

UJALA MAURYA

MUMBAI, MAHARASHTRA | +91-8850766421 | ujalamaurya.um@gmail.com | [GITHUB](#) | [LINKEDIN](#)

EDUCATION

St. Francis Institute of Technology, Mumbai | University of Mumbai

B.E (Computer Science) | GPA: 3.77/4

2018 – 2022

St. Francis D'Assisi Junior College, Mumbai | Maharashtra Board

12th Standard (Science) | 80.88 %

2015 – 2017

EXPERIENCES

Senior Machine Learning Engineer

Nov 2025 - Present

Quantiphi Analytics Solutions Pvt Ltd. Mumbai.

Care AI Agents | Google ADK | Gemini | Vertex AI | GCP | FastAPI | Multi-Agent Systems

- Engineered and deployed AI agents on GCP using Google ADK, Vertex AI and Gemini, enabling autonomous decision-making and multilingual query handling; improved resolution efficiency by 40%.
- Built a Guard Agent to filter irrelevant queries and govern agent actions, improving response accuracy by 30% and enforcing controlled action execution.
- Cut AI response latency by 50% by introducing Pub/Sub-based asynchronous processing and deploying services on Cloud Run.
- Monitored live production data through Grafana to accelerate debugging and issue resolution across AI services.
- Led end-to-end multilingual AI support across 40+ languages, covering language detection, conversation handling and graceful resolution of language mismatches in production.
- Engineered a real-time guardrails safety framework to detect policy violations and sensitive topics, with configurable escalate, reject and resolve fallbacks across live and preview modes.
- Contributed to the multi-agent orchestration pipeline spanning issue-navigation, core-response and action agents using Google ADK.
- Owned LLM prompt design and knowledge-base / FAQ retrieval through the internal middleware layer (Setu).

Product Engineer (Generative AI)

June 2022 – Nov 2025

Seclore Technologies Pvt Ltd. Mumbai.

Astra AI Platform - Agentic RAG Server LangChain | FastAPI | VespaDB | AWS | Tool Calling | Hugging Face

- Architected and implemented an Agentic RAG pipeline integrating multiple private document sources per team and distributed query handling with LangChain, FastAPI and VespaDB.
- Implemented hybrid search and re-ranking, improving response accuracy by 65%.
- Built FastAPI REST APIs and a Next.js frontend for the AI platform, increasing adoption by 50% across teams.
- Cut LLM hallucinations by 40% through prompt engineering and advanced data filtering.
- Automated resolution of deployment and production queries, reducing dependency on engineering teams by 60%.

KEY PROJECTS

Astra AI Platform - Agentic RAG Server

- Designed and deployed an **Agentic RAG-based internal chatbot** for enterprise teams, enabling fast and accurate contextual responses from private datasets, Reduced hallucinations by **40%**.
- Automated resolution of deployment and production queries, cutting dependency on engineering teams by **60%**.

TECHNICAL SKILLS

Programming Languages: Python, JavaScript, Java, SQL

Frameworks & Platforms: LangChain, LangGraph, FastAPI, Pandas, GCP, Vertex AI, Google ADK

AI/ML: Gen AI, RAG, Transformers, Prompt Engineering, LLM, NLP, Machine Learning, Hugging Face

Tools: Git, Docker, Jenkins, AWS, Linux Shell, CI/CD, Grafana, Sonarqube, Pub/Sub, Gemini