HUSSAIN RAZA SHAH JAFFERY

AI & Backend Engineer | Healthcare AI · LLMs · Speech & Voice Systems

hussainrazashah536@gmail.com | +92 310 5303331 | github.com/hussain-raza-ai | Lahore, Pakistan

Portfolio: hussain-raza-ai.vercel.app | **LinkedIn:** linkedin.com/in/hussain-raza-ai-engineer

PROFESSIONAL SUMMARY

AI Engineer building production healthcare AI for IMC Hospital at Contech International, hired full-time after a five-month internship. Built MedScribe AI, which turns a 30-second doctor dictation into a structured 7-field clinical record in about 2 seconds, with an AI drug-safety review, running entirely on the hospital's own GPU server. Experienced in speech and language models, LLM-based extraction of lab-report data, Urdu voice agents, FastAPI microservices and n8n automation.

TECHNICAL SKILLS

Languages & Frameworks: Python, FastAPI, Pydantic, JavaScript, SQL, React, Tailwind CSS, Chart.js, REST APIs, Webhooks

AI & LLM Systems: MedGemma (4B, 27B), GPT-4o-mini, Google Gemini, Anthropic Claude, LangChain, Ollama, PyTorch (CUDA), Hugging Face Transformers, Docling OCR, RapidFuzz, Prompt Engineering, NLP

Voice AI & Automation: Google MedASR, Groq Whisper, Pipecat, Edge TTS, Real-Time Voice Pipelines, n8n, Gmail API, Google OAuth2

ERP, LMS & Databases: Frappe Framework, ERPNext v15, Frappe Education, Moodle 4.5, PostgreSQL, Supabase, MariaDB

Infrastructure & Practices: Linux (SSH, VPN), Docker, On-Premise GPU Deployment, Vast.ai, Vercel, Git & GitHub, RBAC, Row-Level Security, QA Testing

PROFESSIONAL EXPERIENCE

Contech International Health Consultants — *Lahore, Pakistan*

AI Engineer — Jul 2026 – Present

AI Engineering Intern — Feb 2026 – Jun 2026

Healthcare AI for IMC Hospital: speech AI, document extraction, voice agents, FastAPI backends and n8n workflows

- Promoted from AI Engineering Intern to full-time AI Engineer within 5 months.
- Delivered three healthcare AI systems for IMC Hospital from prototype to functional backend services: the MedScribe AI voice-to-prescription system, a lab-report extraction pipeline and an Urdu voice intake agent.
- Prototyped medical LLM and speech models on Vast.ai cloud GPUs and migrated production inference to on-premise GPU hardware, keeping patient data within the hospital network.
- Set up a university ERP and LMS on Frappe ERPNext and Moodle for admissions, fees, academics and exams.

KEY PROJECTS

Full case studies with architecture and results: hussain-raza-ai.vercel.app

MedScribe AI — Voice-to-Structured-Prescription System — *IMC Hospital*

Python · FastAPI · PyTorch (CUDA) · Google MedASR · MedGemma 27B · RapidFuzz · On-Premise GPU

- Built a FastAPI backend that turns a single 30-second doctor dictation into a complete structured clinical record across 7 EMR fields (vitals, history, allergies, examination, diagnosis, prescription and lab orders), with doctor review and approval in the hospital's Flutter app.
- Engineered a rule-based fuzzy-matching engine (RapidFuzz) that resolves every dictated drug and test against the hospital's own catalogue of 12,640 medicines and 1,694 lab tests, down to the exact item ID the pharmacy dispenses against.
- Deployed Google MedASR and MedGemma 27B fully on-premise on the hospital's GPU server, keeping patient data within the hospital network; MedGemma reviews every prescription for dosing errors, drug interactions and contraindications before approval.
- Prototyped on Vast.ai cloud GPUs, then moved inference on-premise: ~2-second dictation-to-record and sub-5-second safety reviews, with every output schema-validated.

Smart Query Routing & Email Automation System *Final Year Project, Virtual University*

Python · FastAPI · React · Supabase · LangChain · Google Gemini · Anthropic Claude · n8n · Google OAuth2

- Engineered an AI-powered query management hub that classifies incoming web and email queries by intent and routes each one to the right department automatically.
- Integrated Google Gemini and Anthropic Claude through LangChain to generate contextual draft replies for staff, with a keyword-based fallback that keeps routing running when AI services are unavailable.
- Developed a FastAPI REST backend on Supabase with passwordless Google OAuth2 sign-in, role-based access control for 5 user roles, a 3NF schema with row-level security, and append-only audit logging.
- Automated Gmail ingestion, 24-hour SLA escalations and daily summaries with self-hosted n8n, and built a React dashboard with editable AI drafts and Chart.js analytics.

Lab Report Extraction Pipeline *IMC Hospital (Telemedicine)*

Python · FastAPI · n8n · Docling OCR · MedGemma-4B · Vast.ai · Docker

- Built an end-to-end pipeline, deployed at IMC Hospital, that pairs Docling OCR with MedGemma-4B to turn unstructured PDF lab reports into structured, doctor-verified summaries.
- Resolved failing n8n multipart file uploads by redesigning document transfer as base64-encoded JSON payloads, restoring reliable document ingestion.
- Engineered MedGemma-4B prompts for structured clinical output and enforced mandatory doctor verification of every extracted result as a patient-safety control.

Integrated University ERP & LMS Platform *GHURKI University (Ongoing)*

Frappe Framework · ERPNext v15 · Frappe Education · Moodle 4.5 · MariaDB · REST APIs · Linux

- Architected and deployed a self-hosted university platform on Frappe ERPNext, Frappe Education and Moodle, digitizing the full student lifecycle: admissions, fees, academics, examinations and degree completion.
- Engineered a rule-based academic compliance layer that enforces attendance, fee-clearance and course-prerequisite policies, replacing manual exam-eligibility verification with a single automated check.
- Designed the API-first integration: the admissions portal pushes admitted students into the ERP, which syncs students and enrolments to Moodle, removing duplicate data entry.
- Led platform decisions: chose Frappe + Moodle over OpenEduCat (Odoo) and secured server access over VPN.

“Dr. Sara” Urdu Voice Medical Intake Agent *IMC Hospital*

Python · Pipecat · Groq Whisper · GPT-4o-mini · Edge TTS · n8n

- Built an Urdu-language voice agent for patient medical intake, chaining Groq Whisper speech-to-text, GPT-4o-mini reasoning and Edge TTS speech output in a real-time Pipecat pipeline.
- Debugged cascading n8n workflow failures (broken node connections, URL redirect errors and malformed JSON payloads), restoring reliable delivery of intake data to the IMC backend.

OTHER PROJECTS

- **AI CCTV Workplace Monitoring (In Progress):** Computer-vision system that measures employee idle time from seat presence and PC activity with open-source models; object-pickup alerts are next.

EDUCATION & CERTIFICATION

Bachelor of Science in Computer Science *Expected Dec 2026*
Virtual University of Pakistan

ICS (Intermediate in Computer Science) *2019 – 2021*
Federal Government College, Jhelum

CCNA: Introduction to Networks *Dec 2025*
Cisco Networking Academy