Sai Sumanth Reddy Kachi

saisumanthreddy216@gmail.com | 667-445-9499 | linkedin.com/in/sumanth14 | Portfolio

AI Engineer with 3+ years of experience with production LLM systems. Experienced in RAG pipelines, AI integration, vector databases, and multi-agent workflow design with a track record of shipping AI features at scale across healthcare and enterprise platforms.

EDUCATION

University of Maryland, Baltimore County, Baltimore, MD — *May 2026*
Master of Science - Information Systems — GPA: 3.85
Relevant Coursework: Artificial Intelligence, Natural Language Processing, Data Structures & Algorithms, Cloud Computing, System Design

Dayananda Sagar Institutions, Bangalore, India — *Aug 2018 - July 2022*
Bachelor of Science - Computer Science — GPA: 3.77

SKILLS

Programming Languages: Python, Javascript, TypeScript, HTML, CSS, C, C++, C#
Frameworks & Libraries: FastAPI, LangChain, React.js, NestJS, Node.js, Express.js, Next.js, .NET, n8n
Databases & Storage: MySQL, PostgreSQL, MongoDB, Supabase, Firebase, Oracle, Redis
Cloud & DevOps: AWS (S3, Lambda, EC2), Azure , GCP, Docker, Kubernetes, Jenkins, CI/CD Pipelines, GitHub
AI & ML: RAG Pipelines, LLMs, Prompt Engineering,Multi-Agent Orchestration, Model Context Protocol(MCP)
Core Competencies: LLM System Design, API Development, Microservices Architecture, Context Window Management, Retrieval-Augmented Generation, Agent Prompt Design, REST APIs, System Design

EXPERIENCE

Minnodi LLC — *Baltimore, MD*
Software Engineer Intern — *Nov 2025 – Present*

- Built MCP AI agents leveraging Claude Sonnet 4.6 (Amazon Bedrock) for healthcare portal analyzing 500+ patient histories, generating pre-consultation summaries, and GitHub Assistant for commit summarization.
- Implemented 10 RESTful APIs for document workflows, patient records, and appointment scheduling with role-based access control across 4 user roles, cutting 20+ hours of weekly administrative workload.
- Designed normalized database schema with optimized indexing and relational queries, reducing data retrieval latency by 35% while supporting 1,000+ active records.

RSM US LLP — *Bangalore, India*
Software Development Engineer — *Aug 2022 – Jun 2024*

- Led end-to-end design and Azure-hosted deployment of Enterprise Mobility Solution (React.js, Node.js, Express.js) within an Agile/Scrum framework, automating 95% of scheduling workflows for 5,000+ employees, saving 80+ admin hours monthly and earning the organization-wide "Shining Star" award.
- Optimized system performance across 10 enterprise applications through configuration improvements and code refactoring, reducing latency by 30% and earning organization-wide recognition award.
- Engineered multi-functional API system with role-based access across 10 permission levels, serving 2,000+ users and 10 teams across the enterprise, improving cross-team development efficiency by 30%.

Cognizant Technology Solutions — *Bangalore, India*
Programmer Analyst — *Mar 2022 – Jul 2022*

- Developed 20+ reusable application components with TypeScript following established coding standards, reducing development time by 40% and supporting consistent enterprise application architecture.
- Designed efficient database models and implemented indexing strategies, reducing query execution time by 40% and enhancing backend scalability for high-traffic enterprise operations.

PROJECTS

Kernel: AI-Native SDLC Framework- *GitHub* — *Multi-Agent Orchestration, Claude API, SDLC Design*

- Architected a multi-agent SDLC runtime that orchestrates 8 specialized AI agents (PM, BA, Architect, Developer, QA, Security, DevOps) across 3 phase-gated workflows, enforcing mandatory deliverables and human approval gates before each phase transition.

Invenio: AI-Powered Job Discovery Platform- *GitHub* — *React.js, Node.js, Express.js, AWS, Supabase, Gemini*

- Built a full-stack job discovery platform using Claude Code with vibe coding and context engineering workflows, intelligent multi-ATS scraping, real-time alerts and role-based filtering by serving 20+ users.

We Zaap: AI Interview Simulator - *GitHub* — *React.js, FastAPI, RAG, PostgreSQL, LLMs*

- Developed the frontend for an AI-powered interview simulator supporting 10 job roles, implementing lazy loading and RAG-based context retrieval for role-specific questions, currently serving 50+ students.