MUHAMMAD FAHAD

AI ENGINEER

Building production AI systems: RAG, agentic workflows, speech AI, LLM applications

www.fahadai.site
Islamabad, Pakistan | UTC+5

CONTACT

mfahad2405@gmail.com
+92 349 532 6867
linkedin.com/in/fahad2703
github.com/M-Fahad27
www.fahadai.site

EDUCATION

BSc Computer Science
NUML — Islamabad
2023 — 2027

CERTIFICATIONS

AWS Solutions Architect
Amazon Web Services

AWS Cloud Practitioner
Amazon Web Services

Certified AI Engineer Assoc.
School of AI — Udemy

CORE SKILLS

LLM / AI
Python | LangChain | LangGraph | RAG | Prompt Eng. | Groq | Ollama

DATA / VECTOR
Qdrant | pgvector | Pinecone | Postgres | Redis

INFRA / OPS
Docker | Nginx | AWS EC2/S3 | Lambda | FastAPI | n8n | Git

SPEECH AI
Whisper | Pyannote | Diarization

TOOLS
Streamlit | Flowise | REST APIs | ClickUp API | Gmail API

LANGUAGES

English — Professional
Urdu — Native

INTERESTS

Production AI systems
Open-source tooling
System design
Cloud architecture

WORK EXPERIENCE

AI Engineer — Onsite

Dec 2025 — Present

Veeivs

- Deployed **DocuMind**, a production RAG chatbot (Qdrant + Ollama + Groq) on Ubuntu 24 with Docker, Nginx, and HTTPS — **99.4% uptime**, <2s median latency across a 10K+ document corpus.
- Built an **AI Meeting Assistant** (agentic pipeline) that auto-transcribes recordings via Whisper, extracts summaries, decisions & action items, emails recaps, and creates ClickUp tasks — **3-min turnaround per 60-min meeting, 100% auto-delivery rate**.
- Engineered a **Speaker Separation & Transcription** pipeline enabling accurate multi-speaker call diarization for downstream AI analysis.
- Automated **2–3 core business workflows** using Agentic AI bots and built a RAG chatbot with LLMs, cutting repetitive manual effort across teams.

AI Developer — Remote (Project-Based)

Nov 2025 — Present

Provoxio Ltd.

- Engineered **Speak-Align**, processing **100 call recordings/day** — auto-evaluates script compliance, order violations, and agent deviations using Whisper + pyannote diarization — **reduced manual QA review time by 55%**.
- Built a **38-criterion scoring rubric** with timestamped flags at violation points and a daily ops digest delivered by 9 AM.

KEY PROJECTS

SupportIQ — AI Customer Support Agent

LIVE

aisupport.fahadai.site

Full-stack AI support agent: Groq reasoning, Qdrant vector search, Ollama embeddings, RAG, intent/sentiment routing, order lookup, auto-ticketing, escalation logic, and human-agent handoff summaries. **8-stage AI pipeline, 3 Docker services, deployed on AWS EC2 behind Nginx.**
Groq · Qdrant · Ollama · Streamlit · Docker · Nginx · AWS EC2

DocuMind — Source-Grounded Document Q&A

LIVE

Production RAG chatbot with hybrid retrieval (dense + BM25), Cohere rerank, citation-grounding over **10K+ documents**. Per-tenant isolation, hot-swappable embedding models. **99.4% uptime, <2s p50 latency.**
Streamlit · Flowise · Qdrant · Ollama · Groq · Docker

AI Meeting Assistant — Meeting Intelligence

PROD

Agentic pipeline: auto-detects recordings, transcribes via Whisper, extracts decisions & action items, emails recaps, creates ClickUp tasks — **zero human intervention, 3-min turnaround.**
n8n · Groq Whisper · Cloudinary · Gmail API · ClickUp

AI Sentiment Analyzer — Interview Analysis

OSS

Records candidate speech, transcribes audio, evaluates sentiment & confidence with transformer LLMs, generates structured JSON feedback with scoring across 3 evaluation axes.
Python · Streamlit · Gemini API · Transformers

5+	99.4%	55%	100/d	3 min	<2s
AI Systems Shipped	Production Uptime	QA Time Automated	Calls Scored	Meeting Recap	RAG Latency

TECHNICAL HIGHLIGHTS

- Built **end-to-end RAG pipelines** (embed → chunk → retrieve → rerank → cite) serving real users with sub-2s latency.
- Designed **agentic AI workflows** with tool execution, memory, routing, and human-in-the-loop checkpoints in production.
- Implemented **AI intent/sentiment classification** with structured JSON output parsing for automated routing decisions.
- Deployed **5 AI systems on AWS EC2** using Docker Compose, Nginx reverse proxy, TLS, and custom domains.