

Usmar Haider

AI Engineer — Voice Agents · RAG · Multi-Agent Systems

Lahore, Pakistan · officialusmarhaider@gmail.com · +92 310 7277147 · linkedin.com/in/usmarhaider · github.com/UsmarHaider

EDUCATION

MS in Artificial Intelligence — LUMS, Lahore 2026
BS in Data Science — FAST NUCES, Lahore 2024

EXPERIENCE

Full Stack AI Engineer · DFX5 USA · Remote | Jun 2025 – Jun 2026

HealthAI — multi-tenant speech-to-speech healthcare consultation platform used by real clinics

- Built a **real-time speech-to-speech** clinical consultation platform using **Amazon Nova Sonic**, **Bedrock**, and **Claude**, supporting English, Spanish, and Hindi.
- Designed a **multi-agent architecture** with **4 healthcare specialists** (nurse intake, cardiology, neurology, functional medicine) using **MCP tools**, with smart triage, red-flag overrides, and outbound calling.
- Developed the **FastAPI** backend with **WebSocket streaming** and **JWT/Cognito auth**, plus a **React/Vite** frontend with push-to-talk UI.
- Shipped **AI-generated SOAP notes**, prescription drafting with doctor sign-off, and a **Clinical Intelligence Engine** that turns intake, documents, labs, and vitals into clinical briefs.
- Hardened the platform for **HIPAA**: tenant isolation at the database layer, **PHI encryption**, audit logging, and **300+ automated tests** across backend, frontend, and e2e.

Senior AI Engineer · Eye4Tech Remote | Sep 2023 – Jun 2025

Led RAG and conversational AI work across psychology, sales, and WhatsApp commerce products

- Built the **HAI chatbot** for Humanop Quantum Psychology: a **RAG-into-RAG** system on **FAISS** generating personalized wellness plans from user assessments, with **admin feedback** folded back into retrieval.
- Deployed **FAISS vector databases** on **AWS S3** with automatic re-embedding on new data, plus **multi-file search** in FastAPI returning similarity scores and source attribution.
- Developed a **sales AI chatbot** that classifies queries as general or data-related, routing data queries through **ML retrieval** and general ones through an **S3 knowledge base — multi-vendor** with real-time updates.
- Integrated a **WhatsApp RAG chatbot** over **Twilio** with a **Pinecone** pipeline and **MongoDB**-backed per-user conversation history.

AI Engineer (Contract) · VitalsVault Healthcare | 2024

Voice-based medical assistant with fine-tuned medical LLMs

- Cut voice response latency from **15s to ~3.4s (75% improvement)** by rebuilding the streaming pipeline.
- Fine-tuned MedGemma 4B and 27B** models for accurate medical knowledge responses.
- Implemented **GraphRAG** over structured healthcare datasets and automated **SOAP note generation** with secure email delivery.

Junior Data Analyst · BornGreat USA · Remote | Sep 2022 – Mar 2023

- Extracted and analyzed data from **Instagram, Facebook, Reddit, and Twitter** for real-time market and competitor insight.
- Applied **NLP and sentiment analysis** to track consumer opinion, sharpening campaigns and engagement.

SELECTED PROJECTS

ConnectHealth Workspace — clinical operations platform Nova Sonic · Bedrock Claude · Amazon Connect · React · FastAPI · AWS CDK

- Live call console with **real-time speech-to-speech patient check-in** (16 kHz in / 24 kHz out, **barge-in**) — the intake form fills itself as the patient speaks.
- Ambient medical scribe**: dual-channel transcription with Bedrock Claude **SOAP drafts** handed to the EHR, plus AI-suggested **ICD-10/CPT codes** with confidence scores and a full accept/reject **audit trail**.
- Infrastructure as code**: CloudFront + S3 + ALB + **ECS Fargate** via **AWS CDK** with keyless IAM task roles.

Agentic Bank — voice + chat banking agent

Nova Sonic · Bedrock · MCP · BM25 · DynamoDB · WebSockets

- **One banking agent across two modalities** sharing a single session store — customers start a request on a call and finish it in chat, including real, capped **money transfers**.
- **MCP server** for the voice path plus native **Bedrock tool calling** for chat; custom question-weighted **BM25 retrieval** over a 677-entry knowledge base.
- **Security**: PBKDF2 hashing, httpOnly cookies, **60-second scoped WebSocket tokens**, per-route rate limits.

