

SIDDHANT RAI

siddhant.rai.5686@gmail.com

s1d.is-a.dev | github.com/siiddhant | linkedin.com/in/siddhant-rai21

+91 73170 27719

India

EXPERIENCE

DocsGPT -- Arc53

UK -- Remote

Software Engineer, Core Team / Python, TypeScript, LLM Infrastructure

Apr 2024 -- May 2026

- Architected routing services for LLM workloads that supported 50% user growth while cutting P99 latency from 8.2s to 5.1s.
- Improved Qdrant retrieval concurrency 5× and reduced query latency by 1.3s through caching and optimized data access.
- Built an agent workflow builder and CEL-based execution engine with conditional branches, state updates, graph validation, and full-stack configuration.
- Standardized streaming and non-streaming LLM execution with typed response/tool-call contracts and provider fallback; developed ReAct, retrieval, API-tool, and tool-call tracing workflows.
- Integrated remote Model Context Protocol (MCP) clients across OAuth and SSE transports; introduced CI test coverage that reduced reported production regressions by 40%.

PropReturns (YC S21)

Mumbai, India

Full Stack Engineer Intern / Next.js, React, AWS

Apr 2023 -- Jun 2023

- Built a Bit.dev UI library used across three applications, accelerating feature delivery by 40%; reduced Largest Contentful Paint by 400ms using code splitting and CloudFront caching.
- Added visual regression checks to the deployment pipeline, catching 90% of UI defects before production.

Pinch Finance

India -- Remote

Full Stack Engineer Intern / Node.js, React Native, PostgreSQL

Jun 2022 -- Nov 2022

- Built 20+ REST APIs handling 10,000+ daily requests at 99.5% reliability; reduced critical SQL query time by 300ms through indexing.

PROJECTS

MarginTrace | Python, OpenTelemetry, SigNoz, SQLite

2026

- Designed an outcome-aware control plane for AI workflows that correlates model/tool traces with exact provider cost, verified business outcomes, waste, and margin.
- Implemented retry-aware attribution, durable economic ledgers, dashboards, alerts, and pre-spend allow/degrade/deny policies; validated success, semantic failure, and zero-cost denial against real provider flows.

ReefWatch | Python, FastAPI, React, Coral, SQLite

2026

- Built an evidence-first incident investigation system that discovers schemas at runtime and correlates read-only SQL evidence across seven operational and security sources with SSE streaming and persisted runs.
- Designed query budgets, duplicate suppression, context compression, evidence-first retries, source-coverage checks, and synthesis gates to constrain autonomous investigations.

ENGINEERING CONTRIBUTIONS & ACHIEVEMENTS

- Coral / WeMakeDevs:** Selected as a PR Track Winner in the Pirates of the Coral-bean Hackathon after shipping five merged data sources exposing 31 queryable tables across Google Drive, Google Docs, Search Console, Netflix, and OSV.
- OLake:** Extended Go database connectivity across DB2, MSSQL, Oracle, and MySQL with SSH tunneling, sync-progress tracking, JDBC parameters, and SSL/TLS; validated tunnel authentication through a Docker bastion plus 20 tests.
- OpenFGA:** Fixed a concurrency defect in request cloning and added concurrent regression coverage to the Go authorization server.
- mcp-use:** Added configurable base URLs across the inspector UI and server, enabling OpenAI-compatible providers including LM Studio, OpenRouter, and Ollama.
- Competitive Programming:** LeetCode Knight, 2040+ rating (Top 1% globally).

TECHNICAL SKILLS

Languages: Python, Go, TypeScript/JavaScript, SQL, C/C++

Backend & Web: FastAPI, Flask, Django, Node.js, React, Next.js, REST APIs, SSE, MCP

Data & Infrastructure: PostgreSQL, MySQL, MongoDB, Redis, Qdrant, SQLite, AWS, Azure, Docker, Linux, GitHub Actions, CI/CD

Architecture & AI: Distributed systems, system design, microservices, API design, OAuth 2.0/2.1, OpenTelemetry, SigNoz, LLM/RAG systems

EDUCATION

PDPM Indian Institute of Information Technology, Design and Manufacturing

Jabalpur, India

B.Tech, Electronics and Communication Engineering | CPI: 8.5/10

Nov 2020 -- May 2024