

Abhishek Kumar Singh

Senior Software Developer — Applied AI & Platform Engineering
Bengaluru, India | +91 6363695280 | abhishek81121@gmail.com
GitHub Profile | LinkedIn Profile | LeetCode Profile

Summary

Senior Software Developer with experience building and operating production AI systems at scale. Specialized in multi-agent architectures, AI evaluation frameworks, real-time communication platforms, and Kubernetes-based AI infrastructure. Proven track record of reducing AI infrastructure costs, scaling GPU-backed workloads, and improving reliability of AI-powered products supporting hundreds of concurrent sessions.

Skills

Languages: Python, TypeScript, JavaScript

AI Systems: Google ADK, Multi-Agent Systems, Agent Orchestration, AI Evaluation Frameworks, Conversational AI, Prompt Engineering, Prompt Caching

Infrastructure: Kubernetes, GKE Autopilot, HPA, KEDA, Docker, GPU Workloads, AI Inference Systems, LiveKit, WebRTC

Backend: FastAPI, Django, REST APIs, Microservices, WebSockets

Databases: PostgreSQL, MongoDB, Redis

Experience

Recruit41

Bengaluru, India

Senior Software Developer

May 2026 – Present

- Driving Harness Engineering initiatives through AI-assisted development workflows, evaluation mechanisms, and engineering guardrails to improve code generation accuracy and developer productivity.
- Designed and implemented a hierarchical multi-agent orchestration framework using Google ADK, introducing supervisor agents for tool governance, interview flow control, and candidate evaluation, improving reliability of AI-driven interview workflows.
- Led the evolution of a real-time AI interview platform from a custom WebSocket architecture to a WebRTC-based stack powered by LiveKit and Google ADK.
- Built an on-the-fly AI evaluation framework that continuously monitored agent behavior, tool usage, and interview progression, enabling rapid detection and correction of workflow failures in production.
- Architected and operated GPU-backed AI infrastructure on GKE Autopilot, deploying production-scale speech and inference workloads while reducing dependency on third-party providers.

Software Developer

Sep 2024 – May 2026

- Spearheaded migration from Daily.co to a self-managed LiveKit deployment supporting **350+ concurrent sessions** while reducing infrastructure costs by approximately **75%**.
- Achieved approximately **30x speech-processing cost efficiency** by replacing external speech APIs with self-hosted inference infrastructure deployed on GPU-backed Kubernetes workloads.
- Designed autoscaling infrastructure using Kubernetes HPA and KEDA to dynamically scale communication, speech-processing, and AI workloads during burst traffic conditions.
- Implemented prompt caching and inference optimization strategies across hosted LLM providers to reduce latency and AI operating expenses.
- Conducted load testing and performance tuning of distributed AI systems, identifying and resolving scaling bottlenecks affecting high-concurrency workloads.

Think41

Bengaluru, India

Full Stack Developer Intern

Jul 2024 – Aug 2024

- Built React and Django applications for an AI-driven interview scheduling platform and recruiter workflow automation.
- Developed LLM-powered resume screening workflows and implemented automated backend testing using Pytest to improve candidate shortlisting efficiency and API reliability.

Open Source Contributions

LiveKit Rust SDK

- Diagnosed and resolved an authentication-token propagation issue affecting fallback routing under elevated traffic conditions and contributed the fix upstream, improving reliability for production deployments.

Education

CMR Institute of Technology

Bengaluru, India

Bachelor of Engineering in Computer Science

2020 – 2024