Syed Muhammad Hassan Raza

Software Engineer — AI Systems & Backend — Islamabad, Pakistan

smhassanraza1.com | smhassanraza1@gmail.com | linkedin.com/in/smhassanraza1 | github.com/smhassanraza1

Software engineer who takes rough problems to production, across the whole stack: the interface people use and the backend, cloud infrastructure and AI behind it. I move fast with AI tools, but the engineering decisions are mine. At Noctal I went from intern to mid-level in about a year. I also do independent research and publish it as preprints.

Experience

Noctal — AI Sound Design Platform (noctal.xyz) Remote — Los Angeles, CA

AI Engineer *March 2026 – July 2026*

Junior AI Engineer *June 2025 – February 2026*

AI Engineering Intern *March 2025 – May 2025*

- Built the platform's core **AI generation pipeline**: it analyzes video and audio with Gemini, matches sounds from a Pinecone library (**RAG**), and runs the steps in parallel on Celery with per-step status and cost tracking.
- Designed and built an **LLM agent** (LangChain, LangGraph) that turns plain-language requests into multi-step actions, with schema-checked, retry-safe tool calls streamed live to the browser.
- Shipped retrieval with hybrid search, reranking and metadata filtering (Pinecone), and built an async pipeline that generated 210K+ multimodal **embeddings** to upgrade it.
- Shipped **prompt-injection** guardrails after benchmarking four classifiers on 1,000+ adversarial prompts.
- Designed the background-processing tier as four autoscaled Celery worker pools on **GKE**, and cut compute cost about 38% while usage grew to about 5,000 users.
- Built and owned **Stripe** billing and multi-tenant team accounts (four-level roles, seat billing, failed-payment recovery), and fixed critical access-control and billing vulnerabilities.

UpStart Commerce — Final Year Project Remote — Manchester, NH

ML Engineer Consultant *September 2024 – May 2025*

- Built CatBoost revenue-forecasting and Transformers4Rec purchase-prediction models, and served them as **FastAPI** services for real-time inference.

Projects

Voice receptionist | *Python, LiveKit Agents, OpenAI Realtime, MCP, A2A* code

- Built a **conversational voice agent** for a fictional clinic that books through its own MCP server and hands billing questions to a second agent over A2A. Speech-to-speech replied in 1.9 s, a three-model pipeline in 4.0 s.

Hijack-proof agent | *Python, AgentDojo, Qwen3.5-4B, Ollama* preprint | code

- Built a red-teaming harness with an automated LLM attacker, and a defense for a small local agent that cut attack success from 38% to 11% on 158 held-out cases.

Technical Skills

Languages & Frontend: Python, SQL, TypeScript, React, Next.js, Tailwind CSS

LLMs & Agents: LangChain, LangGraph, tool calling, MCP, A2A, RAG, Pinecone, Claude, Gemini, LLM evals

Voice & Multimodal AI: LiveKit, WebRTC, OpenAI Realtime, speech-to-text, text-to-speech, Gemini vision, PyTorch

Backend: Django, Django REST Framework, FastAPI, Celery, PostgreSQL, Redis, WebSockets, Stripe, JWT, RBAC

Cloud & DevOps: Google Cloud (GKE, Cloud Run, Vertex AI), Kubernetes, Docker, Helm, CI/CD, Datadog

Data Engineering: Kafka, Airflow, PySpark, dbt, BigQuery

Education

FAST NUCES Islamabad, Pakistan

B.S. Computer Science — CGPA: 3.65/4.00, cum laude, Dean's List (6/8 semesters) *2021 – 2025*