

HAMZA QURESHI

Senior Full Stack AI Engineer

📞 +92-313-4681517 ✉ qhamza4532@gmail.com 📍 Lahore, Pakistan  [LinkedIn](#)  [Github](#)

ABOUT ME

Senior Full Stack AI Engineer with 5 years shipping production agentic systems – currently team lead at Evren AI architecting LangGraph-based platforms for US clients, including a live workplace mediation product in active pilot and a HIPAA-aware multi-agent financial assistant integrating 5 healthcare platforms.

WORK EXPERIENCE

Nov 2024 - Present

Lead AI Engineer - Evren AI

- Serve as technical team lead architecting and delivering LangGraph-based agentic systems for US clients – including Common Ground, a live AI workplace mediation platform in active pilot, and Jarvis, a HIPAA-aware multi-agent financial assistant integrating 5 external healthcare platforms with SSE streaming and anti-hallucination validation
- Led AI engineering on Converscraft across a 5-person team delivering in 3 months – architected a real-time voice pipeline (Deepgram STT → GPT-4o → ElevenLabs TTS with interruption support) and broke the latency ceiling to sub-2-second end-to-end response by migrating the WebSocket layer from FastAPI to Next.js
- Engineered ISeedoc as lead backend engineer – a telehealth platform with GPT-4o Vision medical report analysis, real-time WebSocket consultation chat, and automated Google Meet session integration; independently delivered Business Incubator, an AI entrepreneurial pipeline system, solo from requirements to production in 3 months
- Lead and mentor a cross-functional team of 6 engineers across multiple concurrent client projects – owning technical architecture decisions, code reviews, sprint planning, and client-facing engineering communication for US-based stakeholders

Jan 2024 - Present

Senior AI Engineer - DevHexa

- Independently architect and deliver production AI systems for European and US clients – designing stateful multi-agent LangChain pipelines, RAG systems with pgvector semantic search, and GPT-4o powered FastAPI backends with PostgreSQL across healthcare, enterprise, and B2B domains
- Engineered and deployed full-stack client platforms end-to-end on AWS – covering system design, API architecture, Next.js frontend delivery, Docker containerization, and CI/CD pipeline setup – independently owning every layer from database schema through production deployment
- Benchmark and select LLM models across OpenAI, Anthropic, and OpenRouter – owning prompt engineering strategy, agent architecture decisions, and cost-latency tradeoff analysis that directly determine client deployment stack and production performance

Jan 2023 - Oct 2024

Full Stack AI Engineer- Vareya

- Delivered GPT-powered SaaS applications and production REST APIs for international clients during the early LLM adoption wave – progressing from frontend delivery to independently owning complete AI product development within 6 months of joining
- Built and deployed full-stack applications using FastAPI, PostgreSQL, and Docker on AWS – integrating early LangChain workflows and OpenAI APIs into client business platforms as agentic tooling matured through 2023-2024
- Automated complex client business workflows using N8N and Make.com, maintained CI/CD pipelines via GitHub Actions, and managed complete project lifecycle across multiple concurrent engagements independently

Apr 2021 - Dec 2022

Full Stack Developer - Freelance

- Built and delivered web applications for small business clients independently – handling frontend (Next.js, React), backend (Python, Node.js), database design, and cloud deployment across multiple concurrent engagements
- Developed Python-based automation tools and intelligent data processing workflows, laying the technical foundation for production AI engineering work that followed

PROJECTS:

Common Ground: AI Workplace Mediation Platform

- Architected a stateful LangGraph conversation engine handling intent classification, QoR routing, pgvector-based excerpt retrieval, and dynamic term-criterion mapping across structured multi-party mediation sessions – with a real-time crisis detection gate (lexical + LLM classifier with fail-safe fallback) as a safety layer
- Overhauled the memory system from LLM-paraphrased summaries to raw content storage, eliminating meaning drift across long sessions – platform currently live in active pilot with real organizational users

TrackRec: AI Sales Platform

- Rebuilt the core matching intelligence from a rule-based pipeline to a fully LLM-powered semantic matching system – reasoning about candidate-role fit the way an experienced recruiter would; platform now serves 18,000+ verified sellers with a 50% reduction in recruiter hiring time and 94% one-year retention on hires
- Engineered end-to-end resume parsing feature using LLM-based structured extraction of sales metrics (deal sizes, quota attainment, ICPs, territories) directly into candidate profiles, dramatically reducing manual onboarding friction

Facial Track: AI Attendance & Classroom Monitoring System

- Engineered a two-phase facial recognition pipeline – MediaPipe quality gating (blur, pose, liveness) for enrollment and dlib 128-dim euclidean distance embeddings for 1:N real-time batch identification – achieving 90%+ accuracy in live classroom conditions deployed on AWS EC2
- Built a classroom capture client posting webcam frames every 3 seconds to a server-side batch recognition endpoint, with APScheduler background jobs auto-managing sessions from timetable data – live at University of Lahore.

SKILLS

AI & LLMs: LangChain, LangGraph, LangSmith, OpenAI GPT-4o, Anthropic Claude, OpenRouter, Prompt Engineering, RAG, Multi-Agent Systems, MCP (Model Context Protocol), AI Agents

Backend & API: FastAPI, Python, Node.js, RESTful APIs, SQLAlchemy, Alembic, APScheduler, N8N

Voice & Real-Time: Deepgram (STT), ElevenLabs (TTS), WebSockets, SSE (Server-Sent Events)

Frontend: Next.js, React, TypeScript, Tailwind CSS, Framer Motion

Databases: PostgreSQL, pgvector, Redis, MongoDB, NeonDB, FAISS, Pinecone

DevOps & Cloud: AWS (EC2, S3, RDS, Lambda), Docker, Docker Compose, Nginx, GitHub Actions, CI/CD, DigitalOcean, Linux

Computer Vision & ML: OpenCV, MediaPipe, dlib, YOLOv8, PyTorch, CNN, FaceNet

EDUCATION

BS Computer Science

2017 – 2021

Virtual University of Pakistan

CERTIFICATION

AWS Certified Generative AI Developer

AWS

Nov 2025

Agentic AI Developer

PIAIC

2023 – 2025

AWS Certified Machine Learning Engineer

AWS

Aug 2024

Professional Certificate in Machine Learning & AI

MIT

Sep 2021