill-gap analysis, and cover letters.
- Aggregated jobs from four sources (6-hour refresh) with Supabase auth (RLS) and a Kanban tracker; hosted on Vercel/Render.

TeleTabib – AI Telehealth Agent (Final Year Project)

- Architected an AI-driven telehealth voice agent using LangGraph and open-source LLMs to automate patient interactions.
- Built a FAISS + Sentence Transformers RAG pipeline and ran Qwen on GCP (NVIDIA L4 GPU) as part of a 4-person team.

Risk Scoring Platform – End-to-End ML System: [GitHub](#)

- Built a fraud risk scoring platform (FastAPI, PostgreSQL, React) with sub-50ms scoring, model versioning, and drift monitoring.

EDUCATION

NED University of Engineering & Technology | B.S. Computer Science (Data Science)

Sep 2022 – Jun 2026

- CGPA: 3.885/4.00 (1st Position, Data Science Batch)
- Coursework: Machine Learning, Cloud Computing, Database Systems, Algorithms

Intermediate — Adamjee Government Science College

Aug 2020–Aug 2022

Matriculation — Usman Public School System

Apr 2018– Apr 2020

CERTIFICATIONS & LANGUAGES

Certifications: Machine Learning Scientist in Python (DataCamp), Google Data Analytics Professional Certificate (Coursera), Machine Learning Engineer (DataCamp), Python for Data Science, AI & Development (Coursera), Data Engineering Essentials (Coursera), Huawei HCIA (Datacom).

Languages: English (Professional), Urdu (Native), German (Elementary), Japanese (Elementary)