Suraj Khonde

Full Stack Engineer · Node.js · TypeScript · React / Next.js · LLM Integration · AWS

surajrkhonde@gmail.com · +91 7348887772 · Portfolio · GitHub · LinkedIn

SUMMARY

Full Stack Engineer with 3+ years shipping production Node.js / TypeScript systems end-to-end — React / Next.js frontends over queue, caching, and encryption backends, with Postgres schema design, AWS deploys, and CI/CD. Recent work includes production LLM integration with the Claude API (vision + text) behind confidence gates and guardrails. Currently building and running **Pilooopu**, a live WhatsApp bulk-invite platform at pilooopu.shop, and **Dawa Saathi**, an AI medicine-awareness Telegram bot.

EXPERIENCE

Software Engineer — Kore.ai (via Advento Technologies), Hyderabad — Sep 2025 – Jan 2026

Conversational-AI assistant for banking — home-loan & credit-card support

- Worked on the team behind an LLM support assistant that answers banking queries — home loans, credit cards, interest rates, loan tenure — using RAG over product docs to deflect routine questions from human agents.
- Built the bot-to-agent handoff over WebSocket with a Redis-backed queue, so a session escalates to a live agent when user intent is high or model confidence is low, instead of guessing.
- Added a confidence gate with grounded, doc-only prompting so the assistant routes uncertain answers to a human rather than risk a wrong reply.
- Built a Claude vision agent for the WhatsApp channel that reads uploaded bills (image or PDF), extracts the amounts into the accounting (Tally) workflow, and asks the user to resend when the image is unclear.

Full Stack Engineer — Technoloader Pvt Ltd, Jaipur — May 2024 – Jul 2025

Real-time trading platform

- Re-architected the live-price feed from a 1-second MySQL poll to an event-driven Redis publisher/subscriber model — filtered Binance ticks land in Redis and fan out to subscribed Socket.IO clients, cutting price-delivery delay from up to ~1s to near real-time.
- Took live prices off the database hot path — stored only the latest value per coin in Redis and snapshotted to MySQL on an interval instead of writing every tick, removing continuous per-tick writes and per-second polling reads.
- Cut infrastructure cost ~40% by serving real-time data over Redis pub/sub instead of a heavier message-broker setup (e.g. Kafka) and removing the bulk of recurring database calls — sized deliberately for the platform's ~100–500 active users.
- Moved signup email verification off the request–response cycle into a Redis-backed async job — the API enqueues and responds immediately instead of blocking on the email send, cutting signup response time from ~1–2s to under ~100ms.
- Restructured the app into a modular monolith — an API server for admin/client requests and background jobs, plus a dedicated socket server holding the single Binance connection — filtering to the ~20 most-traded coins and exposing a Redis "market-live" flag so clients always knew prices were fresh.

Full Stack Engineer — Avisirah Technologies, Hyderabad — Mar 2022 – May 2024

Real-time social platform

- Built feed and messaging APIs on Node/Express + MongoDB with Redis as read-cache for hot timelines.
- Implemented JWT access/refresh auth with HttpOnly cookies and RBAC across admin / user / moderator routes.
- Extracted shared middleware (auth, validation, pagination, error formatting) into a reusable layer used across every module.
- Cached high-read endpoints in Redis with TTL invalidation; p50 response time down ~20%.
- Delivered real-time chat, presence, and read receipts over Socket.IO with Redis-backed fan-out across instances.

Financial Advisor — Aditya Birla Capital — 2019 – Feb 2022

Financial advisory — pivoted to software engineering in 2022

TECHNICAL SKILLS

Languages	JavaScript (ES6+), TypeScript
Frontend	React, Next.js, Redux Toolkit (RTK Query), Tailwind CSS
Backend	Node.js, Express, REST, WebSockets / Socket.IO, BullMQ, RabbitMQ, JWT, RBAC
AI / LLM	Claude API (vision + text), prompt design, RAG, confidence gating & guardrails
Databases	PostgreSQL (Drizzle ORM), MongoDB, MySQL, Redis, Firebase Firestore
Infra & DevOps	AWS (EC2, S3), Docker, Nginx, PM2, CircleCI (CI/CD), Swagger / OpenAPI
Tools	Cloudinary, Puppeteer, Razorpay, Firebase (Auth, FCM), Pino logging

PROJECTS

Pilooopu — Bulk Personalized WhatsApp Invite Platform — pilooopu.shop · GitHub

Node.js · TypeScript · BullMQ · Redis · PostgreSQL · Drizzle · Next.js · Puppeteer · WhatsApp Cloud API · Razorpay · Docker · Nginx

- Designed a 3-stage queue pipeline (render → upload → finalize) across 7 BullMQ queues, checkpointed in Redis — a worker crash resumes from the last completed step, not the start.
- Encrypted every guest phone with AES-256-CBC and a unique IV per record; decrypted only inside the send worker, masked as +91*****1234 in every API response and log.
- Split retries into retryable vs non-retryable errors — bad data routes straight to a dead-letter queue instead of burning the 3-attempt budget.
- Reused a single Puppeteer browser across render jobs, producing 400×600 thumbnail and 800×1200 full card from one HTML pass.
- Built a Redis-backed rate limiter with per-endpoint quotas (login 5 / 15 min, signup 3 / hr, event-create 20 / hr) and standard X-RateLimit headers.
- Deployed as two Docker services (API + Worker) behind Nginx; graceful SIGTERM shutdown with 10s force-kill timeout.

Dawa Saathi — AI Medicine Awareness Telegram Bot — t.me/dawasaathi_bot · GitHub

Node.js · TypeScript · Claude Vision + Text · Redis · PostgreSQL · openFDA · Telegram Bot API · Docker

- Built a Telegram bot that reads a medicine-strip photo and returns an elder-friendly drug summary; Claude vision extracts the label at ~80% accuracy.
- Tightened guardrails with a 0.75 confidence gate and ingredient-only prompts that refuse uncertain reads — cutting misleading or wrong drug info ~80%.
- Cached analyzed medicines in Redis → PostgreSQL so repeat scans skip the model — cutting token consumption ~30% and serving answers instantly.
- Dockerized the full app with health checks and graceful shutdown for one-command deploys and fast scaling.

EDUCATION

B.E. Mechanical Engineering — VTU, Belagavi (2018)