

Ahmed Raza

AI Engineer | Full Stack AI Engineer | AI Software Developer

Karachi, Pakistan | ahmedrz.amjad@gmail.com | +92 330 7051591

[linkedin.com/in/ahmedraza101](https://www.linkedin.com/in/ahmedraza101) | github.com/ahmedraza-96 | ahmedraza.is-a.dev

Summary

AI Engineer (BS in Artificial Intelligence, FAST-NUCES) building production LLM systems – RAG pipelines, tool-using agents, and evaluation pipelines – shipped end to end with FastAPI, Next.js, and AWS. Multiple products live.

Technical Skills

Generative AI: LLM APIs (OpenAI, Claude, Gemini, Groq), LangChain, LangGraph, RAG pipelines, AI agents, tool calling, LLM evals (LLM-as-judge, CI quality gates), Chroma, FAISS, LangSmith

Engineering: Python, TypeScript, SQL; FastAPI, Next.js, React, Node.js; PostgreSQL, MongoDB

Cloud & DevOps: AWS (EC2, S3, Lambda), Docker, Vercel, Terraform, GitHub Actions, Linux, Git

Experience

AI Engineer, HashOne Global – Karachi, Pakistan

Jul 2025 – Jun 2026

- Built and shipped AI chatbots and tool-using agents end to end with LLM APIs and local LLMs.
- Designed RAG pipelines – chunking, embedding, retrieval tuning – to ground answers in source documents.
- Served LangGraph agents through FastAPI APIs; deployed with Docker and CI/CD on AWS.

AI Engineer Intern, Tech Hunt – Karachi, Pakistan

Jun 2024 – Aug 2024

- Built RAG document pipelines in Python with LangChain, the OpenAI API, and FAISS.

Projects

NovaNet Eval Harness – AI support agent with production-grade LLM evals

[Live](#) | [Demo](#) | [Code](#)

FastAPI, LangGraph, Groq Llama 3.3, Chroma, LLM-as-judge, GitHub Actions CI, Next.js, AWS EC2, Vercel

A customer-support RAG agent wrapped in a full evaluation pipeline: 100-case golden dataset, deterministic assertions plus LLM-as-judge grading, response caching, a CI quality gate that blocks prompt regressions before merge, and a Next.js metrics dashboard. Baseline evals caught an intermittent prompt-injection vulnerability.

PSX MCP Server – Published Model Context Protocol server for LLM agents

[PyPI](#) | [Demo](#) | [Code](#)

Python, MCP (FastMCP), httpx, BeautifulSoup, pytest, GitHub Actions CI, PyPI trusted publishing

An open-source MCP server that gives Claude and other LLM agents live Pakistan Stock Exchange data – quotes, indices (KSE-100), company fundamentals, dividends, and announcements – as 10 tool-callable endpoints. Built test-first (78 tests), with CI across Python 3.10–3.13 on Linux and Windows, and published to PyPI for one-line install (`uvx psx-mcp-server`).

Convey AI – Conversational AI interview agent (Final Year Project)

[Live](#) | [Demo](#) | [Code](#)

Next.js, Node.js, OpenAI API, REST APIs

An AI agent that conducts and evaluates mock interviews autonomously in Urdu and English, with an avatar-based interview mode. Deployed and live at conveyai.live.

Cognos – AI agent app (RAG + tools)

[Live](#) | [Demo](#) | [Code](#)

FastAPI, LangChain / LangGraph, Chroma, PostgreSQL, Next.js, JWT, SSE, Terraform, AWS EC2

Lets users chat with a tool-using AI agent over their own documents, with built-in web search. Features JWT authentication, per-user Chroma vector stores, Postgres-backed chat memory, and token-by-token response streaming via SSE. Deployed on AWS EC2, provisioned with Terraform.

CareShare – Medicine donation platform (MERN)

[Live](#) | [Demo](#) | [Code](#)

React, Node.js, Express, MongoDB

Connects medicine donors with people in need. Includes an admin panel that moderates donations and user accounts before listings go live. Deployed and accessible online.

IT Inventory Management System

[Live](#) | [Demo](#) | [Code](#)

FastAPI, MySQL, React.js, REST APIs

Internal tool for tracking company IT assets and maintenance, with a React dashboard on a FastAPI and MySQL backend. Deployed and accessible online.

Education

Bachelor of Science in Artificial Intelligence (BSAI), FAST National University (NUCES), Karachi

2022 – 2026

Final Year Project: Convey AI – conversational AI interview agent (see Projects).