

ANIRUDH SINGH

+91 7014319880 | anirudhsing308@gmail.com | [LinkedIn](#) | [GitHub](#) | Jaipur

SUMMARY

Software Engineering Intern candidate with proven experience building **scalable, distributed systems** (10,000+ users) across full-stack and AI-integrated software. Skilled in microservices architecture, caching, RAG pipelines, real-time services, and system design with strong foundations in **data structures & algorithms**. Seeking to contribute to large-scale infrastructure challenges as a **Software Engineering Intern**.

EDUCATION

Vellore Institute of Technology, Chennai · B.Tech Computer Science & Engineering · CGPA: **8.9 / 10**

Aug 2023 – Present

BSER – Senior Secondary (12th) · 82% · Jaipur, Rajasthan

Jul 2022

WORK EXPERIENCE

Software Engineer Intern – Project Developer · *Taiyarri24*

Sept 2025 – Present

- Owned end-to-end **distributed system design** for an edtech platform serving 1,200+ concurrent users: designed **multi-tier architecture (API gateway → service layer → DB)** with **L7 load balancing**, **horizontal pod autoscaling**, and **Redis write-through cache** on hot-path reads, cutting DB load by ~45%.
- Engineered a **custom DSA IDE** with **sandboxed multi-language code execution (Docker isolation)**, **async job queue (Celery)** for test-case evaluation, and **WebSocket-based real-time output streaming** — Analysed to process 500+ concurrent submissions without head-of-line blocking.
- Designed **LLM inference pipeline with request batching**, **response streaming (SSE)**, and **vector store-backed context retrieval (RAG)**; introduced **per-user token-bucket rate limiting** at the API gateway and **idempotency keys on payment webhooks** to ensure exactly-once billing under retries.

PROJECTS

SAHAYAK– Multilingual Voice Translator · *Python · PyTorch · Whisper · FastAPI · Redis · Celery · PostgreSQL*

A real-time healthcare communication system that translates spoken language across 10+ languages using a voice-in, voice-out pipeline — designed for low-resource medical settings where language barriers delay diagnosis.

- Developed a **3-tier stateless microservice pipeline** (STT → LLM → TTS) behind a FastAPI gateway with per-stage **horizontal scaling and fault isolation**; async task queue (Celery + Redis broker) with **priority lanes** for urgent requests and **TTL-based result caching** cutting redundant LLM calls ~40%, p99 latency 4s → 2.4s.
- Enforced **idempotency via request-ID hashing in Redis** to prevent duplicate processing under retries; implemented **circuit-breaker back-pressure** via queue depth monitoring and **exponential backoff** on downstream failures; persisted transcripts with **indexed composite keys (user_id, session_id, timestamp)** for O(log n) audit-trail range scans.

TaiyarriHire – Interview Prep & Job Matching Platform · *Python · FastAPI · PostgreSQL · pgvector · Redis · LLMs · RAG*

A full-stack platform that matches candidates to job listings using semantic search and prepares them for interviews via AI-driven mock sessions with personalized, feedback-driven question adaptation.

- Architected **RAG pipeline with pgvector + HNSW indexing** (512-token overlapping chunks, sentence-transformers) for sub-100ms **ANN retrieval** at 50k+ docs; decoupled ingest → embed → query stages via **async message passing** to eliminate head-of-line blocking.
- Applied **write-through caching with event-driven invalidation** on user profiles (60% read latency drop on hot paths); enforced **token-bucket rate limiting at API gateway** and **multi-tenant row-level data isolation with PgBouncer connection pooling** maintaining sub-200ms p95 response time under load.

CognIQ – Adaptive Cognitive Assessment Engine · *Python · FastAPI · PostgreSQL · Scikit-learn · WebSocket*

An online assessment engine that dynamically adjusts question difficulty in real time based on a test-taker's ongoing performance, delivering precise cognitive profiling for educational and hiring use cases.

- Built **adaptive testing engine (IRT)** as a stateless scoring service with Redis session state; modelled sessions as an **event-sourced state machine (append-only log)** for exactly-once scoring on reconnect; exposed **WebSocket endpoint with graceful degradation to HTTP long-polling fallback** on connection loss.
- Implemented write-through caching with event-driven cache invalidation for user profiles, reducing hot-path read latency by 60%; optimized PostgreSQL scalability through PgBouncer connection pooling and row-level tenant isolation, sustaining sub-200ms p95 latency under peak traffic.

Codski – Developer Collaboration Platform · *React · FastAPI · PostgreSQL · Redis · LLMs · GitHub API · Node.js · WebSocket*

A developer tooling platform that enables teams to perform AI-powered code reviews, visualize codebase dependency maps, and get DevOps guidance — all connected to their local repo via a published npm CLI.

- Implemented **microservices backend (Code Review, DevOps AI, Repo Map)** with independent data stores and async event bus; modelled repo dependency trees as a **DAG with adjacency-list storage in PostgreSQL** and **DFS-based dependency resolution** for Repo Map graph traversals.
- Published **CLI npm package (codski-link)** that authenticates via **JWT handshake**, traverses the local filesystem to build a **file-dependency graph**, serialises it as a **zlib-compressed JSON payload**, and streams it over **chunked HTTP upload** — enabling remote AI code review without manual file uploads.
- Implemented code review as a **parallel-fanout SSE streaming endpoint** (diffs chunked per file, fanned to LLM concurrently) cutting time-to-first-feedback ~55%; applied **optimistic locking via version vectors** and **normalized schema with soft deletes** for concurrent PR reviews.

TECHNICAL SKILLS

Languages: C++, Python, Java, JavaScript, SQL, HTML/CSS,

Frameworks: React, Flask, FastAPI, Scikit-learn, PyTorch, OpenCV, Flutter, Node.js

Systems: Distributed Systems, Load Balancing, Caching (Redis), REST APIs, Microservices, WebSocket, Full-Stack development

AI / ML: LLMs, RAG, Vector Databases (pgvector), HNSW, NLP, Deep Learning, Transformers

Tools: Git, GitHub, MongoDB, PostgreSQL, Docker (Microsoft Certified DevOps Engineer)

Soft-Skills: leadership ,Innovation, Creativity