SARVAGYA SAXENA

AI Systems Engineer

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

SUMMARY

AI Systems Engineer with 6 years building and shipping production AI platforms at enterprise scale (Walmart, Sears). Architect production Agentic AI, RAG, and Model Context Protocol (MCP) systems end to end in Python — orchestration, retrieval, evaluation, deployment, and the distributed backends they run on. Most recently owned an agentic analytics platform serving 50,000+ daily active users at Walmart, the evaluation infrastructure behind its AI systems, and a CQRS order service handling 15M+ events/day across Python, Java, and Node.js runtimes. Led teams of up to 8 engineers across AI, backend, and platform work.

TECHNICAL SKILLS

Languages: Python, TypeScript / JavaScript, Java, SQL

AI Systems & Agents: Agentic AI, Multi-Agent Orchestration, LangGraph, LangChain, Model Context Protocol (MCP) Servers, Tool / Function Calling, Prompt Engineering, Reinforcement Learning from LLM-Generated Reward Functions, 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, Agent & RAG Evaluation, Eval Gating in CI/CD, Token & Cost Optimization

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

Backend & Distributed Systems: Python, FastAPI, Java, Spring Boot, Node.js, Nest.js, Express.js, Microservices, CQRS, Backend-for-Frontend (BFF), Transactional Outbox, Event-Driven Architecture, Kafka, RabbitMQ, REST, GraphQL, WebSockets, Idempotency, Eventual Consistency

System Design & Scaling: Distributed System Design, Read/Write Segregation, Database Partitioning & Sharding, Consumer Groups, Edge & Client-Side Caching, Dead-Letter Queues, Backpressure, Latency & Throughput Optimization

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

Cloud & DevOps: Azure, AWS, Kubernetes, K8s Dashboards, Global Server Load Balancers

Frontend: React.js, Next.js, TypeScript, React Native, Angular, Micro-Frontends, Server-Side Rendering (SSR), Redux, Tailwind CSS

Leadership: System Architecture, Technical Design Reviews, Team Leadership (5–8 engineers), Mentoring, Agile / Scrum, Cross-Team Delivery

PROFESSIONAL EXPERIENCE

Walmart Global Tech India **December 2023 – August 2026**

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

- Designed and deployed the production Agentic AI layer of Dynamic Dashboard Creation App in Python (FastAPI + LangGraph), routing natural-language queries to REST APIs, retrieving structured database records, and auto-generating dashboard widgets — cut dashboard build time from 2–3 days to 2–3 minutes for a platform serving 50,000+ daily active users; validated to 50,000 QPS under concurrency testing.
- Designed and built an internal evaluation platform for AI systems — performance benchmarking and RL-based accuracy testing across agents, RAG pipelines, and autonomous forklifts — using LLM-generated reward functions (Eureka-style reward design) refined on every iteration to optimize agent behaviour.
- Architected an AI-powered warehouse audit pipeline replacing fully manual inventory counts across 100+ facilities — Python / OpenCV image preprocessing, AI-driven carton counting with Ti-Hi pallet configuration inference (OpenAI Vision), and automated reconciliation against lodged quantities — reaching 85% audit accuracy.
- Built a Python-based RAG observability pipeline over Kubernetes and cloud-console logs, retrieving against historical failures and runbooks to diagnose CI/CD and cluster incidents in real time — surfacing probable cause and suggested fixes instead of manual log triage.
- Engineered an LLM-powered test automation framework (Python, Playwright + OpenAI) generating executable test suites from natural-language specs; adopted by 3 product teams, improving QA velocity 60%.
- Led CQRS re-architecture of a 29-endpoint order service, segregating write-heavy and read-heavy traffic onto separate runtimes and scaling paths to hit latency targets neither could reach alone.
- Optimized the Java / Spring Boot write path by keeping only the one or two primary database writes synchronous and publishing all eventually-consistent state updates through a Kafka transactional outbox; partitioned the database along US regional cluster boundaries to cut cross-region round-trips — reducing order-update API response time from 1–5s to under 500ms under heavy CRUD load, with at-least-once delivery and no dual-write inconsistency.
- Moved read-heavy endpoints behind a Node.js / NestJS backend-for-frontend with edge and client-side DNS caching — cutting API runtime 50% and p95 latency 12–15%; defined Kafka partitioning strategy, consumer group topology, and dead-letter handling across a supply chain processing 15M+ events/day.
- Built a micro-frontend architecture (React + Angular) on a dynamic shell application, enabling 50+ teams to ship independently deployed, composable dashboard widgets; delivered SSR templates dynamically from blob storage based on workflow state.
- Owned Kubernetes platform operations — pod tuning and autoscaling, multi-environment provisioning, load balancer / DNS — and led a team of 5 engineers end to end, from product discovery through production deployment, on a logistics application.

