Chitransh Saxena

Senior Staff Software Engineer — GenAI & Distributed Systems

chitransh033@gmail.com | LinkedIn | GitHub | LeetCode | Portfolio

Senior engineer with **3+ years building production GenAI systems** — RAG, embedding pipelines, and synthetic data generation — on enterprise-scale distributed infrastructure. **Avid Claude Code user (1.5+ years)**, shipping rapid AI-assisted POCs across both enterprise and personal projects.

TECHNICAL SKILLS

Languages: Java, Python, C++, Go, SQL, NoSQL

GenAI & ML: Retrieval-Augmented Generation (RAG), Synthetic Data Generation, LLM Fine-tuning & Extended Pre-training, Embedding Models, Semantic & Hybrid Search, Dense/Sparse Retrieval (BM25, SPLADE), Re-ranking, Multimodal Ingestion, Prompt Engineering, Agentic Coding (Claude Code), Model Context Protocol (MCP)

Vector & Search: MilvusDB, OpenSearch, Elasticsearch, Faiss, ANN Indexing (HNSW, IVF, Quantization)

Streaming & Messaging: Apache Kafka (Consumer Groups, Partitioning, Exactly-Once Semantics, Offset Management), Redis Streams, Dead Letter Queues

Distributed Systems: Event-Driven Architecture, Stream Processing, Microservices, High Availability, Observability, SRE Practices, Design Patterns, TDD, CI/CD

Infrastructure: Kubernetes, Docker, Helm, Azure Cloud, MongoDB, Grafana, Prometheus, Jaeger

PROFESSIONAL EXPERIENCE

IBM India Software Labs — Senior Staff Software Engineer

Aug 2024 – Present | Bangalore

Configurable SaaS Embedding Engine & RAG Platform — watsonx Code Assistant for Z (Bob-IDE)

Workstream 1 — Embedding Ingestion & RAG

- Architected and led (team of 4 developers and 2 testers) an end-to-end embedding engine and Retrieval-Augmented Generation (RAG) pipeline processing **8K+ code pairs, 50+ documents (500+ pages each), and 250+ runtime examples** with multimodal ingestion across multiple embedding models.
- Scaled the document ingestion pipeline across **3 message-queue streams and 4 decoupled services** (orchestrator, PDF-processor, chunker, embedding generator) with dual-store persistence in Elasticsearch and OpenSearch for hybrid retrieval.
- Led in-depth investigation of BM25 and sparse embeddings and integrated **SPLADE** for learned sparse retrieval, improving **Recall@20 by 35%**.
- Designed production-grade MilvusDB vector infrastructure achieving **<900ms inference** via optimized ANN indexing (HNSW with quantization, IVF) and semantic search for enterprise-scale code assistance.
- Built Redis Streams infrastructure (consumer groups, DLQ, auto-scaling, Grafana observability) at **sub-100ms latency and 99.95% delivery**; developed a novel LSP-based chunking algorithm over **10K+ code artifacts** exposed through custom MCP tools (Python, MongoDB, GraphQL).
- Re-architected a monolith with Java Generics and the Strategy pattern for **3+ extensible language pairs (90% reuse)** and **100% throughput gain** via JVM profiling and 8-core parallelization; pipeline adopted by **4+ teams**, cutting development time **60%** — awarded Star of the Month within 2 months.

Workstream 2 — Synthetic Data Generation

- Built a synthetic data-generation pipeline producing **50K training code pairs from a 5K base (10x augmentation)** through pattern-level augmentation to fuel LLM extended pre-training.
- Applied a TDD methodology to validate generated pairs, enforcing syntactic and behavioral correctness across the augmented corpus and raising downstream training-data quality.

Walmart Global Tech India — Software Engineer 3

Oct 2021 – Jul 2024 | Bangalore

- Designed a fault-tolerant, event-driven stream-processing platform sustaining **1.5M transactions/day at 50 TPS** on a 3-broker Kafka cluster (RF=3) with idempotent producers, exactly-once processing semantics, consumer-group coordination, and dead-letter-queue isolation, holding a **99.9% availability SLO**.
- Tuned the consumer topology to cut **p99 consumer lag 2s → 0.5s** via partition-rebalancing strategy, batch-size and fetch tuning, and offset-commit optimization; collapsed 15+ consumer queries into 3 with composite indexing.
- Compressed the wire format **67% (150 → 50 bytes)** and eliminated 2 single points of failure through hot-path database indexing on 500K+ records and topology hardening.
- Engineered an extensible, plugin-based stream-processing framework handling **50+ event types across 15+ topics** using Strategy, Factory, and Decorator patterns with Resilience4j circuit breakers, bulkhead isolation, exponential-backoff retries, and async error queues for graceful degradation.
- Operated **12+ horizontally autoscaled microservices** on Kubernetes (Helm, HPA, rolling deployments) and built a MERN-stack control plane that automated GitHub release workflows, compressing manual effort from **7–10 days to 45 seconds** across 20+ repositories.
- Drove SRE-led P1 incident response — root-cause analysis via centralized log aggregation, distributed tracing (Jaeger), and metric correlation — reducing **MTTR by 40%**, shrinking incident blast radius, and sustaining zero-defect releases.

NielsenIQ — Software Engineer

Jul 2019 – Oct 2021 | Chennai

- Decomposed 3 monolithic services into a horizontally scalable, distributed microservices architecture with leader-election–based coordination, eliminating 2 single points of failure and lifting throughput **60% (100 → 160 RPS)**.
- Re-engineered latency-critical C++ data pipelines for a **55% runtime reduction (20s → 9s)** through CPU/memory profiling, memory-leak elimination, and hot-path restructuring across 15K+ LOC, while owning on-call rotations for production debugging, log analysis, and customer-escalation resolution.

EDUCATION

B.Tech in Information Technology — 88% SRM Institute of Science and Technology | 2015 – 2019