

FARAZ MIRZA

+91 9170370343 | farazmirza1023@gmail.com | [LinkedIn](#) | [Portfolio Website](#) | [GitHub](#)

EXPERIENCE

AI Engineer Intern

Dec 2025 – Mar 2026

Ascezen Consulting Private Limited

Remote / Hybrid

- Built **multi-agent LLM systems** (LangGraph, LangChain) with role-based agents for autonomous data workflows, enabling iterative validation and leakage-free ML pipeline generation using Qwen via Ollama.
- Developed **RAG pipelines** (FAISS, embeddings) for conversational querying over documents with context-aware memory.
- Implemented **LLMOps workflows** including structured output parsing, validation loops, and prompt refinement to ensure reliable, interpretable, and consistent model responses.
- Built **ML model APIs using FastAPI** (PyTorch/Keras, Vision Transformers) for structured and scalable inference.
- Improved inference latency through **efficient model selection** within multi-step AI workflows.

Open Source Contributor

Jul 2025 – Oct 2025

GirlScript Summer of Code (GSSoC'25)

Remote

- Contributed to open-source projects with **10+ merged PRs** across frontend, backend, observability, and developer tooling ecosystems.
- Implemented privacy-focused **OpenTelemetry** integration with configurable tracing, metrics, and error reporting for improved observability workflows.
- Built scalable real-time video calling infrastructure using **mediasoup SFU**, replacing P2P mesh topology with server-coordinated multi-user streaming.

PROJECTS

Orra - Open-Source Agent Tracing SDK & Dashboard | Python, FastAPI, LangGraph, PyPI

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

- Engineered an open-source telemetry platform with a lightweight **Python tracing SDK** on PyPI and a Next.js dashboard.
- Designed a **SQLAlchemy and SQLite** database layer utilizing absolute path routing to ensure persistent serverless tracking.
- Architected a FastAPI and LangGraph streaming backend to route agent execution steps into a dynamic node canvas.
- Implemented automated CI/CD pipelines via **GitHub Actions & PyPI Trusted Publishers** utilizing secure OIDC handshakes.

Pipeline.ai - Multi-Agent ML System | Python, FastAPI, LangGraph, Next.js, TypeScript, scikit-learn, Ollama

[GitHub](#)

- Built a **multi-agent LangGraph system** with role-based Ollama agents to automate end-to-end ML pipeline generation.
- Implemented **QA validation loops** to detect data leakage and improper splits with automated retries.
- Developed **regex-based extraction pipelines** (re.DOTALL) to convert LLM outputs into executable Python code.
- Designed a **Next.js-based frontend** enabling interactive pipeline generation and code export.

ContextAI - Low-Latency RAG engine | FastAPI, React, LangChain, FAISS

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

- Built a **Retrieval-Augmented Generation (RAG) system** achieving **sub-200ms retrieval latency** using FAISS and HuggingFace embeddings.
- Integrated conversational memory with LangChain, enabling context retention over **10+ interaction turns**.
- Implemented Server-Sent Events (SSE) for streaming responses, reducing perceived latency by **50%**.
- Optimized deployment on Render using CPU-specific PyTorch builds, reducing container size by **300MB**.

TECHNICAL SKILLS

Languages: Python, JavaScript (ES6+), TypeScript, Java, HTML, CSS

Frameworks: FastAPI, React.js, Next.js, Node.js, Express.js, Tailwind CSS

Machine & Deep Learning: PyTorch, TensorFlow, Keras, Scikit-learn, Pandas, NumPy

Agentic AI & Workflows: LangChain, LangGraph, n8n, Hugging Face, Ollama

Databases: PostgreSQL, MongoDB, Vector Databases (FAISS, Pinecone)

Developer Tools & Cloud: AWS, Docker, Git, GitHub Actions, Linux, Vercel, Postman

CERTIFICATIONS & ACHIEVEMENTS

Certifications: AI Engineer Associate & Deep Learning (DataCamp), AI Agents (Hugging Face), Postman API Expert, Backend Architectures & Open Source AI (Frontend Masters).

Leadership & Achievements: Mentored an international bioinformatics student in Python; SIH 2025 Team Lead (CNN Breed Classifier); Hacktoberfest Supercontributor; GDG Cloud & AI Facilitator.

EDUCATION

Integral University

Lucknow, India

Bachelor of Computer Applications — CGPA: 8.6

Aug 2023 – Present

- Relevant Coursework: Data Structures, Algorithms, Operating Systems, DBMS, AI, Computer Networks, Web Dev