PREM SAI GADWAL

+91-6360231174 — premsaig1605@gmail.com — linkedin.com/in/prem-sai-gadwal-b665811a1/ — github.com/premg16

Summary — Full-Stack Developer and AI Engineer with 2+ years of experience building production-scale AI systems and web applications. Expertise in **Generative AI, LLM integrations, and real-time applications** using Next.js, TypeScript, and Python. Demonstrated success designing and implementing high-performance systems supporting thousands of concurrent users while delivering exceptional user experiences.

Skills

- **Programming Languages:** TypeScript, Python, C++, Java, Bash
- **Frameworks & Libraries:** Next.js, React.js, Node.js, FastAPI, Flask, LangChain
- **AI/ML:** LLM Integration (OpenAI, Llama, HuggingFace), RAG, Stable Diffusion, Imagen, Flux
- **Cloud & Infrastructure:** AWS, GCP, Azure, Kubernetes, Docker, Terraform
- **Developer Tools:** Git, CI/CD, Jenkins, VS Code Extension Development
- **Design:** Figma, Blender, Spline, Illustrator

Professional Experience

Publicis Sapient **June 2024 – Present**
Associate Software Development Engineer L2

AI Studio

- **Led frontend development** for an enterprise-grade AI multimedia studio processing 10,000+ image generation requests daily using **Stable Diffusion, Flux, and Imagen 3**
- **Developed "PromptBoost"**, an internal AI tool that auto-tuned user prompts based on user behavior analytics, improving image generation success rate by 25
- **Reduced load time by 40%** by implementing efficient state management and optimizing API calls for high-volume asset rendering
- **Designed and built** a real-time streaming Chat UI with multi-modal capabilities (text, image, audio) that increased user engagement by 35%
- **Implemented custom React hooks** to streamline LLM integration, reducing boilerplate code by 50% and accelerating feature development
- **Established performance benchmarks** and implemented monitoring for AI systems, maintaining 99.9% uptime and sub-200ms response times

Engage Next

- **Scaled Quiz Next platform** to handle 5,000+ concurrent users with real-time leaderboards and interactive elements
- **Achieved 98% test coverage** while integrating new AI features, resulting in zero critical bugs during production launch
- **Engineered multilingual virtual assistant** with context-aware responses using vector databases and RAG, supporting 8 languages
- **Implemented real-time data synchronization** using WebSockets, reducing latency by 60% for competitive quiz features
- **Optimized image generation pipeline**, cutting processing time from 5s to 1.2s while maintaining high-quality outputs

Publicis Sapient **June 2022 – June 2024**
Associate Software Development Engineer L1

Quiz Next

- **Developed end-to-end AI-powered quiz generation system** using RAG to automatically create quizzes from transcripts and session content
- **Built scalable backend architecture** on GCP that processed 500+ concurrent quiz sessions with real-time scoring and analytics
- **Reduced question generation latency by 65%** through caching strategies and optimized model inference
- **Created dashboard with real-time analytics** that increased client adoption of the platform by 45%

Codebuddy - VS Code Extension

- **Co-developed VS Code extension** for AI-assisted coding that achieved 3,000+ installs within first month
- **Architected integration layer** for multiple LLM providers, allowing seamless switching between models
- **Implemented token-efficient code chunking** algorithm that reduced API costs by 30% while improving code context understanding
- **Built streaming response UI** with syntax highlighting and code completion that reduced development time for users by 20%

Education

Indian Institute of Information Technology, Surat **2017 – 2022**
Bachelor of Technology in Electronics and Communication

Certifications

- AWS Cloud Practitioner
- Google Cloud Digital Leader

Projects & Contributions

AI Technology Writer — Medium **Mar 2023 – Present**

- Published 12+ technical articles on Generative AI applications, garnering 50,000+ total views
- Created practical tutorials for implementing LLMs in production environments
- Analyzed emerging trends in text-to-image/video models and their industry applications