

Yatheswar Ethalapaka

Software Engineer | Full Stack | Gen & LLM Applications

+91 8374178371 | yatheswar546@gmail.com | [LinkedIn](#) | [Github](#) | [Portfolio](#)

Professional Summary

Software Engineer with hands-on experience building full-stack and AI-powered applications using JavaScript, TypeScript, Python, React, Next.js, Node.js, MongoDB, and REST APIs. Experienced in developing and deploying LLM-powered applications using OpenAI and Google Gemini APIs, with practical knowledge of RAG, prompt engineering, embeddings, vector search, conversational memory, and AI agents. Strong AWS Cloud and production troubleshooting background from resolving 200+ critical customer issues and supporting 100+ customers, combined with experience delivering client applications, optimizing performance, and integrating authentication, databases, and third-party services. Comfortable working across the software and AI application stack, from frontend and backend development to cloud deployment and intelligent application workflows.

Skills

OS: Windows, Linux

Languages: JavaScript, TypeScript, Python, C, C++

Technologies/Framework: React.js, Next.js, Node.js, Express.js, Tailwind CSS, MongoDB, MySQL, PHP, DSA, OOPs

Cloud: AWS (EC2, S3, IAM, VPC, CloudWatch, SNS, EMR), AWS Solutions Architect Associate

GenAI/LLM: LangChain, LangGraph, OpenAI API, Google Gemini API, RAG, Prompt Engineering, Embeddings, Vector Search, Conversational Memory, AI Agents

Process: Agile, SDLC, Debugging, Performance Optimization

Tools: Git, GitHub, VS Code, Postman, Cursor, GitHub Copilot

Experience

Freelancing Full Stack Developer & Technical Trainer | Visakhapatnam, India

[May 2025 - Present]

Freelancer Full Stack Developer and Trainer

- Developed and delivered full-stack web applications for clients and independent projects using JavaScript, TypeScript, React, Next.js, Node.js, PHP, MySQL, MongoDB, and Tailwind CSS.
- Built AI-powered applications integrating LLMs and GenAI capabilities, including conversational memory, RAG-based workflows, prompt engineering, and Google Gemini APIs.
- Designed and implemented responsive, production-oriented applications with authentication, databases, APIs, performance optimization, and third-party service integrations.
- Improved web application performance through image optimization, WebP conversion, lazy loading, prefetching, and progressive loading, achieving significant reductions in page weight and improved PageSpeed scores.
- Worked with AI-assisted development tools such as Cursor, GitHub Copilot, Lovable, Bolt, ChatGPT, and Perplexity to accelerate development while maintaining code quality and maintainability.
- Conducted practical web development training and mentored learners through hands-on projects, helping them understand development workflows, debugging, and real-world application development.

Amazon Web Services | Bangalore, India

[Dec 2023 – May 2025]

Cloud Engineer

- Resolved 200+ customer-reported critical performance issues across EC2, IAM, S3, VPC, CloudWatch, and SNS by analyzing distributed system logs and cluster behavior to isolate root cause.
- Supported 100+ customers running big-data workloads on AWS EMR by troubleshooting distributed Hadoop and Spark cluster configurations and data pipeline issues.
- Prevented recurring runtime failures for 100+ customers by proactively configuring CloudWatch monitoring and alerting, enabling issues to be caught and resolved before customer impact.
- Improved customer self-sufficiency on big-data tooling by guiding 100+ customers through Hadoop, Spark, and EMR cluster-configuration concepts during support engagements.
- Earned the AWS Solutions Architect Associate certification while in role, strengthening cloud architecture judgment applied directly to customer troubleshooting.

Projects

Full Stack AI Voice Assistant [\[Link\]](#)

[Jun 2026 - Present]

Skills: Next.js, TypeScript, MongoDB, Google Gemini API, JWT Auth

- Built and deployed a full-stack, voice-enabled AI assistant with authentication, persistent multi-session chat, and contextual memory.
- Eliminated context loss across multi-turn conversations by designing a MongoDB-backed conversational memory system that retrieves and chronologically orders the most recent 20 messages per session before each Gemini API call.
- Delivered secure, isolated multi-user chat history by implementing JWT authentication with bcrypt password hashing and a Users/Conversations/Messages data model in MongoDB via Mongoose.
- Enabled hands-free interaction by integrating browser-native Web Speech APIs for voice input and text-to-speech output, alongside a responsive UI with dynamic, recency-grouped chat session management.

NAKAMA Network — Full Stack Application

[Jun 2026 – July 2026]

Skills: Next.js, TypeScript, Tailwind CSS, Firebase Auth, Razorpay

- Co-developed the official website for a growing anime community as one of three engineers, translating client reference designs into a responsive, modular frontend with Firebase authentication and Razorpay payments.
- Cut per-image page weight by 94% (~5 MB to ~300 KB) by implementing image compression and WebP conversion, directly reducing load time for the platform's largely mobile audience.
- Reduced unnecessary API calls and improved perceived load performance by architecting lazy loading, progressive loading, and prefetching into the component structure from the start of the build.
- Enabled parallel development across a 3-person team without merge conflicts by structuring the frontend around Separation of Concerns, modular components, and reusability principles.

AbsIncGroup | Biopharmaceuticals Landing Page [\[Link\]](#)

[Aug 2025 – Oct 2025]

Skills: HTML, CSS, JavaScript, AI-Assisted Development (Lovable, Bolt)

- Built a responsive biopharma consulting landing page for a U.S.-based client using an AI-assisted ("Vibe Coding") workflow.
- Raised the Google PageSpeed performance score from ~55% to ~90% (a 35-point improvement) by implementing lazy loading, prefetching, and aggressive media optimization.
- Cut image payload by 97.5% (~20 MB to ~500 KB) through compression and WebP conversion, the single largest contributor to the site's performance gain.
- Eliminated video-driven page bloat by offloading large AI-generated video assets to Vimeo and embedding them via iframe instead of hosting them directly on the site.

Certifications

- Gen AI Program — Codekerdos | [Mar 2026 – July 2026]
- Gen AI Bootcamp — Outskills | [May 2026]
- Crio Fellowship in Software Development – Frontend | [Mar 2025 – Apr 2026]
- AWS Solutions Architect Associate | [Dec 2024]
- Devtown — Full Stack Web Development | [Nov 2022 – Feb 2023]
- CodeChef Certified Data Structures & Algorithms (CCDSA) | [July 2022]

Education

B.Tech, Computer Science and Engineering [2019 - 2023]

Vignan's Institute of Information Technology (A)

CGPA: 8.51 / 10