

ADITYA SARADE

+91 84510 25156 | aditya.sarade2003@gmail.com | linkedin.com/in/adityasarade | github.com/adityasarade

SUMMARY

AI Engineer focused on taking systems from prototype to production. I build multi-agent, low-latency services, and vector-search pipelines serving real users. As founding member of **Medikabazaar's AI initiative**, I shipped a **multilingual 11-language voice-AI platform** and open-sourced **AI Command Center** (an LLM gateway), **Asterix** (a persistent-memory layer for AI agents), and **OSCAR** (an agentic coding assistant) across npm + PyPI. Thriving in fast-paced environments!

EDUCATION

AISSMS INSTITUTE OF INFORMATION TECHNOLOGY

Bachelor of Technology in Artificial Intelligence and Data Science

Pune, Maharashtra

Expected May 2026

WORK EXPERIENCE

MEDIKABAZAAR

AI Engineer

Mumbai, Maharashtra

Feb 2026 – Present

The founding member of the company's AI initiative.

- Built and deployed a production **multilingual (11 Indian languages) AI voice-calling platform** (Python, LiveKit) that autonomously places and answers live phone calls, integrating real-time **ASR/STT (Sarvam Saaras)**, **TTS (Sarvam Bulbul)**, and **speech-to-speech (Google Gemini Live)** pipelines with **Gemini LLMs**, function-calling, and **RAG grounding (ChromaDB; Cohere Embed v4 on AWS Bedrock)**, plus automatic language detection.
- Delivered its **GenAI analytics layer** - LLM-based call summarization, **3-class sentiment analysis**, **BANT hot/warm/cold lead scoring**, and structured **entity extraction** - on a **FastAPI / Next.js / Docker** stack with scalable workers, **S3** call recording, and per-call cost metering.
- Built and shipped a production **FastAPI image-enhancement service** that scores SKU images on a multi-metric quality composite (sharpness, resolution, contrast, exposure) and enhances flagged assets via **Google Gemini AI**; backed by **SQLAlchemy/MySQL on AWS EC2**, **S3** uploads, **Kafka** publish to the Catalyst CMS, and **Jenkins CI/CD**, with **Postman** wired against the deployed production endpoints - now live on the **Medikabazaar marketplace**.
- Architected an end-to-end **SEO blog generation pipeline** (Python backend + React/TypeScript dashboard) that applies **unsupervised hierarchical clustering (scikit-learn AgglomerativeClustering)** over embeddings to **segment 21K Ahrefs keywords across 1K+ SKUs into 624 clusters**, then auto-generates and publishes blogs and banners via **Gemini 2.5 Pro + 3.1 Flash Image**; blogs now ship continuously to the marketplace, replacing a previously manual category-manager workflow.
- Re-engineered the KYC OCR endpoint into a **Spring Boot / Java** microservice (**Tess4J + OpenCV**) with byte-for-byte parity across 14 proof types; conduct monthly AI audits across **27 codebases**, with recommendations actioned by 4 Engineering Managers.
- Delivered **₹3L+ in direct cost savings** via Gemini Batch API, caching, and Google Cloud credits; integrated **Google Ads, Analytics, and Search Console** with Claude Code via **MCP** and drove company-wide AI-tooling adoption.

WASSERSTOFF INNOVATIONS

AI Engineer Intern

Gurugram, Haryana

June 2025 – Oct 2025

- Built **FastAPI backend** integrating **Google Gemini API & OpenAI embeddings** for legal document analysis, semantic search, and **text-to-SQL query generation on DuckDB** via ETL pipelines along with automated visualizations.
- Designed vector embedding workflows with **Qdrant and Redis caching**, reducing query latency by **~90%**; offloaded background workflows to **TaskIQ/Celery with RabbitMQ** for assessment generation and email automation.
- Developed a **multi-agent HR automation system using LangGraph and MongoDB** for candidate evaluation, automated scoring, and behavioral analysis.
- Built supporting data pipelines for web scraping, data cleaning, and database integration alongside the core HR and document-analysis services.
- Created **QMem**, a Python package for vector search that automates Qdrant collection and embedding setup via CLI.

ADGAMA DIGITAL

Data Science Intern

Pune, Maharashtra

Feb 2025 – April 2025

- Built and fine-tuned **15+ statistical and ML/DL models** across **regression, classification**, and image-based tasks on diverse datasets (~100K+ rows), with reusable preprocessing and **feature-engineering** pipelines.
- Deployed 10+ trained models into real-time web apps (**Streamlit, Gradio, FastAPI, Django**) with interactive UIs and dynamic visualizations that improved model accessibility and interpretability.

