

C S Sahil

📞 +91 9074879639

✉️ sahilcs1111@gmail.com

🌐 [Linkedin](#)

🐙 [Github](#)

Experience

Siemba

Software Development Engineer

Feb 2023 – Present

Bangalore, India

- **Architected an end-to-end AI system** on ECS Fargate and AWS Batch with workload-aware auto-scaling, **cutting infra cost by 50–80%** and holding sub-second latency through cost and latency tuning of model workloads.
- **Deployed Sift (AI review agent) into the production AIDLC pipeline** with risk-based routing, **cutting review turnaround by 50%** and owning its reliability across every pull request.
- **Decomposed a monolith into microservices** via service managers and abstract interfaces, running isolated Docker-in-Docker scan jobs per provider with retry-safe, re-runnable execution.
- Built **RAG pipelines with vector search and Bedrock Agents**, grounding output in retrieved CVE and codebase context and measuring quality to reduce hallucinated severity calls.
- **Engineered low-latency streaming with Redis and WebSockets** for live scan status and result delivery, instrumented for run-level visibility.

Gravystack

Software Development Engineer Intern

Mar 2022 – Feb 2023

Phoenix, Arizona, United States

- Built **scalable NestJS (TypeScript) APIs** integrating third-party SaaS services over DynamoDB, using PartiQL and tuned access patterns to **cut query latency by 40%**.

Projects

Sift - AI-Powered PR Review Agent | Python, FastAPI, LiteLLM, PostgreSQL, pgvector, Semgrep, GitHub API

- Built an **open-source FastAPI agent for automated code review** over GitHub webhooks and Actions, with a multi-pass pipeline (per-file, critic, holistic) and a configurable primary/review model split for **prompt and cost tuning**.
- Designed a **risk-based routing and merge-gate layer** over path heuristics, diff complexity, and historical feedback, with content-hash caching to cut redundant LLM calls and **reduce per-review latency and cost**.
- Built a **feedback-driven eval loop and RAG pipeline** from reactions and slash commands, computing quality scores that gate future routing, with tree-sitter AST extraction and **pgvector HNSW embeddings** for retrieval.

Collaborative Text Editor | React, Firebase, CRDTs, Tailwind CSS

- Designed a **CRDT-based collaborative editor** on Firebase, holding eventual consistency and distributed state across clients for real-time multi-user editing with no central merge authority.
- Built a custom **Fractional Indexing algorithm** and multi-field synchronization layer, supporting efficient arbitrary insertions/deletions ($O(\log N)$) and granular updates for Title, Content, and Tags.
- **Engineered Undo/Redo, batching, and real-time presence systems**, preserving concurrent edits, low-latency cursor tracking, and session management across multiple collaborators.

Technical Skills

Languages: Python, TypeScript, Go, SQL, Java, JavaScript

Technologies: FastAPI, LiteLLM, RAG (chunking, embeddings, reranking), pgvector / vector search, Bedrock Agents, PostgreSQL, Redis, MySQL, MongoDB, OAuth / third-party SaaS APIs, Docker, Kubernetes, observability tooling

Education

Vellore Institute of Technology

Bachelor of Technology in Computer Science Engineering (CGPA: 8.64 / 10)

Vellore, Tamil Nadu

Certifications

AWS Certified Cloud Practitioner | Supervised Machine Learning: Regression & Classification | Neural Networks & Deep Learning