

KEDAR SATHE

✉ kedarsthe49@gmail.com 📞 +91 9561233236 📍 Pune, India [in LinkedIn](#) [🔗 Portfolio](#) [🐙 GitHub](#) [🔗 Leetcode](#)

EXPERIENCE

SDE Intern (AI)

Jan 2026 – May 2026

Unique Schools App

Pune

- Built a production multi-agent RAG system using LangGraph with 3 specialized agents retrieval, reasoning, and response, reducing hallucination rate to near-zero on domain QA tasks.
- Developed and shipped an AI-powered React Native mobile application; integrated LLM-based features for student engagement.
- Implemented system health monitoring and performance tracking for LangGraph pipeline, enabling real-time observability of agent state transitions.

Open-Source Maintainer

Present

Mermaid [🔗](#)

- Maintaining Mermaid.js 87,000+ GitHub stars, industry-standard diagramming library used by GitHub, GitLab, Notion, and millions of developers globally.
- Diagnosed and fixed 8+ critical bugs through systematic root cause analysis across multiple parsers, improving diagram rendering compatibility and reducing failure rates.
- Contributed multiple PRs introducing new diagram functionality and performance optimizations, working collaboratively across a distributed open-source team.

EDUCATION

B.Tech(Computer Engineering)

Jul 2022 – Jun 2026

Vishwakarma Institute of Technology Pune

7.94 CGPA

SB College of Science, Aurangabad

Jun 2020 – Apr 2022

SKILLS

Programming Languages

Python, JavaScript (Node.js), C/C++, Bash/Shell, Assembly (x86, ARM)

AI & LLMs

angChain, LangGraph, LlamaIndex, Hugging Face, DeepSeek, RAG, Multi-Agent Systems, Prompt Engineering

ML-DL

PyTorch, Scikit-learn, Computer Vision, RLHF, RLAI, LLM Fine-tuning (Unsloth, PEFT, LoRA, DPO), Machine Learning, Deep Learning, GAN, CNN, SVM,

MLOps

MLflow, Evidently, Prefect, Docker, Kubernetes, GitHub Actions CI/CD, Model Monitoring, Data Drift Detection

Cloud & Infra

GCP (Vertex AI, Cloud Run, Pub/Sub), Linux/Bash, Prometheus, Grafana, FastAPI, PostgreSQL, pgvector

ACHIEVEMENTS

Paper Published: Enhanced Railway Safety System [🔗](#)

IEEE

Eyantra 2023 (IIT Bombay) [🔗](#)

Ranked AIR 1 in Stage 1; top 10 globally in Stage 2. - [Link](#)

Paper Published: Social Distancing ID Card [🔗](#)

- Published in a Scopus-indexed journal

Paper Published: Efficient Information Extraction from Large PDFs Using Retrieval-Augmented Generation and Large Language Models [🔗](#)

- Published in Nature (Springer) Book

Certifications

- Google Cloud Computing Foundations - [Link](#) [🔗](#)
- Natural Language Processing & Deep Learning - Certified by DeepLearning.AI - [Link](#) [🔗](#)
- Fundamentals of Deep Learning - Certified by NVIDIA - [Link](#) [🔗](#)

PROJECTS

LocalRaG [🔗](#)

Full-Stack AI App · Python · FastAPI · PostgreSQL · Docker · LangChain

- Production RAG application on LLaMA 3 (70B) with PostgreSQL vector store; HNSW indexing delivers 25% faster vector retrieval.
- Dockerized deployment achieves sub-1s response latency for 90% of queries; zero hallucination rate on domain QA tasks measured against ground-truth benchmark.

Snakes & The Golden Apple [🔗](#)

Multi-Agent Deep RL · Python · PyTorch · Deep Q-Learning · Pygame

- Built competitive multi-agent snake game using Double DQN, Dueling DQN, and Prioritized Experience Replay.
- 25D state space with epsilon-greedy policy, 50,000-experience replay buffer; agents achieve 60–70% win rate over 2,000–5,000 training episodes.
- Tracked all experiments with MLflow: reward curves, epsilon decay, win rate per episode — fully reproducible training runs.

Pelvix [🔗](#)

Multimodal AI Chatbot · DeepSeek-R1 · Flux · Streamlit

- Built multimodal AI chatbot supporting text and high-quality image generation with personality-driven conversations.
- Achieved 99.9% uptime; image generation under 10 seconds for 95%+ of requests through performance optimization and efficient API usage.
- Privacy-first design: session-only storage, zero data persistence, zero tracking.