Sears Global Technologies India **June 2022 – July 2023**

Software Developer | Bengaluru, India (Remote)

- Hands-on experience deploying applications on AWS EC2 via remote SSH; built enterprise Agentic AI apps on AWS Bedrock.
- Built Legal Data Analyzer, a GPT-4 and LangChain-powered analytics application on a Python backend, converting raw data into business insights via Node.js GraphQL APIs and GCP-backed vector storage, serving 5,000+ internal users.
- Developed the Python data and embedding pipeline behind it — document and record ingestion, cleaning, and chunking into vector storage — so retrieval ran against current, well-formed context rather than raw exports.
- Scaled the payments platform 10x — from 5,000 to 50,000 transactions/day — by re-architecting settlement on PayPal APIs with a double-entry ledger and enforced idempotency keys, eliminating duplicate charges and reconciliation drift under retries and concurrent requests.
- Delivered an end-to-end NFT ticketing marketplace (React, TypeScript, Node.js / Express.js, Solidity) supporting on-chain publishing, auctions, and secondary sales.

GreyB Research Private Limited **November 2020 – June 2022**

Technology Consultant | Mohali, India

- Led architecture for client-facing analytics applications (React, Next.js, React Native) — service boundaries, REST API contracts, data models, caching strategy — reducing manual reporting effort 80% and client consultancy spend up to 50%.
- Mentored 10 junior developers across web engineering, Node.js and Python automation, and early LLM-based solutions, raising code quality and review standards across the team.

SELECTED PROJECTS & OPEN SOURCE

ComplyMint — DPDP Compliance SaaS Platform · *Founder / Sole Engineer · comply-mint.cloud*

- Live multi-tenant SaaS platform for India's Digital Personal Data Protection (DPDP) Act compliance — automated gap assessment, compliance dashboard, privacy notice generator, and Data Principal Rights (DPR) request management workflow.
- Ships LLM-powered compliance intelligence — automated vulnerability analysis of client data practices, regulatory fine prediction, and a privacy-notice analyser flagging non-conformant clauses; owns the full stack solo, including per-tenant data isolation, Shopify App Store integration, and Razorpay-backed subscription flows.

Shipyard — Multi-Agent Portfolio Operating System

- Portfolio OS that lets a solo developer operate and build several products at once, ranking apps by net margin × autonomy so effort compounds instead of scattering.
- Multi-agent execution layer built on a propose-don't-decide model — cheap task-specific models handle extraction and research while expensive reasoning is reserved for code — cutting per-task cost while keeping every decision human-approved.

RAG Lens — Agentic RAG Evaluation Pipeline · *github.com/serverf21/rag_lens*

- Python evaluation system (LangSmith + RAGAS) benchmarking retrieval quality, hallucination rate, and answer faithfulness across multiple document corpora; wired into CI/CD as a regression gate so retrieval quality drops block deploys.

Zerodha Claude MCP Server · *github.com/serverf21/zerodha-mcp-claude*

- Modular Model Context Protocol server (TypeScript, Bun) built in 2025, while MCP tooling was still early — bridges Claude with the Zerodha Kite brokerage API for real-time portfolio visualization, trade execution, and analytics over WebSockets with schema-validated tool definitions.

GenomeAI — Genomic Analysis Platform · *In development (private)*

- Python / FastAPI service applying transformer models to predict medication repurposing for gene-linked disorders; extending toward disease-risk forecasting from combined genomic, health, and dietary inputs.

Technical Writing — TechTunes Blog *(Next.js + Sanity)*: engineering deep dives on production RAG pipelines, retrieval evaluation, and MCP server design. sarvagya-nextsanity-blog.vercel.app

EDUCATION

National Institute of Technology Silchar **2016 – 2020**

B.Tech, Mechanical Engineering | Silchar, Assam, India

- Final year project: autonomous driving prototype — CNN-based steering prediction with road-sign interpretation, deployed on Raspberry Pi at ~88% real-world accuracy (Python, TensorFlow, OpenCV).

LEADERSHIP

- General Secretary, Students' Union Body, NIT Silchar (2019–20) — managed an annual events budget of ₹1 crore (~USD 120K) across the institute's student organizations.
- Chief Coordinator, Incandescence Cultural Festival; Convener, SPIC MACAY Chapter, NIT Silchar.