

Milan Soni

AI Engineer | Backend & Full Stack Developer

+91-9694677393 | milansoni96946@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#) | Churu, Rajasthan

SUMMARY

AI Engineer and Full Stack Developer who builds production RAG pipelines, multi-agent LLM systems, and scalable backends in Python and TypeScript. SIH 2023 National Winner (top 1% of 44,000+ teams) and Scopus-indexed researcher, with 3 internships and merged open-source contributions to a 10.8k-star AI gateway. Strong in data structures, algorithms, REST API design, and shipping reliable systems on free-tier infrastructure.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript (ES6+), C++, SQL
AI / ML: RAG Pipelines, Multi-Agent Orchestration, LangChain, LangGraph, Hybrid Search (Vector + BM25), Cross-Encoder Reranking, Prompt Engineering, LLM APIs (OpenAI, Anthropic, Gemini), Vector Embeddings, BERT, scikit-learn
Backend: FastAPI, Node.js, Express.js, REST APIs, GraphQL, SQLAlchemy, Pydantic, JWT/Clerk Auth, RBAC, Microservices
Frontend: React 19, Next.js, TanStack, Redux, Tailwind CSS, Framer Motion
Databases: PostgreSQL, pgvector, Redis (HNSW), MongoDB, Supabase, Prisma, FAISS
Cloud / DevOps: Docker, GitHub Actions CI/CD, Vercel, HuggingFace Spaces, AWS, Linux, Git, Nginx

EXPERIENCE

OmniRoute — Open-Source AI Gateway (10.8k stars) Jul 2026 – Present
Open Source Contributor · Remote

- Diagnosed and fixed an HTTP 400 regression affecting strict LLM providers; proposed a declarative Zod schema flag adopted by the maintainer into the shipped fix, with 55/55 passing tests (PR #6225).
- Shipped 5+ merged PRs across a 230+ provider codebase — a Claude provider registry integration, an accessible provider-rankings filter (168 additions, 9/9 tests), and docs normalization across 42 locales.

nTheta Works Pvt. Ltd. Oct 2025 – Dec 2025
Full Stack Developer Intern · Remote

- Built a two-stage semantic retrieval pipeline (Ollama embeddings → Redis HNSW → FlashRank reranking) for an enterprise LLM API-testing platform, improving template-matching accuracy by 40% and cutting manual QA effort by 60%.
- Shipped async FastAPI microservices and Next.js/TypeScript dashboards, containerized with Docker Compose and CI/CD on Linux.

OBG Outsourcing Pvt. Ltd. May 2025 – Jul 2025
Full Stack Developer Intern · Jaipur

- Led FinSageAI360, a multi-tenant financial-intelligence SaaS, cutting monthly close-reporting time by 45% and manual effort by 30% via AI anomaly detection and real-time KPI dashboards.
- Designed a JWT-authenticated REST API (Node.js/Express/MongoDB) with granular RBAC for multi-tenant data isolation.

Om Logistics Ltd. Jun 2024 – Aug 2024
Software Developer Intern · Delhi

- Optimized enterprise document search with LangChain + FAISS embeddings, cutting query latency by 70% and improving retrieval accuracy by 40% across 10,000+ documents.

PROJECTS

MiningNiti · Multi-Agent RAG Document Intelligence Platform | *SIH 2023 National Winner* | [GitHub](#) · [Live](#) Dec 2023

- Engineered a 6-agent AI pipeline (classifier, safety analyzer, entity extractor, summarizer, compliance auditor, RAG chat) across 4 LLM providers, cutting end-to-end latency from 14s to 3.2s and reaching 98.7% pipeline completion.
- Built production hybrid RAG (pgvector + pg_trgm via Reciprocal Rank Fusion, cross-encoder reranking) that reduced manual compliance analysis by 90%, deployed on \$0/month free-tier infrastructure.

NLPForge · AI NLP Dataset Generator & Semantic Search Platform | [GitHub](#) 2025

- Built a full-stack RAG platform (async FastAPI, Next.js/TypeScript) converting natural-language queries into executable API test cases via a two-stage Redis HNSW → FlashRank reranking pipeline.
- Integrated 8 LLM providers and 15+ embedding models via a factory pattern; Dockerized microservices with GitHub Actions CI/CD and 70 backend unit tests.

ACHIEVEMENTS & CREDENTIALS

- SIH 2023 National Winner** — Ministry of Coal; top 1% of 44,000+ teams, recognized by Coal India Limited & CMPDI.
- Scopus-Indexed Publication** — PICET-2026 (IET): hybrid attention-based temporal modeling.
- Open-Source** — 5+ PRs merged into OmniRoute (10.8k stars). **NASSCOM Certified Full Stack Developer (2024)**. **1st Place**, Jigyasa Tech Event (GIT). **Hackathon Organizer**, CodeFiesta (2023–2025).

EDUCATION

Global Institute of Technology, Jaipur · B.Tech, Computer Science & Engineering — CGPA: 8.10/10Oct 2022 – May 2026