Souvik De

Software Engineer, AI Systems | LLMs · RAG · AI Agents · Backend

Gurugram (Delhi NCR), India | Open to remote | +91 98305 05728 | souvikde.tech@gmail.com
souvik.de | linkedin.com/in/souvik-de-a2b941169 | github.com/Souvikns

Summary

Software engineer with 4+ years in backend and developer tooling, now building production LLM systems in Python, Go and TypeScript. At XANE.AI I work on the agent behind Maruti Suzuki's AI assistant, where adding evals to CI helped take answer relevance from 64% to 89%. I also maintain AsyncAPI tools with 36,000+ downloads, wrote the project's MCP server and mentor for Google Summer of Code.

Skills

Languages: Python, Go, TypeScript, JavaScript (Node.js), SQL
LLMs & agents: RAG (hybrid search, reranking, chunking, embeddings), tool calling, MCP, multimodal (vision) inputs, Mastra, LangGraph, LangChain, deepagents, OpenAI, Claude, Ollama, MinerU OCR
Evaluation & observability: RAGAS, DeepEval, evals in CI, Langfuse tracing
Backend & data: REST / OpenAPI design, FastAPI, Gin, Qdrant, PostgreSQL
Infrastructure, cloud & frontend: AWS (EC2, S3, RDS, Bedrock, SageMaker, CloudFront, Route 53, API Gateway), Azure, Docker, Kubernetes, GitHub Actions CI/CD, Linux, React

Experience

Software Engineer II, XANE.AI | *Gurugram, India* Jul 2025 – Present

- Built the Mastra agent behind Maruti Suzuki's AI assistant, which handles 100,000+ queries a day from 4.37M+ users across 34 vehicle models. It began as a RAG system over vehicle manuals and now also checks insurance policy status, handles renewals and quotes, and finds and books workshops.
- For retrieval, manuals go through MinerU OCR instead of manual data entry and are then chunked, embedded and searched in Qdrant with hybrid search and reranking.
- Built a separate deepagents system that sorts raw OCR output into CRM records.
- Added RAGAS evals on a golden set to CI (GitHub Actions) to flag quality regressions and set up self-hosted Langfuse for tracing. Answer relevance rose from 64% to 89%.
- Routed customer photos of car parts to a vision model whose output drives retrieval, and attached provenance metadata to every chunk so answers cite their sources.
- Deployed it on AWS, with Bedrock for the assistant's LLM and the part-identification vision model, EC2 for the backend, and S3 and RDS for vehicle manuals and metadata.

Lead Software Engineer, CodeMate.AI | *Noida, India* Dec 2024 – May 2025

- Wrote a custom search algorithm for retrieval over Swagger/OpenAPI specs that improved precision in the RAG system.
- Cut startup time of the VS Code extension's local server by over 30% by loading models in the background instead of at boot.
- Set up GitHub Actions CI/CD for macOS, Linux and Windows releases and automated deploys to an Azure production VM, saving the team 2–5 hours of release work a week.
- Led a two-engineer team working on the local server and VS Code extension.

Software Engineer, Postman | *Bangalore, India* Feb 2022 – Jun 2024

- Took the official AsyncAPI CLI from an early skeleton to the standard tool for the spec (20,000+ downloads), putting the project's separate tools behind one command.
- Shipped AsyncAPI Bundler (16,000+ downloads), which resolves JSON $ref references across specification files into a single document.
- Maintained 3+ open-source projects, reviewed 185+ pull requests and brought two contributors up to maintainer.
- Mentored 3–5 Google Summer of Code contributors and co-led an effort to standardize parsers across languages.

Maintainer, AsyncAPI Initiative | *Remote, Linux Foundation project* Sep 2021 – Present

- Built asyncapi-mcp, an MCP server with four tools that lets AI agents work with AsyncAPI documents (Apache 2.0).
- Won the AsyncAPI Hackathon 2021 ($5,000 prize) with the Bundler, donated the project to AsyncAPI and still maintain it.

Backend Developer Intern, Mage | *Remote* Nov 2021 – Feb 2022

- Built Go microservices that provision cloud infrastructure from templated configuration and a React low-code interface for them, deployed with Docker and Kubernetes.

Projects

angelone-mcp | TypeScript, Bun, MCP, WebSockets github.com/Souvikns/angelone-mcp

- MCP server that connects Claude to an Angel One brokerage account (portfolio, orders, live market data), used on a live account to automate investing. Orders Claude proposes are sent only after approval in a desktop dialog.

JobScout | Go, Wails, LLMs, scrapers findmejob.xyz

- Cross-platform desktop app that uses LLMs to match job seekers with openings. Wellfound and Greenhouse scrapers refresh listings every 6 hours into a self-hosted backend.

Kitsu | LangGraph, deepagents, GitHub Actions github.com/Souvikns/kitsu

- AI agent packaged as a GitHub Action that reviews pull requests and writes summaries. It filters out low-signal changes and chunks large diffs to fit context limits.

Education

Chandigarh University | B.E., Computer Science Jun 2018 – Jun 2022