Gokul JS

jsgokul123@gmail.com https://gokuljs.com/ https://github.com/gokuljs https://www.linkedin.com/in/gokul-js/
https://twitter.com/gokul_js029

Software engineer with 4.5 years of startup experience building LLM-based products from scratch and running them at scale. At YC-backed companies, I worked across the stack, scaling systems, reducing latency, and hardening them for real-world production use.

PROFESSIONAL EXPERIENCE

Rime | Senior Software Engineer 07/2025 – 02/2026 | San Francisco, CA (Remote)

- Built and owned end-to-end real-time voice pipelines for Rime Text-to-Speech using LiveKit and Pipecat over WebRTC.
- Integrated Rime TTS plugins across LiveKit and Pipecat to deliver low-latency, production-grade real-time audio, contributing to a 150% revenue increase.
- Diagnosed and resolved latency, reliability, and failure issues across the WebRTC stack and backend WebSocket infrastructure, significantly improving audio stability and response times.
- Designed custom chunking and streaming algorithms to optimize real-time audio delivery and reduce playback inconsistencies.
- Led incident response as on-call owner for real-time audio systems, shipping rapid fixes and implementing safeguards to prevent regressions.
- Built and supported production-grade customer-facing voice agents on top of the real-time infrastructure, driving enterprise adoption.
- Contributed across the frontend stack to ship customer-facing voice features and improve overall product performance.

Teamble | Lead Engineer 06/2024 – 03/2025 | New York, NY (Remote)

- Architected and maintained high-performance web applications using Node.js and React, ensuring scalability and reliability across production systems.
- Designed and implemented a frontend design system from scratch, standardizing UI components and accelerating feature development.
- Optimized backend APIs by refactoring business logic and migrating inefficient SQL queries, significantly reducing response latency.
- Built LLM-powered agents using LangGraph for internal and customer-facing workflows, adopted by 100,000+ enterprise users.
- Developed a conversational AI interface with role-based permissions and real-time data access for secure enterprise deployment.
- Implemented a multi-agent RAG architecture for scalable performance review generation, improving contextual accuracy and response quality.
- Built a One-on-One AI agent that aggregates quarterly performance data to generate structured, actionable insights for managers.

Aerotime, Y Combinator (W21) | Founding Engineer 11/2022 – 02/2024 | San Francisco, CA (Remote)

- Owned frontend architecture and infrastructure end to end, improving performance, scalability, rendering efficiency, and long term maintainability.
- Built a scalable design system and reusable component library, standardizing UI patterns and accelerating developer velocity across products.
- Architected a production grade JavaScript SDK to enable seamless third party integrations and expand platform extensibility.
- Drove customer discovery and feedback loops, translating insights into high impact product improvements and influencing roadmap decisions.
- Deployed and productionized open source LLMs including LLaMA and Stable Diffusion, integrating AI capabilities into core workflows.
- Fine tuned models using the Hugging Face ecosystem with techniques such as DreamBooth and Textual Inversion for domain specific optimization.
- Implemented analytics, telemetry, and a high performance virtualized data table to support large datasets and enable data driven UX improvements.

EDUCATION

CMR Institute Of Technology 2017 – 2021 | Bangalore, India
B.E Computer Science and Engineering

SKILLS

Languages: JavaScript, TypeScript, Python

Frontend: React JS, Next JS, SolidJS, Redux, React Query, Tailwind CSS, shadcn/ui, GSAP, HTML5, CSS, SASS.

Backend: Node.js, Express, Fast API, Redis, SQL, Prisma, Django, tRPC, gRPC, GraphQL, ElasticSearch, Logstash, Kibana, Suricata, Zeek, Bull MQ, TypeORM, Drizzle, Clerk(Auth)

Databases: PostgreSQL, MySQL, MongoDB, DynamoDB,

Large Language Models and AI: LangGraph, LlamaIndex, Langchain, RAG, Stable Diffusion, Prompt Engineering, Prompt Evaluation,

Cloud and DevOps: AWS (S3, ECR, EC2, etc...), Docker, GitHub CI/CD, Vercel (serverless)

Realtime: LiveKit, Pipecat, WebRTC, Twilio(SIP+SDK)