ADARSH RAO

+91 9845402182 ◇ Bengaluru, Karnataka
adrao18020@gmail.com ◇ linkedin ◇ Github ◇ Portfolio

EXPERIENCE

Software Development Engineer 2 Mar 2026 - Current
Mastercard *Pune, India*

- Worked on a **production RAG chatbot** for report search, with **generative UI via entity matching**: report IDs in answers are matched and rendered as different cards based on the user's **subscription state**.
- Prototyped an **on-device AI search agent** POC on Chrome's built-in LLM with **zero server round-trip**, implementing **structured tool calling** on a platform with no native function-calling: a **JSON-Schema constrained** intent extractor resolves natural language into a typed query, invokes the **real backend API**, and feeds results back for summarization; also exposed a **WebMCP** tool schema.
- Built and demoed a series of internal **LLM agents**, including a **defect-solver bot**, a **feature-implementation bot**, and **automated testing agents**, driving adoption of AI tooling into the team's development workflow.
- Quantified a **69-81% reduction in logging tail latency** (p99, p99.9 at 8 threads) with a **JMH** benchmark harness, then delivered a **risk-assessed hybrid design** backed by crash-loss simulation and a phased canary rollout adopted into follow-up work.
- Reduced load time by **75%** (**6s** to **1.5s** LCP) on a reporting platform serving **20,000 users** via lazy loading, bundle optimization, and API call reduction, and code-split a reporting tool into a standalone **micro-frontend** for **500+ users** with its own **Jenkins CI/CD** pipelines.
- Part of the move of the reporting platform's services to **microservices on AWS** under **PCI-DSS** constraints (nothing stored unencrypted at rest), keeping live services running.

Software Development Engineer 1 Jun 2024 - Mar 2026
Mastercard *Pune, India*

- Built a **text-to-SQL and RAG** proof-of-concept over an enterprise reporting catalog of **400+ reports** on a **self-hosted LLM** with **pgvector** semantic search, engineering the retrieval and prompting pipeline from natural language to SQL to summarized result; demoed to **engineering directors**.
- Designed and shipped a full-stack admin tool in **Java**, **Spring Boot**, and **Oracle SQL** with schema design, **REST APIs**, **4-role RBAC**, and a multi-stage approval workflow with rollback and audit trail, replacing a quarterly release cycle with on-demand self-service.

EDUCATION

B.Tech, Computer Science and Engineering, National Institute of Technology Warangal 2020 - 2024

PROJECTS

Remi - Open-Source LLM Observability Platform (Java 21, Spring Boot, ClickHouse, PostgreSQL, React, Docker, LangGraph) (Github) (Website) Solo-built **OpenTelemetry-native** tracing and evaluation platform for LLM agents, in the same problem space as **LangSmith and Langfuse**. Ingests OTLP traces from **LangChain/LangGraph** through an authenticated **multi-tenant** proxy into **ClickHouse**, serving flame-chart timelines, per-release regression comparison, and an **LLM-as-a-judge evaluation pipeline**, with **zero-SDK** integration. Cut aggregate query latency **234ms to 26ms** on **560K rows**.

Relay Memory - Graph-Memory Support Agent (Python, LangGraph, Neo4j, Cypher, OpenRouter) (Github) Built a customer-support agent for a **Neo4j hackathon** that remembers returning customers across sessions. A **12-node LangGraph** state machine keeps **four memory layers** in Neo4j (short-term, long-term, reasoning traces, **vector-indexed semantic memory**) over **8,469 tickets** and **8,320 customers**. A **policy controller** turns each message into a bounded structured decision (ticket action, response mode, next step), so the LLM only writes the reply and never mutates data. Added a **confidence gate** that asks one focused clarifying question instead of guessing, and persists an **audit trace** before reasoning so every turn is reviewable.

umarag - Client-Side RAG Chatbot SDK (TypeScript, WebGPU, WebLLM, transformers.js, Orama) Building an open-source SDK where **retrieval and generation both run in the user's browser** with zero backend calls at query time. Does client-side vector search over a prebuilt knowledge export, **auto-selects between WebLLM (WebGPU) and Chrome's on-device Prompt API** with fallback and live engine hot-swap, bounds multi-turn memory for small context windows, and fails with a typed error rather than falling back to slow CPU inference. **Vitest** suite covers engine selection, retrieval, and prompt building.

Amma App - Accessibility-First Language Learning App (Flutter, Firebase) (The Hindu) Built a flashcard learning flow with **image recognition** and **speech feedback** for children with hearing difficulties, developed with speech therapists and tested with **10 children** across **10+ Indian languages**.

ACHIEVEMENTS

- Placed **3rd Pan-India** among all interns in **Mastercard's Innovation Challenge** (2023 internship) for a real-time payment fraud detection build.
- **Technical Team Lead**, Centre for Innovation and Incubation: ran workshops for **1800+ students**; built **Voter Buddy**, a WhatsApp chatbot for **500 users** across **Hanumakonda and Warangal** (Lok Sabha 2024 campaign).