Azaan Khan

+91-9987813771 | mr.khanazaan@gmail.com | LinkedIn | GitHub | HuggingFace

Summary

AI researcher with hands-on experience designing and deploying LLM-powered systems, including production-grade RAG apps, agentic workflows, and domain-specific model fine-tuning. Skilled in end-to-end pipeline development using LangChain, FastAPI, FAISS, and Docker. Strong at building MVPs rapidly and optimizing LLM performance through evaluation, retrieval tuning, and prompt architecture.Also share technical breakdowns and insights on LLMs, RAG pipelines, and AI engineering through LinkedIn articles.

Education

Father Conceicao Rodrigues College of Engineering — Bandra(W), Mumbai
B.E in Computer Engineering — 2022-2026

Mithibai College — Andheri(W), Mumbai
HSC, Class XII — 2020-2022

Experience

Nienna Labs — June 2024 - July 2025
Research Intern

- Evaluated 20+ flagship LLMs (GPT-4o, Grok 3, Qwen 2.5, Claude) to analyze accuracy, hallucination patterns, and domain suitability.
- Designed 10+ optimized prompt architectures, improving content quality by 20% and reducing hallucinations in internal tests.
- Integrated LLMs into an agentic data-extraction pipeline used by 8 researchers, reducing research time for revenue insights by 35%.
- Tech: Python, LangChain, CrewAI, OpenAI API, FastAPI, FAISS, PyTorch, OpenRouter API

TEDxCRCE and Mozilla Campus Club — July 2023 - June 2024
Technical Head — July 2024 - May 2025

- Built a website for the club.
- Set up automation for auto-sending emails.
- Crafted a dynamic email UI template for auto-sending personalized emails.

Projects

fAInance

Built a high-accuracy PDF pipeline using Tesseract OCR, PyMuPDF4LLM, and LangChain to separate text/images and process them with LLMs & VLMs via the OpenRouter API. Delivered reliable financial insights through a Streamlit UI with integrated TTS for voice interaction.

Doc RAG — Production Medical RAG System (actively used by a doctor) **[Link]**

Designed a textbook-strict RAG chatbot for medical students and practitioners using LangChain, FAISS, PyPDF2, and Gradio UI. Implemented custom embeddings, retrievers, and context filters to ensure answers come only from the provided medical textbook. Fully built and deployed at zero cost on HuggingFace Spaces.

Spanish Reasoning LLM — Fine-Tuning (UnSloth) **[Article Link]**

Fine-tuned GPT-OSS using UnSloth & QLoRA to boost reasoning accuracy in Spanish on Google Colab GPU.

Technical skills

Languages : Python, C++, C, HTML, CSS, JavaScript, Docker, MySQL, PostgreSQL, MongoDB, FAISS, ChromaDB, Qdrant, PineCone, Redis

Frameworks and Libraries : Langchain, CrewAI, Agno, LlamaIndex, Flask, FastAPI, Gradio, PyTorch

Tools : Git, Github, Notion, HuggingFace, YOLO, MCP, A2A, Prompt Engineer, n8n

Relevant coursework : DBMS, Operating System, Artificial Intelligence, Machine Learning, NLP

Achievements and Certifications

Achievements : 1st Runner Up at Hackspark 1.0, State-level Hackathon

Specialization : LLM integration, Agentic frameworks, MCPs, RAG, AI calling agents, MVPs

Student of the Year 2020 : Gyan Kendra Secondary School (Best all rounder)

Certifications : Google Cloud Data Analytics Certificate, HF Agents Course, Working with HF, AWS NLP, AWS for CI/CD & Microservices, AWS Cloud Architecting, Python Programming Masterclass, NTSE Exam Qualifier

Participation : Bit N Build Hackathon 2024, Tech-A-Thon 3.0, SIH 2024, SIH 2023