Shivam Malpani

Full Stack AI Engineer
Bengaluru, India | mr.shivammalpani@gmail.com | +91-9979173770
LinkedIn | GitHub

Summary

Software Engineer II and former founding engineer with 3+ years building production SaaS and AI-enabled products. Experienced across React, TypeScript, Python, Node.js, and PostgreSQL, with end-to-end ownership of user interfaces, APIs, asynchronous pipelines, and LLM integrations. Delivered measurable improvements in product adoption, developer productivity, latency, and cost.

Experience

Karbon Business Bengaluru, India
Software Engineer II *December 2025 – Present*

- Built and launched the **Accounts Receivable module** across frontend and backend, covering AI-assisted invoice creation, allocation matching, and Tally synchronization; reached **200+ uploads and 50+ active users** in its first week.
- Rebuilt invoice ingestion for **40,000+ row files** using streaming parsing and chunked asynchronous processing; paired it with a virtualized review grid, inline validation, and bulk correction.
- Built **Amazon SQS-based workflows** for asynchronous processing and service integrations, including Tally communication and accounting-data synchronization.
- Owned end-to-end architecture of **FlowTax**, a full-stack CPA tax-filing platform with agent-driven orchestration for multi-step return processing and client management.
- Built **Shadow PM**, a multi-agent platform integrating Jira, customer transcripts, and codebase analysis for product research and UI prototyping, with **Temporal** orchestration, **FastAPI**, and **LangSmith** tracing.

Kusho.ai - Salesmonk.ai Bengaluru, India
Founding Software Engineer *August 2023 – November 2025*

- Architected an **AI-powered API testing platform** with OpenAPI and Postman integrations, achieving **95% API coverage** and reducing test-authoring time by **75%**.
- Developed a **cross-framework middleware SDK** capturing live API traffic for anomaly detection and automated test generation, reducing manual debugging effort by **60%**.
- Built a **RAG-based knowledge pipeline** connecting transcript ingestion, Cloudflare R2, and AutoRAG to retrieve historical test runs and failure patterns for investigations.
- Optimized production LLM services with request batching, response caching, and prompt trimming, reducing **time to first token by 50%** and **per-request token spend by 30%**.
- Implemented real-time model-consumption tracking, pre-flight cost checks, and threshold alerts to monitor usage and control spending across the AI testing pipeline.

Technical Skills

Languages: TypeScript, JavaScript, Python, SQL
Frontend: React, Next.js, React Hook Form, Zod, TanStack Query
Backend: Node.js, NestJS, Django, Django REST Framework, FastAPI, REST APIs, GraphQL, Webhooks
AI Engineering: LLM APIs, LangChain, LangGraph, RAG, Structured Outputs, Tool Calling, Embeddings, Vector Search
Data and Messaging: PostgreSQL, Redis, MongoDB, Elasticsearch, pgvector, Amazon SQS
Cloud and Quality: AWS, GCP, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Jest, Playwright

Projects & Open Source

DocuSage — Document Knowledge Platform *Spring Boot, PostgreSQL, Docker*

- Built semantic search across document pipelines, integrating LlamaCloud processing, vector retrieval, pipeline management, and scheduled synchronization.

StackHub — Git Workflow Extension Open VSX | GitHub

- Built and published a VS Code/Cursor extension for stacked branches and pull requests, with a visual branch tree, automated restacking, conflict resolution, and GitHub review/CI status integration.

Education

Indian Institute of Information Technology Vadodara Gandhinagar, India
B.Tech in Computer Science; CPI: 8.26 *August 2019 – May 2023*