

RISHI RAJ

+91 7909050035 | rishiraj438gt@gmail.com | linkedin.com/in/rishi-raj-3488432ab
github.com/rishiraj38 | rishiraj38.github.io | leetcode.com/u/rishiraj438

EDUCATION

Newton School of Technology, Rishihood University

B.Tech in Computer Science & Artificial Intelligence

Aug 2024 – Jun 2028

Third-Year Student

Sainik School Tilaiya

Class 6 – Class 12

Aug 2017 – Mar 2024

EXPERIENCE

Maintainer & LFX Mentee – Meshery (CNCF)

Feb 2026 – Oct 2026

Maintainer across Meshery repos; LFX Mentee: Models Support for OCI Registries (2026 Term 2)

Remote

- Authored **100+** merged PRs; reviewed **300+** and merged **100+** as a maintainer across meshery, meshery/schemas and sistent, keeping shared interfaces consistent across repos
- Built OCI-registry support for Meshery's model registry, implementing **ORAS**-based push/pull workflows so models ship as standard OCI artifacts (manifests, blobs, digests, media types)
- Led the permission-shield migration: centralized RBAC permission checks into the shared Sistent library, so Meshery UI, Meshery Cloud and Kanvas pass only a permission key and action
- Fixed a recurring Registry-page crash spanning multiple dependent repos; onboarded new contributors, triaged issues and improved contributor docs

PROJECTS

Quiver – Vector Database from Scratch | *Go, HNSW, ARM64 NEON, WAL, HTTP API*

[Live](#) | [GitHub](#)

- Built a vector database in Go with zero third-party dependencies (8.7k lines): HNSW index, write-ahead log, snapshots, filtered search, HTTP API and a web UI
- Reached **0.99 recall@10 at 3,849 queries/sec per core** on the SIFT1M benchmark (1M vectors), about 75% of Meta's FAISS and 20% faster than hnswlib on the same hardware
- Wrote ARM64 NEON assembly distance kernels, **2.5x faster** than the compiled Go version
- Implemented group commit in the write-ahead log, raising durable write throughput **35x** (420 to 14,900 writes/sec); verified zero acknowledged writes lost across repeated kill -9 crashes
- Deployed a live semantic search demo over 20,000 news articles, with in-browser query embedding and sub-2 ms results

HELIX – Autonomous AI News Pipeline | *Python, LLMs (Grok), PostgreSQL, Next.js, Prisma*

[Live](#) | [GitHub](#)

- Ingests ~400 articles/day from 13 sources, summarizes them with LLMs and ranks them per subscriber, running unattended on a GitHub Actions cron
- Sends each subscriber a personalized daily email, auto-posts top stories as an Instagram carousel, and serves a Next.js dashboard with Clerk auth and Prisma
- Cut query time **8.5s → 3.7s** by replacing a ~32k-row in-memory scan with a SQL anti-join and indexes; added 106 offline tests, CI and alerts for zero-delivery runs
- Traced a silent production outage to 3 root causes, restoring ingest from 5 to ~420 articles/run; enforced Free/Pro/Admin access server-side, closing a flaw that exposed subscriber emails

SMART EMERGENCY – RL Dispatch Environment | *Python, TRL, QLoRA, FastAPI, Docker*

[Live](#) | [GitHub](#)

- Co-built an OpenEnv RL environment where an LLM agent triages and dispatches 911 calls on procedural city graphs, with a 5-part shaped reward that penalizes exploits like hallucinated unit IDs
- Fine-tuned Qwen3-1.7B with QLoRA via SFT (loss **2.32 → 0.31**), then GRPO against environment rewards
- Built the FastAPI grader and baseline API; deployed to Hugging Face Spaces via Docker and GitHub Actions

TECHNICAL SKILLS

Languages: Python, Go, TypeScript, JavaScript, SQL

AI/ML: LLM Agents, LangGraph, Hugging Face (TRL, Unsloth), QLoRA, SFT, GRPO, Vector Search (HNSW, FAISS)

Backend & DB: FastAPI, Node.js, Express.js, Prisma, PostgreSQL, MongoDB, MySQL

Cloud & DevOps: Docker, Kubernetes, GitHub Actions, AWS, Vercel, OCI Registries / ORAS

Frontend: React.js, Next.js, Tailwind CSS

ACHIEVEMENTS

- Finalist, Meta × Hugging Face OpenEnv Hackathon
- CNCF LFX Mentee (2026 Term 2); Certified Meshery Contributor