PROJECT EXPERIENCE

AI COMMAND CENTER | Zero-Dependency LLM Gateway + Self-Hosted Cost & Usage Dashboard | [GitHub](#) | [Website](#) | [npm](#) | [PyPI](#) Jul 2026

- Designed, built, and open-sourced a **zero-dependency Node.js LLM gateway** + self-hosted cost/usage dashboard - point any project at it with a one-line base-URL swap (language-agnostic) and every token, cost, and millisecond lands in one place. Proxies **10+ providers** (OpenAI, Anthropic, Gemini, Groq, Ollama, + any OpenAI-compatible) with **SSE streaming passthrough** at **<0.25 ms p50 overhead**; shipped as **3 packages across npm + PyPI** with a live **Next.js** docs site + in-browser demo.
- Delivered exact per-request **cost accounting** (multi-currency + live FX over a 2,000+ model pricing dataset; **0 mismatches / 20 cases**), a multi-view dashboard (traces, prompts, budgets, anomaly alerts), **RBAC** with per-project gateway keys, and opt-in **model tiering + prompt caching + smart routing** (projected **40-60% LLM-spend cut**); hardened via adversarial review - fixed **CORS confused-deputy, CSRF, and backpressure** bugs, with zero-dep **script + HMAC** auth. **73 tests, CI on Node 18/20/22**.
- Authored the underlying **"AI Box" platform architecture** - a 6-layer **LangGraph** multi-agent AI-ops design (orchestrator + cost, safety/compliance, anomaly, evaluation, and incident-response agents) - and pitched it with a 16-week roadmap to company leadership as a standard for AI cost, guardrails, and eval pipelines.

OSCAR | Agentic AI Coding Assistant (VS Code + CLI) | [GitHub](#) | [PyPI](#) | [Website](#) May 2026

- Built an **agentic AI coding assistant** (VS Code extension + Python CLI) on my own **Asterix** framework, running **Gemini 2.5 Flash via Vertex AI** through a **ReAct loop** with persistent session memory; exposes 15 tools across git operations, shell, **Tavily** web search, and **Playwright** browser automation through 12 **FastAPI** endpoints with **Server-Sent Events** progress streaming and a **human-in-the-loop safety layer** gating sensitive operations.
- Designed an **eval-driven development** workflow with a custom benchmark harness measuring **100% safety-classifier accuracy** and **100% tool-routing exact-match** on Gemini, alongside live runtime telemetry tracking per-tool latency, token usage, and per-prompt cost (**median 6s end-to-end latency, ~\$0.005 per prompt** on Vertex AI). Open-source under MIT.

ASTERIX | Python Library for AI Agents with Persistent Memory | [GitHub](#) | [PyPI](#) Oct 2025

- Developed and published a Python library enabling AI Agents to read/write **editable memory blocks**, persist state across sessions via **SQLite backends**, and perform semantic retrieval with **Qdrant Cloud** for long-term memory.
- Designed a decorator-driven tool system that auto-registers agent tools with input validation and self-generated documentation, supporting multiple LLMs (Gemini, Groq, OpenAI API) through a modular architecture; now powers **OSCAR** as the runtime foundation for stateful agent workflows.

QUERYPILOT | Python, LLaMA 3, Pinecone, GroqCloud, MLflow, Docker, YAML, RAG | [GitHub](#) April 2025

- Built an intelligent RAG-based SQL Copilot using **Groq-hosted LLM**, sentence-transformers, and **Pinecone** (after testing FAISS and ChromaDB), achieving real-time autocompletion inside MySQL Workbench and PostgreSQL.
- Implemented an **MLOps pipeline with MLflow** tracking acceptance metrics (**95%+ acceptance across 20+ sessions**), with session memory caching and YAML-driven configuration for reproducible, hypothesis-driven experimentation.
- Dockerized the entire project with system dependencies, clipboard hooks, and preloaded models, **reducing latency by over 50%** and enabling scalable deployment and CI/CD through GitHub Actions.

CERTIFICATIONS

- Large Language Model Agents | UC Berkeley
- LLMs as OS: Agent Memory | DeepLearning.ai
- Neural Networks & Deep Learning | Coursera, DeepLearning.ai
- Career Essentials in Generative AI | Microsoft, LinkedIn

TECHNICAL SKILLS & LEADERSHIP

Languages: Python, SQL, Java, JavaScript (React, Next.js), C/C++, HTML/CSS

ML/AI: Scikit-learn, TensorFlow, PyTorch, XGBoost, OpenCV, NLTK, SpaCy, Vector Embeddings, LangChain, LangGraph, MCP, LiteLLM, NeMo Guardrails

Tools: Docker, Jenkins, MLflow, FastAPI, Spring Boot, Flask, Django, Streamlit, Gradio, GitHub Actions, Kubernetes, Kafka, RabbitMQ, OpenTelemetry, Postman, CI/CD pipelines

Data Infrastructure & Platforms: MySQL, MongoDB, DuckDB, SQLite, Redis, Qdrant, ChromaDB, FAISS, Pinecone, AWS (EC2, S3, Vertex AI, Bedrock), Hugging Face

Leadership: Joint Secretary (IOIT ACM), Deputy Secretary General (IOIT MUN), and Campus Manager (Bullsmart)