NoorAI — source-cited Islamic knowledge assistant

LangGraph · GPT-4o-mini · Pinecone · Next.js · SSE

- **LangGraph pipeline** tuned to **3 LLM calls per answer**; **42 parallel Pinecone queries** with trilingual EN/UR/AR keywords over a **22,000+ document corpus** (Quran, Hadith, Tafseer).
- **SSE streams** the agent's reasoning, each retrieved source with relevance score, and a final **confidence rating**.

LMS-AI — AI-native learning platform

LangChain · Gemini · Pinecone · FastAPI · MongoDB · React

- **Four cooperating agents** (planning, web search, content, PDF export) turn course material into **PPTX decks across 19 typed layouts** — one LLM pass per slide, 5 themes applied cheaply.
- **RAG-grounded tutoring** with per-resource **Pinecone namespaces** and **ReAct chatbots**; resilient Gemini client with **multi-key rotation** and 4-model fallback.

Production voice agents — healthcare & restaurant

Voice-to-voice · WebSockets · GPT · Twilio

- **Hospital appointment agent**: patients book by phone against live physician availability — **98% intent accuracy**, sub-2s responses, **99.9% uptime**, email confirmations and secure cancellation codes.
- **Cheezyous restaurant ordering agent**: full-menu voice ordering with customizations and live order status — **96% order accuracy** at ~90s average order time.

PERSONAL AI ASSISTANT — OPENCLAW × HERMES AGENT

Self-hosted, self-improving personal assistant reachable on WhatsApp

OpenClaw · Hermes Agent (Nous Research) · WhatsApp · Email · OneDrive · Calendar

- Deployed **OpenClaw**, an open-source **self-hosted AI assistant gateway**, connected to **WhatsApp** and configured with **email, OneDrive, and calendar** integrations.
- Paired it with **Hermes Agent** (Nous Research) — a **self-improving agent** with persistent memory that writes **skill documents** from completed tasks, so it **automatically learns** my routines and preferences over time.
- **Email autopilot**: the agent triages the inbox, **drafts and sends emails** on my behalf, and pushes important updates straight to **WhatsApp notifications**.
- **Conversational scheduling**: create, **schedule, and reschedule calendar events entirely through WhatsApp**, with conflict checks and confirmations before anything is booked; grounded access to **OneDrive files** for context.

AWARDS

🏆 **2nd Place** — /execute: **AI Genesis Hackathon (Global · Dubai, 2025)** — **HealthSync.AI**: multi-agent medical assistant; **fine-tuned Mistral 7B** for blood-report analysis, agents for hospital location, appointment booking, and skin disease detection.

🏆 **3rd Place** — Traversaal × **Optimized AI Global Hackathon (2025, solo)** — **Blood.AI**: blood-health assistant with **disease-prediction and blood-type models**, hospital and donor-organization locator agents.

🏆 **2nd Place** — Transcend AI Hackathon (Pakistan-wide) · **NETSOL** — **FARM.AI**: agriculture assistant with **plant-disease detection** (AlexNet, 36 labels), live weather and price agents, and a multilingual farmer chatbot.

SKILLS

Voice & Agents	Amazon Nova Sonic, Amazon Connect, MCP, LangGraph, LangChain, OpenClaw, Hermes Agent, multi-agent orchestration, Twilio
LLMs	Claude (Bedrock), GPT-4o, Gemini, Mistral, MedGemma — fine-tuning, tool calling, prompt engineering
Retrieval	RAG, GraphRAG, Pinecone, FAISS, BM25, embedding pipelines
Backend & Cloud	Python, FastAPI, WebSockets, SSE, AWS (Bedrock, S3, DynamoDB, ECS, Cognito, CDK), MongoDB
Frontend	React, Next.js, TypeScript, Vite, Tailwind CSS

CERTIFICATIONS

Claude Code: **A Highly Agentic Coding Assistant** (2025) · **Web Scraping** — freeCodeCamp (2023) · **Data Manipulation & Joining Data with Pandas** — DataCamp (2022)