Sai Ram Sana

Senior Agentic AI Engineer

Mississauga, ON | sanasairamca@gmail.com | linkedin.com/in/sairamsana | sairamsana.com | github.com/sairamsana

SUMMARY

Senior Agentic AI Engineer with 9+ years of experience shipping production software. I design and ship production AI systems (RAG, agent workflows, multi-LLM routing and MCP) with the guardrails and verification that make them trustworthy.

CORE SKILLS

LLMs & Agents: Claude, GPT / OpenAI, Gemini / Vertex AI, OpenRouter, Multi-agent orchestration, Tool calling, MCP servers, Claude Code, Structured outputs

RAG & Retrieval: Vector embeddings, Pinecone, MongoDB Atlas Vector Search, LangChain, Chunking strategies, Metadata filtering, Web search grounding (Tavily, Perplexity)

Speech, Voice & Multimodal: Deepgram, AssemblyAI, AWS Transcribe, Azure Speech, Live translation, LiveKit voice agents, ElevenLabs / OpenAI TTS, Gemini audio & video, NER & sentiment

Sales Tools & Automation: Clay, Apollo.io, AI Ark, HubSpot, Outbound sequencing, Lead enrichment, ICP scoring, n8n / Zapier / Make

Backend: Node.js, TypeScript, Express, Python, Flask, REST & WebSockets, Socket.IO, Queues & schedulers, Redis, Kafka

Frontend & Mobile: React 19, Next.js, React Native / Expo, Swift & Kotlin modules, Angular, Vue.js, React Flow, Zustand / TanStack, Capacitor

Cloud & Data: AWS (S3, ECS, Lambda, CloudFront, Cognito, SES, SSM), GCP (Vertex AI, GCS), Cloudflare, Docker, MongoDB, PostgreSQL, FFmpeg

Foundations: Java, Spring Boot, Hibernate, OAuth 2.0 / SSO, Stripe / Paddle / PayPal, Design patterns, Testing (Vitest, Playwright, JUnit), Agile

EXPERIENCE

Senior Full Stack Developer (AI Platform), Speak AI Inc. Mississauga, Canada · Jul 2023 – Present

- Designed and built the AI systems listed under Selected AI Systems: RAG and semantic search, the branching agent workflow engine, the multi-LLM registry and router, multimodal Gemini analysis, MCP tools and the grounded web-research step.
- Built a real-time transcription gateway (WebSocket/Socket.IO, streaming ASR, 48 live languages) and live translation into 51 languages.
- Built an Apollo-driven prospecting agent that auto-generates a personalised voice-agent demo from each prospect's website.
- Shipped a native iOS/Android recorder (Expo, Swift, Kotlin), an embeddable recorder with white-label theming, and a macOS desktop recorder.
- Integrated meeting bots with Google/Outlook calendar auto-join, plus Google Drive, OneDrive, SharePoint, Slack, ClickUp and Zoom workflow actions.
- Implemented Python sentiment analysis (+25% insights) and deployed NER models on autoscaling AWS ECS behind an ALB.
- Integrated Stripe, PayPal and Paddle billing, OAuth 2.0/SSO, multi-region S3 data residency, and SSRF-safe outbound requests.
- Cut latency by 40% with Redis caching. I lead development cycles, code reviews and mentoring.

Client Service Controller (Individual Customer), Canada Life London, Canada · Mar 2023 – Jul 2023

- Analysed business processes to find opportunities for automation.
- Designed an API that automates processing of individual client and Ingenium policies.
- Implemented Kafka producers and consumers in microservices for real-time data flow.
- Led MongoDB version upgrades and migrations across platforms.

Full Stack Developer, Justkapture Innovations London, Canada · Apr 2021 – Feb 2023

- Built FOMO, an app that tracks end-user interactions on web applications to drive sales.
- Built Justkapture, which gathers tagged content from social media sources to improve customer sales.
- Delivered full stack features with Node.js, Express, MongoDB, Angular, TypeScript and AWS.
- Designed PostgreSQL schemas with complex stored procedures and functions.

Associate Developer, Cognizant Technology Solutions Chennai, India · Dec 2018 – Apr 2021

- Designed and developed the Investment Data Service Controller (IDS), which aggregates Bloomberg, Flex and other feeds for traders.
- Automated manual checks, saving about 15% of support team effort annually (~$120,000).
- Built an internal Spring Boot testing library for TWS (IBM Scheduler) QA.
- Used Adobe Experience Manager (AEM) for dynamic content delivery.

Software Engineer, Medoske Technology Solutions Hyderabad, India · Jun 2017 – Dec 2018

- Built doctor appointment and patient case-sheet apps with React and MobX.
- Built a dental analysis platform that improves diagnostic efficiency.
- Built the Redheal Care platform with React and Next.js, with real-time communication and live updates.

SELECTED AI SYSTEMS

RAG & Semantic Search over Conversations, Docs and Websites: Chat with your data across media, documents and websites, with cited sources and indexes that stay fresh without full re-indexing. *(OpenAI Embeddings, Pinecone, MongoDB Atlas Vector Search, LangChain, Node.js)*

Branching Agentic Workflow Engine + Visual Builder: No-code AI agents that branch on what the model finds, so every recording can be routed, enriched and acted on automatically. *(TypeScript, Node.js, MongoDB, React 19, React Flow)*

Multi-LLM Model Registry & Router (with BYOK): New models launch by adding one registry entry, and every routing decision is explainable, cost-aware and resilient to provider limits. *(Claude, GPT, Gemini, OpenRouter, TypeScript)*

Audio, Video & Image Understanding on Gemini / Vertex AI: Users can ask questions about what happens in a video, not just what was said, with predictable cost and data residency. *(Gemini, Vertex AI, Google Cloud Storage, AWS S3, FFmpeg)*

MCP Server: Let Claude & ChatGPT Operate the Platform: Anyone can say "when a sales call lands, extract the objections and post them to Slack" in Claude and get a working automation. *(Model Context Protocol, TypeScript, Zod, Vitest, Claude)*

EDUCATION

Post-Graduate Diploma, Business and Information Systems Architecture, Fanshawe College (President's Honours) London, Ontario, Canada

Bachelor of Technology, Computer Science Engineering, Jawaharlal Nehru Technological University India