

JYOTIRADITYA SINGH

+91 8791978912 | jyotiraditya.singh.tech@gmail.com | jyotiradityasingh.vercel.app
github.com/jyotir07 | linkedin.com/in/jyotir07 | leetcode.com/jyotir07

OPEN-SOURCE & RESEARCH ENGINEERING

- Loom** — Unified Multi-Provider AI Routing Framework | *Python, Redis, PostgreSQL, LLM Provider SDKs* loom-weaves.vercel.app
- Shipped **loom-router** to PyPI — an open-source Python framework collapsing **14+ AI providers** (OpenAI, Anthropic, Gemini) behind one **generate()** contract, so a model swap is a config change, not a rewrite.
 - Engineered the layer that makes it worth adopting: response caching, cheap-first model routing, batch APIs, and cross-vendor failover that survives a provider outage mid-request.
 - Adopted as the standard LLM/TTS/image integration layer across **11 projects** at Prometrix AI — one dependency replacing per-project SDK glue.
- Technical Writing** blog
- Building a RAG Pipeline*, *RabbitMQ Unpacked*, and *Building a Real-Time Collaborative Editor* — long-form write-ups dissecting system design and the trade-offs behind each implementation choice.

EXPERIENCE

Prometrix AI Aug 2025 - July 2026
AI Developer Intern

- Drove domain-specific retrieval up **9% Recall@5** by benchmarking cosine, dot-product, and L2 similarity against Manhattan (L1) distance, then adopting the winning metric.
- Built AIVA, a generative-engine-optimization (GEO) platform that tracks and improves how often brands surface in AI-assistant answers to real buyer queries.
- Authored Loom (above), adopted company-wide as the standard LLM/TTS/image integration layer across **11 projects**.
- Shipped AI Podcast Studio, an LLM+TTS pipeline generating multi-speaker podcasts with streaming output and speaker orchestration that assigns personas dynamically across a cast.
- Cut dashboard load time **from ~1s to ~400ms** via Redis caching and query tuning.

SYSTEMS PROJECTS

DAWA — Vernacular Medication Voice Agent | *Python, FastAPI, Sarvam AI (STT/TTS)* Team of 4 — owned the voice layer

- Built the patient voice layer: a Hindi dose-time call running Sarvam Saaras STT into a policy-bound 5-state dialogue and back out through Bulbul TTS, verified live end-to-end.
- Designed the agent so it *cannot* give medical advice: every spoken line is a fixed template with the drug name injected from the med graph, so the model never writes words the patient hears.
- Kept the safety path deterministic — Latin/Devanagari regex decides adherence; the LLM refines only unknowns and can raise but never clear a refusal flag, so a dead LLM still completes the call.

Pingr — Distributed Multi-Region Uptime Monitor | *Go, WebSockets, PostgreSQL, Redis, Docker, Fly.io* GitHub

- Engineered a Go monitor probing any target from **three continents at once** — containerized US/EU/Asia Fly.io workers run HTTP/TCP/DNS checks with full **httptrace** timing (DNS, TCP, TLS, TTFB).
- Threat-modeled the abuse surface first — scheme allow-listing, 11 blocked private/metadata CIDR ranges, and DNS-rebinding defense via a dialer that revalidates the dialed IP, not the hostname.
- Fanned out probes with **errgroup**, streaming per-region results over WebSockets as they land; Redis rate-limits per IP while PostgreSQL rollups serve 24h/7d/30d uptime and p50/p95/p99 latency.

OrderMesh — Distributed Order Orchestration | *FastAPI, PostgreSQL, Redis, RabbitMQ, K8s* ordermesh.vercel.app

- Orchestrated inventory, payment, and delivery services over RabbitMQ with retry-safe message handling, so a failed step replays instead of corrupting an order.
- Cut API latency **35%** by moving hot reads to Redis and pushing slow work into rate-limited background jobs.

TECHNICAL SKILLS

Languages: Python (primary), Go, Java, JavaScript, C++, SQL, Bash

AI & LLM Systems: RAG, AI agents, voice agents (STT/TTS), vector databases, retrieval eval, OpenAI / Anthropic / Gemini APIs

Backend & APIs: FastAPI, REST APIs, WebSockets, PostgreSQL, Redis, RabbitMQ, MongoDB, Webhooks

Infrastructure: Docker, Kubernetes, Linux/UNIX, CI/CD, Nginx, Git/GitHub

EDUCATION

University of Petroleum and Energy Studies (UPES) Dehradun, India
B.Tech in Computer Science Engineering Aug 2022 – May 2026

ACHIEVEMENTS

Smart India Hackathon (SIH) 2023 — National Finalist | WebCrux 2024 Web3 Hackathon — Winner | CSI Hackathon — Runner-up (solo)