

WALEED ALI

Full Stack AI Engineer

Sialkot, Pakistan • +92 347 092 5565 • waleedali09az@gmail.com • [LinkedIn](#) • [GitHub](#) • Open to Remote

PROFESSIONAL SUMMARY

Full Stack AI Engineer with hands-on experience designing, building, and deploying production-grade AI applications. Specialized in Generative AI, Retrieval-Augmented Generation (RAG), Voice AI, Large Language Models (LLMs), and scalable backend systems. Experienced in real-time conversational platforms, AI voice assistants, adaptive learning systems, and intelligent automation workflows using FastAPI, PostgreSQL, WebSockets, and cloud infrastructure. Proven ability to take AI products from concept to deployment, integrating speech technologies, conversational agents, and modern AI frameworks into reliable, user-focused solutions.

SKILLS & TECHNOLOGIES

Programming Languages: Python, SQL, JavaScript, HTML, CSS

AI & Generative AI: LLMs, RAG, AI Agents, Prompt Engineering, Conversational AI, Voice AI, Workflow Automation

Machine Learning & Deep Learning: Machine Learning, Deep Learning, NLP, Computer Vision, Transfer Learning

Frameworks & Libraries: FastAPI, Django, Flask, LangChain, LangGraph, TensorFlow, Keras, Scikit-learn, Pandas, NumPy, FAISS

Speech & Voice Technologies: STT, TTS, Whisper, Real-Time Voice Pipelines, LiveKit, Vapi, ElevenLabs, Deepgram

Databases & Caching: PostgreSQL, Redis, ChromaDB, Pinecone

APIs & Real-Time Systems: REST APIs, WebSockets, Asynchronous Processing, API Integration

Cloud & Deployment: Runpod, Vercel, Linux, GPU Deployment

Tools & Platforms: Git, GitHub, VS Code, Google Colab, Hugging Face, Streamlit, n8n, Webhooks, Airtable, Langfuse

PROFESSIONAL EXPERIENCE

Full Stack AI Engineer | Entropik Labs | Remote

January 2026 – Present

- Built and deployed production-grade Generative AI systems including RAG pipelines, voice AI agents, and real-time conversational applications.
- Developed FastAPI-based backend services and WebSocket communication systems integrated with LLM-powered workflows.
- Worked on GPU-based deployments and production AI systems including voice-to-voice RAG agents and conversational assistants.

AI Engineer | GulzarSoft | Full-Time

September 2025 – January 2026

- Developed AI applications using LLMs, LangChain, and agent-based architectures for production-ready systems.
- Built automation workflows, embedding-based retrieval systems, and machine learning pipelines.
- Contributed to backend services and AI integration for multiple client-facing applications.

AI Engineer Intern | GulzarSoft | Full-Time

June 2025 – August 2025

- Worked on machine learning models, data preprocessing, and LangChain-based applications.
- Assisted in building RAG prototypes, embeddings workflows, and basic AI automation systems.

PROJECTS

Voilo – AI-Powered Voice Language Learning Platform

Developed backend services using FastAPI for a real-time AI language-learning application powered by OpenAI Realtime models and WebSocket-based conversational pipelines. Built adaptive learning features including vocabulary tracking, mistake memory, and personalized difficulty adjustment supported by PostgreSQL-backed learning analytics. Deployed scalable backend infrastructure and integrated AI services into a cross-platform mobile application.

Stack: FastAPI, OpenAI Realtime API, WebSockets, PostgreSQL

Urdu Voice-to-Voice RAG Agent

Built an end-to-end Urdu conversational AI assistant using open-source STT, LLM, and TTS models from Hugging Face. Implemented a RAG pipeline for context-aware responses and deployed AI inference on Runpod GPUs with a frontend hosted on Vercel.

Stack: Hugging Face (Whisper, LLM, TTS), RAG, Runpod, Vercel

LiveKit Doctor Receptionist with Langfuse Traces

Built a real-time AI receptionist using LiveKit agents, Google Gemini, ElevenLabs TTS, and Deepgram STT. Integrated Langfuse for tracing, logging, prompt evaluation, and production-grade observability. The agent manages appointments, refreshes patient data, responds conversationally, and handles fallback logic with full traceability.

Stack: LiveKit, Gemini, ElevenLabs, Deepgram, Langfuse

WhatsApp Automation Agent (n8n + Unipile + Airtable)

Built an end-to-end WhatsApp automation system using Unipile and n8n webhooks with an LLM-based classifier to route user queries to specialized agents:

- Order Tracking Agent: Fetches order status from Airtable and responds to users.
- Product Info Agent: Retrieves product knowledge from Pinecone vector database.
- Technical Support Agent: Sends automated emails to support teams and confirms user follow-up.
- Billing Agent: Handles billing-related issues with structured responses.

Stack: n8n, Unipile, Airtable, Pinecone, LLM classifier, Webhooks

Voice AI Sales Agent (Vapi + Airtable)

Developed a Vapi-powered voice agent for an e-commerce store. Integrated Airtable as a dynamic product knowledge base for real-time conversational product discovery, FAQs, and buying assistance at scale.

Stack: Vapi, Airtable, Voice AI

Final Year Project: AI-Powered Learning Assistant

Developed a video summarization and notes-generation system using Hugging Face models (Whisper, Pegasus, DeepSeek). Implemented a transformer-based book recommender using FAISS for personalized study assistance.

Stack: Whisper, Pegasus, DeepSeek, FAISS, Hugging Face

YouTube Video Q&A Chatbot

Built a Streamlit chatbot that extracts multilingual YouTube transcripts, embeds them with MiniLM, and retrieves context-grounded answers using a Groq LLaMA 3.3 70B RAG pipeline.

Stack: LangChain, LLaMA 3.3, MiniLM, Groq, Streamlit

Driver Drowsiness Detection — Computer Vision

Created a real-time system using EAR, facial landmarks, OpenCV, and YOLO to detect driver drowsiness and trigger alerts.

Stack: OpenCV, YOLO, Python

EDUCATION

Bachelor of Science in Computer Science — University of Gujrat

2021 – 2025 • CGPA: 3.37 / 4.0