

SARVAGYA SAXENA

AI Systems Engineer

Bengaluru, India | +91 8630444673 | sarvagyasaxena.2102@gmail.com | sarvagyasaxena.tech
github.com/serverf21 | linkedin.com/in/sarvagya-saxena-a8740195

AI Systems Engineer who ships production AI systems, not demos. Own a production Agentic AI platform at Walmart serving 50,000+ daily active users, plus the RAG, MCP server, and evaluation infrastructure behind it. Founder of ComplyMint, a live AI compliance SaaS. 5+ years across AI, distributed systems, and platform engineering; led teams of up to 8.

TECHNICAL SKILLS

AI Systems & Agents: Agentic AI, Multi-Agent Orchestration, LangGraph, LangChain, Model Context Protocol (MCP) Servers, Tool / Function Calling, Prompt Engineering, OpenAI APIs, Anthropic Claude API, Ollama, FastAPI, LangServe

Retrieval & LLMops: Retrieval-Augmented Generation (RAG), Semantic Search, Embeddings, Hybrid Search (BM25 + Dense Retrieval), Reranking, pgvector, Pinecone, FAISS, Chroma DB, LangSmith, RAGAS, Eval Gating in CI/CD, Token & Cost Optimization

ML & Computer Vision: PyTorch, TensorFlow, Convolutional Neural Networks (CNNs), OpenCV, OpenAI Vision

Backend & Distributed Systems: Python, Node.js, Nest.js, Express.js, Java, Microservices, CQRS, RESTful APIs, GraphQL, Kafka, RabbitMQ, WebSockets

Infrastructure & Data: Kubernetes (K8s), Docker, Helm, Jenkins CI/CD, AWS, Azure, GCP, PostgreSQL, Google BigQuery, Snowflake, MongoDB, Supabase

Frontend: React.js, Next.js, TypeScript, React Native, Micro-Frontends

PROFESSIONAL EXPERIENCE

Walmart Global Tech India

December 2023 – Present

Senior Full Stack Developer | Bengaluru, India (Remote)
(via CA-One Tech Cloud)

- Designed and deployed a production Agentic AI system routing natural-language queries to REST APIs and auto-generating dashboard visualization widgets — cut dashboard build time from 2–3 days to 2–3 minutes for a platform serving 50,000+ daily active users; validated under concurrency testing to 50,000 QPS.
- Architected an AI-powered warehouse audit pipeline replacing fully manual inventory counts across 100+ facilities — OpenCV image preprocessing, AI-driven carton counting with Ti-Hi pallet inference (OpenAI Vision), and automated reconciliation against lodged quantities — reaching 85% audit accuracy.
- Engineered an LLM-powered test automation framework (Playwright + OpenAI) generating executable test suites from natural-language specs; adopted by 3 product teams, improving QA velocity 60%.
- Architected 29 production Nest.js APIs (CQRS) with Kafka event streaming for real-time order monitoring across a supply chain processing 15M+ events/day; led legacy Java-to-NestJS migration cutting API runtime 50% and p95 latency 12–15%.
- Owned Kubernetes platform operations — pod tuning and autoscaling, runtime error handling, multi-environment provisioning, and load balancer / DNS — plus a deployment evaluation pipeline gating releases on health and performance checks. Led a team of 5 engineers end to end on a logistics application.

Sears Global Technologies India

June 2022 – July 2023

Software Developer | Bengaluru, India (Remote)

- Built a GPT-4 and LangChain-powered retail analytics chatbot converting raw transactional data into business insights via GraphQL APIs and GCP-backed vector storage; served 5,000+ internal users.
- Delivered an end-to-end NFT ticketing marketplace (React, TypeScript, Solidity); cut platform transaction costs 50% by migrating payments to Polygon with MetaMask.

GreyB Research Private Limited

November 2020 – June 2022

Technology Consultant | Mohali, India

- Delivered AI-driven client applications automating analytics workflows — reduced manual reporting effort 80% and client consultancy spend up to 50%; mentored 10 junior engineers.

PROJECTS & OPEN SOURCE

ComplyMint · Founder / Sole Engineer · comply-mint.cloud — Live multi-tenant AI SaaS for India's Digital Personal Data Protection (DPDP) Act — LLM-powered vulnerability analysis, fine prediction, and privacy-notice analysis, with Shopify App Store and Razorpay integrations. Shipped solo.

RAG Lens · github.com/serverf21/rag_lens — RAG evaluation system (LangSmith + RAGAS) measuring retrieval quality, hallucination rate, and answer faithfulness; wired into CI/CD as a regression gate so quality drops block deploys.

Zerodha Claude MCP Server · github.com/serverf21/zerodha-mcp-claude — Model Context Protocol server (TypeScript, Bun) built in 2025 while MCP tooling was still early — bridges Claude with the Zerodha Kite API for real-time portfolio analysis and trade execution over WebSockets.

EDUCATION

National Institute of Technology Silchar — B.Tech, Mechanical Engineering

2016 – 2020

Final year project: CNN-based self-driving prototype deployed on Raspberry Pi, ~88% real-world accuracy (TensorFlow, OpenCV).