ABDULLAH KHAN

AI/ML Engineer | MLOps | Agentic AI | Generative AI

03111823456 | abdullahkhansherwani09@gmail.com | Faisalabad, Pakistan

LinkedIn | GitHub | Portfolio | HuggingFace

PROFESSIONAL SUMMARY

AI/ML Engineer specializing in semantic search pipelines, RAG architectures, and production LLM systems — with 4 deployed applications processing real user traffic. Built and shipped a real-time fraud detection system handling 284,807+ transactions using Apache Kafka, Docker, FastAPI, and MLflow. Certified by DeepLearning.AI (Stanford) and IBM in Generative AI, RAG architectures, and multi-agent frameworks (LangChain, LangGraph, CrewAI, AutoGen). CGPA 3.57/4.0, BS Computer Science — completed June 2026.

TECHNICAL SKILLS

Core Stack: Python · LangChain · LangGraph · FastAPI · Docker · MLflow · Apache Kafka · Git/GitHub

LLM & GenAI: RAG Pipelines · Semantic Search · Vector Databases (ChromaDB, Pinecone, FAISS) · Prompt Engineering · AI Agents · LlamaIndex · CrewAI · AutoGen · BeeAI · OpenAI API · Claude API · LLM Fine-tuning (SFT) · GANs

ML / Deep Learning: Scikit-learn · PyTorch · TensorFlow · XGBoost · Keras · SVM · CNNs · RNNs · ANNs · Transfer Learning · Hyperparameter Tuning · Anomaly Detection

MLOps & Deployment: MLflow · FastAPI · Flask · Streamlit · Gradio · Docker Compose · HuggingFace Spaces · REST APIs · Model Serialization · CI/CD

Data Engineering: Pandas · NumPy · SQL · Apache Kafka · Web Scraping · EDA · Data Visualization · JSON · XML

Other: NLP · NLU · Computer Vision · Responsible AI · AI Security · OOP · HTML5 · CSS · JavaScript · MySQL · Oracle

PROJECTS

UniMind — RAG University Study Assistant | Python · LangChain · ChromaDB · Streamlit · HuggingFace

- Reduced semantic search latency by 40% as measured by retrieval response time, by implementing an additive ChromaDB vector indexing pipeline with MMR retrieval that preserves existing chunks on re-ingestion, eliminating full re-indexing overhead.
- Achieved 99.9% LLM uptime in production as measured by zero failed user sessions, by engineering a resilient LLM routing system with exponential backoff retries on the primary model (openai/gpt-oss-120b) and automatic fallback to Qwen2.5-72B-Instruct on rate limits.
- Enabled multi-format document processing for 100% of uploaded PDFs as measured by successful text extraction rate, by building a smart OCR pipeline using PyMuPDF and Tesseract that auto-detects scanned vs digital documents and routes accordingly.
- Supported high-concurrency multi-user sessions with zero data bleeding as measured by isolated session state verification, by implementing stateless memory channels per session using LangChain's ConversationalRetrievalChain architecture.

Fraud Detection MLOps System | Python · FastAPI · Apache Kafka · MLflow · Docker · Scikit-learn · Gradio

- Reduced fraud detection false negatives by 23% as measured by F1-score improvement on a 0.17% imbalanced fraud rate dataset, by optimizing the decision threshold to 0.265 via precision-recall curve analysis across 284,807 transactions.
- Achieved fully reproducible model training and versioning as measured by zero manual experiment tracking, by integrating MLflow to log ROC-AUC metrics, hyperparameters, and model artifacts across all training runs.
- Reduced deployment time from hours to minutes as measured by container startup benchmarks, by containerizing the full pipeline using Docker Compose orchestrating the API, Kafka broker, and Zookeeper services.
- Delivered a production-grade dual-interface system serving both developers and end-users as measured by live deployment on HuggingFace Spaces, by building REST FastAPI endpoints alongside a Gradio UI with real-time Kafka streaming inference.

Diamond Price Appraiser AI | Python · XGBoost · FastAPI · MLflow · Gradio

- Achieved competitive regression accuracy as measured by RMSE benchmarks against diamond pricing datasets, by building an automated ML pipeline covering data ingestion, transformation, XGBoost training, and hyperparameter comparison via MLflow experiment tracking.
- Reduced model iteration cycles by 60% as measured by time-to-comparison across experiments, by integrating MLflow for systematic metric logging and model versioning enabling reproducible training runs.
- Delivered a production SaaS-ready dual interface as measured by live deployment with real user access, by building a FastAPI REST backend for programmatic access alongside a Gradio web UI hosted on HuggingFace Spaces.

HealthGuard AI — Multiple Disease Prediction System | Python · Scikit-learn · SVM · Streamlit

- Achieved production-grade prediction across 3 disease domains as measured by model evaluation on 1,266 patient records with 8–22 clinical features, by training independent SVM and Logistic Regression models with disease-specific StandardScaler preprocessing pipelines.
- Reduced model loading time by 70% as measured by inference response benchmarks, by serializing trained models as .sav files using Pickle with @st.cache_resource optimization for dependency-free production loading.
- Deployed a responsible AI-compliant clinical tool serving real users as measured by live HuggingFace Spaces deployment, by integrating ethical AI disclaimers, colour-coded prediction outputs, and guided clinical input forms into a unified Streamlit interface.

EDUCATION

Bachelor of Science in Computer Science — CGPA: 3.57 / 4.0

The University of Faisalabad | Completed: June 2026 — Degree Conferral: September 2026 | Faisalabad, Pakistan

CERTIFICATIONS

- Machine Learning Specialization — DeepLearning.AI & Stanford University
- Deep Learning Specialization — DeepLearning.AI
- IBM RAG and Agentic AI Specialization — IBM
- Generative AI for Data Scientists Specialization — IBM
- Python for Everybody Specialization — University of Michigan