Suraj Rawat

Software Engineer (Full-Stack AI/ML)
Gurgaon, India — +91-9555219911 — suraj.schs99@gmail.com
github.com/SurajTrs — linkedin.com/in/surajrawat99

Professional Summary

Software Engineer with **2.2 years of experience** designing and deploying scalable full-stack applications and AI-driven systems. Expert in **MERN/Next.js**, **Python**, and **AI/ML** (LLMs, NLP, speech processing), with a focus on building low-latency, cloud-native solutions on AWS. Delivered voice-enabled assistants and optimized APIs, improving automation by **30%** and achieving sub-second response times. Adept at system design, testing, and Agile methodologies, with a passion for solving complex technical challenges.

Technical Skills

Languages: JavaScript (ES6+), TypeScript, Python, SQL
Frontend: React, Next.js, Redux, Tailwind CSS, Vite
Backend: Node.js, Express, REST, GraphQL, WebSockets, JWT
Databases: MongoDB, Firestore, PostgreSQL, Mongoose, Indexing
AI/ML: Transformers, Hugging Face, PyTorch, TensorFlow, Rasa, spaCy, LLM Fine-tuning
Speech Processing: Google STT, Mozilla DeepSpeech, Tacotron TTS
Cloud & DevOps: AWS (EC2, S3, Lambda, API Gateway), Docker, CI/CD (GitHub Actions), Nginx
Testing & Tools: Jest, Playwright, Postman, Unit/Integration Testing
Methodologies: Agile, Clean Architecture, System Design, Code Reviews

Professional Experience

Software Development Engineer May 2024 – Present
Capyngen Pvt. Ltd., Gurgaon, India

- Architected **AI voice bots** using GPT and LLaMA models, increasing automation accuracy by **30%** for client workflows.
- Optimized **low-latency inference APIs** on AWS EC2 and Lambda, achieving **sub-300ms response times** under high load.
- Integrated Google STT and Tacotron TTS for real-time voice interactions, with secure JWT authentication and rate limiting.
- Authored **system design documentation**, reducing team onboarding time by **20%** through clear runbooks.

Software Development Engineer I Jan 2024 – Mar 2024
Codbee.in (Remote)

- Developed a **voice-enabled learning assistant** using Hugging Face Transformers, supporting contextual multi-turn dialogs for **10K+ users**.
- Reduced **ASR latency to sub-second** with Mozilla DeepSpeech and streaming inference, improving user experience.
- Built **microservices** in Node.js and Python for scalable inference and state management, integrated with CI/CD pipelines.
- Collaborated with UX teams to refine intent detection, boosting task completion rates by **25%**.

Full-Stack Developer Intern Sep 2023 – Jan 2024
Advance Career Guide Pvt. Ltd. (Remote)

- Integrated **voice command features** into a chatbot, enhancing course discovery and increasing user engagement by **15%**.
- Engineered **AWS deployments** with observability and security hardening, improving system uptime to **99.9%**.

Web Developer Intern Jun 2023 – Aug 2023
Suvidha Foundation (Remote)

- Enhanced **UI accessibility** and responsiveness, reducing navigation time by **20%** across devices.
- Implemented **NLP-based search**, improving content discoverability for **5K+ monthly users**.

Key Projects

AI Voice Assistant (Tripy) GitHub — Live Demo

- Built a scalable **voice assistant** using GPT-based LLMs, Google STT, and Tacotron TTS for natural multi-turn conversations.
- Deployed **low-latency APIs** on AWS EC2/Lambda with JWT authentication, achieving **sub-300ms** response times.
- *Stack: Python, Node.js, Transformers, PyTorch, AWS, REST APIs*

Student Sharpner Edtech Platform GitHub — Live Demo

- Developed a **scalable Edtech platform** with live classes, AI-driven search, and role-based access for **3 user types**.
- Implemented **JWT-based authentication** and deployed on Vercel for global accessibility.
- *Stack: React, Node.js, Express, MongoDB, Tailwind CSS, Vercel*

Amazon Clone E-Commerce GitHub — Live Demo

- Engineered a **responsive e-commerce platform** with real-time Firestore updates, authentication, and checkout flows.
- Optimized UI with React hooks, improving load times by **15%** across devices.
- *Stack: React, Firebase Auth, Firestore, FakeStore API*

Education

B.Tech, Electronics and Communications Engineering 2020 – 2024
Indian Institute of Information Technology (IIIT) Dharwad
GPA: 7.1/10 — Relevant Coursework: Machine Learning, Distributed Systems, Cloud Computing, Data Structures

Certifications

- AWS Certified Developer – Associate, 2024
- TensorFlow Developer Certificate, 2023