

Sarfraz Khan

Denver, CO | sarfaraz85@gmail.com | 720.984.7058 | asksarfraz.ai — interview my AI twin | linkedin.com/in/asksarfraz

Applied AI Engineer · AI Platform & Systems Architect

Engineer with 15+ years of backend and platform depth, now architecting production LLM systems in healthcare — matching real patients to clinical trials from live EHR data, including a fully air-gapped 120B-parameter model deployment on a university supercomputer and multi-modal AI patient agents serving the public today. Rare combination: retrieval + LLM reasoning + hardened production backends, proven in regulated healthcare. Built asksarfraz.ai, a live voice + video AI twin of myself — interview it.

TECHNICAL PROFICIENCIES

- **Applied AI / LLM:** RAG & hybrid retrieval / vector search (BM25 + FAISS, pgvector + HNSW, MedCPT medical embeddings, rank fusion), agent orchestration & agentic workflows (LangGraph, LangChain, FastAPI), Hugging Face models, LLM evaluation harnesses & per-criterion eligibility reasoning, LLMops (versioned prompts/schemas, CI-gated evals, monitoring), OpenAI APIs, prompt engineering, structured outputs with validation/repair loops, open-weight model serving (GPT-OSS 120B via Ollama on A100 GPUs), **MCP (Model Context Protocol) server** in production, voice/video AI agents (LiveKit, WebRTC), guardrails (PII/PHI redaction, IRB-grade decision audit trails), LLM cost optimization (context compression, token budgeting)
- **Languages:** Python, Java, TypeScript/Node.js, Shell
- **Full-stack & web:** React, Next.js, Tailwind CSS, HTML/CSS, REST API design, server-side rendering, real-time browser clients (WebRTC), responsive UI
- **Platform & cloud:** distributed systems at scale, Kafka, Redis Streams, Docker, Kubernetes, **Terraform / infrastructure-as-code**, Slurm/Apptainer, **AWS (Bedrock, Neptune, Athena)**, GitLab/GitHub CI/CD, event-driven microservices, REST API design, retry/DLQ patterns, zero-downtime deploys, observability & alerting, multi-tenant architecture, token-based SSO
- **Data & healthcare:** FHIR/EHR integration (SMART-on-FHIR, CDS Hooks), MySQL, PostgreSQL, Oracle, MongoDB, Snowflake; HIPAA/PII-aware design

PROFESSIONAL EXPERIENCE

TrialX — Applied AI Engineer

Jan 2025 – present

- Architected a production clinical-AI platform: a 10-service event-driven pipeline (Redis Streams, dead-letter queues, graceful drain for zero-downtime deploys) that ingests patient records in any format (FHIR, PDF, DOCX, CDA, images), extracts structured clinical facts with LLMs, retrieves candidate trials via hybrid search, and evaluates every eligibility criterion with LLM reasoning — with PII-scrubbed logging for HIPAA. Provisioned and deployed the platform's **AWS environment with Terraform and Kubernetes**.
- Led an **air-gapped supercomputer deployment** end-to-end for a university research partnership: replaced cloud LLM APIs with **GPT-OSS 120B served via Ollama on A100 GPUs**, re-architected the cloud microservices platform into Slurm/Apptainer batch jobs, and condensed patient histories **~97% with zero information loss** (validated against real Epic bulk exports) to fit local context windows — a pilot spanning **~14,000 de-identified patients matched in both directions** (patient→trials, trial→patients), with patient data and match results never leaving the university's infrastructure and every LLM decision traceable to source facts for IRB-grade auditability.
- Hardened matching quality through eval-driven iteration: deterministic zero-token validators catching fabricated or mismatched LLM evidence, plus prompt/schema revisions that cut exclusion false-triggers from **55% to 0%** and unknown verdicts by **~85%**.
- Shipped **Judy**, a multi-modal AI patient navigator **live today across 60,000+ public trial listings** — eligibility prescreening by video avatar, voice, or chat from one shared knowledge base (HeyGen, Ultravox, GPT-5): **885**

patient sessions and 204 referrals in its first month in production, with LLM-triggered **warm transfer** (Twilio conference bridge) that hands a patient into a live phone call with a human coordinator mid-conversation.

- Built the agentic operations layer (FastAPI + LangGraph) letting study coordinators run the platform conversationally
 - import trials by registry ID, manage sites/users, generate prescreening questions, every action permission-gated
 - plus the secure plugin rails: encrypted token-based SSO, silent refresh, multi-tenant isolation.
- Delivered a company-wide AI training program (50+ people, separate engineer and business tracks); presented live AI demos to clients and research-institution partners with same-day iteration on feedback.

HubSpot — Senior Software Engineer

Feb 2022 – Sept 2024

- Core contributor (4-person team) to **Hawkeye** — HubSpot Fintech's quality-control platform monitoring Finance processes, external vendor dependencies, and batch pipelines over **100K–1M+ Kafka events daily** — catching mis-payments and bad or failed data updates from internal and external systems before they propagated (Java, Guice, Kafka, AWS, Snowflake, MySQL).
- Built Hawkeye's configurable alerting and reporting layer: rule-based routing of issues to the owning team via Slack, email, and alerts. Mentored an intern engineer.

Xavient Digital / TELUS International — Senior Software Engineer

Mar 2012 – Feb 2022

- Embedded with the client engineering org at **Dish Network** (Xavient acquired by TELUS International, 2018). Led design of Java/Spring Boot REST microservices and the enterprise-wide **Kafka messaging platform** (Dish Event System) — owned its **API design** and added config-driven event filtering (no code change for consumers).
- Tech lead (8–10 engineers) for the WebMethods 9.6→10.3 platform upgrade; introduced GitLab CI + Jenkins CD, replacing manual deployment.
- Built internal automation tooling that lifted team efficiency **10–30%**; solo-built a Python/Django flow-code analyzer cutting service-comprehension time **up to 50%**.

Earlier career: System Analyst / Module Lead, Xavient Software Solutions, India (2008–2012) — SOA/WebMethods services for Dish Network; built the WM Admin tool, raising production-support productivity **70%+**.

FEATURED PROJECT

- **asksarfaraz.ai — Interactive Digital Twin (2026, solo, end-to-end):** built and shipped a live, public voice + video AI version of myself that recruiters talk to instead of just reading this resume — proof I take an LLM product from idea to production alone — **front end to infrastructure, built AI-native with Claude Code as the primary delivery vehicle** (Next.js/React/Tailwind UI, Node agent worker, Terraform/Vercel/Supabase). Full real-time agent pipeline (STT → LLM → TTS → lip-synced avatar) on a consented voice clone, RAG-grounded on my career (pgvector), with cross-visit memory, live meeting-brief generation, LinkedIn/Google SSO, safety guardrails, and cost controls. Ships a public **MCP server** alongside it, so a recruiter's own AI assistant can query my background as tools instead of scraping. Go talk to it — link above.

EDUCATION & CERTIFICATIONS

- **University of Sunderland, UK** — BSc (Hons) Computer Systems Engineering
- **Certifications:** AI Solution Architect (ELVTR, 2025); Deep Learning (DeepLearning.AI / Coursera, 2019)
- **Coursework (not production):** PyTorch, TensorFlow, Go