

SARTHAK SHRIVASTAVA

Bengaluru, India | sarthakshrivastavacontact@gmail.com
linkedin.com/in/sarthak-builds | github.com/SarthakShrivastav-a

EXPERIENCE

Koncepts | Founding Engineer

Bengaluru, India | Feb 2026 – Present

Python, FastAPI, GenAI, PostgreSQL, LiveKit, Stripe, AWS ECS, Docker

- Architected and solely built the **FastAPI** backend for **Koncepts**, an AI learning platform for IB schools (PYP, MYP, DP) serving **12K+ DAU** across students and teachers, deployed on **AWS ECS**.
- Engineered real-time **Voice + Avatar AI Tutors** on **LiveKit** with streaming **ElevenLabs** STT/TTS, VAD, semantic turn detection, and **AI** avatars, syncing a live whiteboard, math widgets, and 3D models to word-level speech timing.
- Designed **LLM content pipelines** for generating IB assessments, cheatsheets, and interactive lessons for primary-school learners, built on a **Generative UI** framework where the LLM composes and live-edits a schema-driven registry of 46 interactive components via validated JSON props and patches, built around **Batch APIs** to cut generation cost by a flat **50%**.
- Exposed the platform's internal tools as **MCP (Model Context Protocol)** servers, letting the team run content, ops, and admin workflows through AI agents and eliminating repetitive manual work.
- Built a **Stripe** billing engine with **exactly-once checkout** (idempotency keys, Postgres advisory locks, frozen price quotes), an order-independent **webhook inbox**, versioned pricing with targeted offers, parent-pays-for-child purchases, secure app-to-web checkout handoff, and automated reconciliation and drift-detection jobs.

Prepairo | Founding Engineer

Bengaluru, India | May 2025 – Present

Java, Spring Boot, Spring Cloud, PostgreSQL, Kafka, AWS, CI/CD, FastAPI, Python

- Implemented a production-grade **Payments and Subscriptions** service integrating **Razorpay**, **Google Play Billing**, and **Apple IAP** with idempotency keys, signature verification, webhook retry safety, event deduplication, order lifecycle state machine, refunds, and reconciliation.
- Secured and optimized paid video delivery via **S3** and a **CDN** layer with signed access and caching, improving reliability and reducing content load time.
- Built and optimized a high-throughput **Notifications** service using **Kafka** and async processing, scaling to **100K+ notifications/day** with batching, dedupe, rate limiting, and failure recovery.
- Engineered **Airo**, a type-safe UPSC AI Tutor on **Pydantic-AI** and **FastAPI** with SSE streaming, intent routing, persistent context, and OCR-based evaluation of handwritten answers, backed by an **LLM circuit breaker** with automatic failover across multiple LLM providers, and **cost-aware model routing** that matches each request to the right model by user tier and usage, keeping response quality high while holding LLM spend predictable at scale.
- Built an automated **Current Affairs pipeline** on **AWS Bedrock** that turns daily news into RAG-grounded MCQs and summaries in English and Hindi, and generates a daily **two-host AI podcast** from it with distinct LLM-scripted host personas, **ElevenLabs** multi-speaker dialogue, and bilingual episodes served via **CloudFront**.
- Maintained CI/CD with **GitHub Actions**, **Docker**, **Amazon ECR**, **Kubernetes (EKS)**, and **Argo CD** for automated build, test, publish, and rollout to production clusters.

PROJECTS

Sentinel — SaaS Website Monitoring Platform

Spring Boot, Golang, MongoDB, SQLite, Docker

- Monitors **uptime** and **SSL expiry** across user endpoints with concurrent **Goroutine** checks, **Quartz**-scheduled reports, cron **heartbeat tracking**, and live **WebSocket** dashboards.

Microservices Observability Platform

Java, Spring Boot, PostgreSQL, Kafka, Redis, Docker

- Trace ingestion of **millions of spans** in **PostgreSQL**, **trace search** / **trace-by-ID** APIs, alert engine (p95, error rate).

EDUCATION

RUNGTA COLLEGE OF TECHNOLOGY, Bhilai, Chhattisgarh

Expected 2027

Bachelor of Technology (GPA: 7.4 / 10.00)

TECHNICAL SKILLS

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

Backend & APIs: Spring Boot, Spring Cloud, FastAPI, Node.js, REST, gRPC, WebSockets

GenAI & LLM Tools: LangChain, LangGraph, Pydantic-AI, RAG, Vector DBs

Data & Messaging: PostgreSQL, MongoDB, Redis, Kafka

Cloud & DevOps: Docker, Kubernetes, AWS (EC2, S3, ECR, ECS, EKS, Lambda), GitHub Actions, Argo CD, CI/CD

Observability & Testing: Grafana, Metrics/Tracing, JUnit, Mockito, TestContainers

ACHIEVEMENTS

7 hackathon wins (2024–2025) across **1 global** and **6 national-level** hackathons: 5 first-place finishes and 2 runner-up finishes.