

Smit Faldu

Ahmedabad, Gujarat

☎ [+91-8140950661](tel:+91-8140950661)

✉ smitfaldy2006@gmail.com

🌐 [smit-faldu](https://www.linkedin.com/in/smit-faldu)

🔗 [smit-faldu](https://github.com/smit-faldu)

Summary

AI/ML Engineer specializing in **Generative AI, LLMs, and Agentic AI systems** with expertise in building **multi-agent frameworks, RAG pipelines, and autonomous workflows**. Skilled in **Python, LangChain, LangGraph, AutoGen, and FastAPI**, with experience deploying scalable AI solutions on cloud platforms. Passionate about advancing **next-generation intelligent systems** that combine reasoning, retrieval, and human-AI collaboration.

Tech Stack

Area of Interest: Generative AI, Agentic AI Systems, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Multi-Agent Architectures, Natural Language Processing (NLP), Conversational AI, AI Infrastructure & Deployment

Programming Languages: Python, JavaScript, SQL

Libraries & Frameworks: LangChain, LangGraph, AutoGen, TensorFlow, PyTorch, Hugging Face Transformers, FAISS, Scikit-learn, Pandas, NumPy, FastAPI, Playwright, Streamlit, React (basic), Tailwind CSS

Soft Skills: Problem-Solving, Research & Experimentation, Collaboration, Adaptability, Rapid Prototyping, Critical Thinking

Experience

VertxAI

AI/ML Intern

Mar 2025 – Aug 2025

Remote, India

- Contributed to the development of **Generative AI and Agentic AI pipelines**, focusing on large language model (LLM) integration and retrieval-augmented generation (RAG).
- Designed and implemented **LangChain- and LangGraph-based workflows** to enable multi-agent reasoning and task automation.
- Built and deployed **FastAPI-based backend services** for AI-powered applications, ensuring scalability and high performance.
- Optimized **vector database retrieval (FAISS)** and embeddings pipelines to improve accuracy and latency in production workloads.
- Collaborated with senior engineers to **deploy AI systems on GCP (Cloud Run, Docker)** and integrate CI/CD best practices for faster experimentation.

Education

Sal institute of tech & Engg research Ahmedabad

Pursuing B.Tech in computer engineering

Ahmedabad, Gujarat

May 2024 – present

Government Polytechnic, Ahmedabad

Completed Diploma in computer engineering

Ahmedabad, Gujarat

Aug 2021 – May 2024

Projects

1) AI-powered Mock Pitch Voice Agent

- Built an AI-driven voice agent to simulate startup pitch Q&A sessions, helping founders practice and refine investor communication.
- Integrated speech-to-text (Whisper) and text-to-speech (elevenlabs TTS) for real-time interactive conversations.
- Designed **LangChain + LangGraph-based agentic workflows** for context-aware questioning, evaluation, and adaptive conversation flow.
- Developed a **FastAPI backend with WebSocket streaming**, ensuring low-latency interactive voice experiences.
- Stored and analyzed conversation data in **MongoDB**, enabling performance tracking and feedback insights for continuous improvement.
- Deployed on GCP (Cloud Run + Docker) with scalable architecture and modular design for easy feature expansion.

Tech Stack: Python, FastAPI, LangChain, LangGraph, WebSockets, Whisper, elevenlabs TTS, MongoDB, Docker, GCP

2) Founder–Investor Matching AI System

- Designed an **AI-powered matching engine** that connects startup founders with suitable investors using **semantic search and vector embeddings**.
- Implemented **LangChain + LangGraph pipelines** to handle query understanding, ranking logic, and multi-agent reasoning for better match quality.
- Used **FAISS vector database** to retrieve the top **100 best matches** for each founder profile based on contextual embeddings.
- Developed a **FastAPI backend with WebSocket streaming**, ensuring low-latency interactive voice experiences.
- Stored and managed match results in **MongoDB**, enabling efficient tracking, filtering, and analytics on investor–founder interactions.
- Containerized with **Docker** and deployed on **GCP (Cloud Run)** for scalable, production-ready usage.

Tech Stack: Python, LangChain, LangGraph, FAISS, MongoDB, Docker, GCP

3) ArXiv Research Agent

- Built an **AI-powered multi-agent research assistant** using **AutoGen**, capable of retrieving, summarizing, and answering questions on **ArXiv papers**.
- Designed a **multi-agent architecture** where specialized agents handled retrieval and summarization collaboratively.
- Leveraged **LLMs** for paper understanding and concise, user-friendly outputs.
- Developed an **interactive Streamlit interface**, enabling seamless research exploration and structured outputs (JSON/Markdown).
- Containerized with **Docker** and deployed on **Hugging Face Spaces** for public accessibility and reproducibility.

Tech Stack: Python, AutoGen, Hugging Face Transformers, Streamlit, Docker, Hugging Face Spaces

4) DataAnalyzerAI

- Built an **AutoGen-based multi-agent system** that lets users upload CSV/JSON data and query it in natural language.
- Implemented agents that **generate, execute, and return Python code** to produce analysis and graph visualizations automatically.
- Delivered an **end-to-end data exploration tool** with reproducible results using **Dockerized deployment**

Tech Stack: Python, AutoGen, LLM, Streamlit, Docker, Hugging Face Spaces