

Rishav Jha

Bengaluru, India | rishav.nitdgp24@gmail.com | 7439876937 | rishavjha.com | linkedin.com/in/rishav-devraj
github.com/rishav-jha-mech

Summary

AI engineer building LLM products end to end: **RAG** pipelines, **MCP** integrations and streaming AI interfaces. Built **Almond AI**, a shared memory layer for AI tools, and won Tessell's hackathon with **Fetch**, a RAG knowledge engine later adopted internally.

Education

National Institute of Technology Durgapur, B.Tech. in Mechanical Engineering

Dec 2020 – May 2024

Projects

Almond AI

almondai.tech

- Built **Almond AI** with a small team: a shared, multiplayer **memory layer for AI tools** where a team records decisions once and every connected tool reads the same searchable memory; in private beta
- Built native integrations for **Claude Code**, **Cursor**, **Figma** and **Figma Make**, plus any **Model Context Protocol (MCP)** compatible tool, so AI assistants read and write shared context instead of resetting it
- Worked across the stack: a **Python** backend with **Neo4j**, a **React** and **TypeScript** frontend, and **Docker** packaging for self-hosted, on-premises deployment

Squid Game

[GitHub](#)

- Built a Squid Game themed event site for 300+ participants: a **Three.js** Green Light Red Light round using **Pico.js** face detection to catch players who moved, served from **AWS EC2** with **Gunicorn**, **Nginx** and **Supervisor** over SSL
- Tools:** Django, Three.js, Pico.js, AWS EC2, Nginx, Gunicorn

VertiView

[Chrome Web Store](#)

- Built a **Chrome extension** that turns any video on a page into a zoomable, pannable 9:16 portrait panel so several videos can run side by side; wrote the zoom, pan-bounds and crop layout engine with **unit tests**, plus a **Next.js** marketing site
- Tools:** TypeScript, Chrome Extension APIs, Vite, Next.js

Experience

Tessell, Member of Technical Staff 2 (Software Engineer) – Bengaluru, IN

Aug 2025 – Present

- Won Tessell's company-wide hackathon** within two weeks of joining by building **Fetch**, a **Python RAG** knowledge engine over company Slack history and documents that answers natural-language questions with cited source permalinks; it continued as an internal tool after the hackathon
- Architected Fetch's retrieval pipeline end to end: paginated **Slack API** ingestion into normalized **SQLite** storage, batch sync into **ChromaDB** with local **sentence-transformers** embeddings, **LLM query rewriting** and dual-collection retrieval that routes ownership questions to a people index
- Extended Fetch ingestion to PDF, DOCX, CSV and Markdown attachments and shipped it as a **Slack Socket Mode** bot with threaded answers and **citation-grounded generation** traceable to the source conversation
- Designed **streamUiQuery**, a centralized **Server-Sent Events (SSE)** streaming layer with a typed event contract, and migrated **six AI chat surfaces** onto it, replacing duplicated per-surface stream handling
- Built the **generative-UI rendering engine** that turns AI-streamed node contracts into **React UI**, then merged it into the design system, collapsing two lockstep-released packages into one and removing version-drift bugs
- Engineered a **two-lane concurrent execution engine in Rust** for chat and background AI workloads with bounded concurrency, **MPMC** channel dispatch and hardware-aware parallelism, plus **SQLite**-persisted execution state with automatic crash recovery of orphaned jobs
- Implemented end-to-end **observability** with request-ID correlation across **TypeScript** and **Rust** into **AWS CloudWatch**, and hardened the **Tauri** security surface with compile-time **IPC** gating and capability restrictions

Cuemath, Software Engineer (Intern to Full-Time) – Bengaluru, IN

Oct 2023 – Aug 2025

- Refactored the 1v1 multiplayer gameplay layer built on **Colyseus**, replacing **setTimeout** workarounds and 10 to 20

events per room state mutation with a single event-driven update, cutting new 1v1 game delivery from 1 to 2 months to **14 days** (~75% faster)

- Implemented player reconnection and AFK handling on **Colyseus** rooms: a 30-second reconnection window, round-level AFK forfeiture and automatic room teardown for bot matches, reducing mid-game drops and orphaned rooms
- Built the **Journey** coachmark and guided-onboarding system with Principal Architects: 8-directional collision detection, absolute and scrollable positioning, gesture-nudge and card formats, and full **Storybook** edge-case coverage, adopted across Circle and Leap

Google Summer of Code (GSoC), Contributor and Mentor – gsof.rishavjha.com

May 2023 – 2025

- Redesigned the Talawa Admin UI for the Palisadoes Foundation in **React** and **TypeScript** from a new Figma design as a responsive, modular interface, deployed incrementally screen by screen as reusable, documented components
- Refactored the Talawa website codebase and added **test coverage**, resolving **cross-browser compatibility** issues and improving long-term maintainability
- Returned as a **GSoC mentor** for the Palisadoes Foundation in the **2024 and 2025** cohorts, guiding contributors

Capricon Technology, React Native and Flutter Developer

Jun 2022 – May 2024

- Built **Would You**, a **Flutter** dating app where on-device smile detection with **Firebase ML** drives swipes, including auth, profile and camera flows
- Built **Treklers**, a **React Native** and **TypeScript** taxi booking app with **50+ screens**: real-time driver tracking, scheduled rides, emergency calling and cash, wallet and card payments; separating UI from business logic let four full redesigns ship without regressions

Skills

Languages: Python, TypeScript, JavaScript, Rust, SQL

AI & LLM: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Model Context Protocol (MCP), Prompt Engineering, Query Rewriting, Embeddings, sentence-transformers, Semantic Search, Azure OpenAI, AI Agents, Generative UI, Computer Vision (Firebase ML, Pico.js face detection)

Data & Infra: ChromaDB (Vector Database), Neo4j, SQLite, Docker, AWS (CloudWatch, S3), Slack API, GitHub Actions

Frameworks: FastAPI, Flask, Django, React, Tauri

Concepts: LLM Streaming (SSE), Citation-Grounded Generation, Shared Memory Systems, Concurrency, Observability