

Dhruvin Rupesh Soni

Research / ML Engineer (Applied & Infra) | Agentic R&D, LLM Evaluation, ML Systems | Math + Distributed-Systems Foundation

dhruvinrsoni@gmail.com • linkedin.com/in/dhruvinrsoni • github.com/dhruvinrsoni • dhruvinrsoni.github.io • Pune, India | Open to relocation

SUMMARY

Software engineer with 5+ years building production ML-adjacent systems, now operating at the applied-research edge of GenAI: agent infrastructure, LLM benchmarking, and Text2SQL/RAG groundwork. Strong mathematical foundation (First Rank in Mathematics; Number Theory, Parallel Computing, Graphs & Networks coursework) paired with hard experimentation rigor – benchmarked leading LLMs into a team-wide model-selection standard, shipped MCP-based agent infra, and authored OSS with 1,233+ tests. Builds the unglamorous infrastructure and evaluation harnesses that make research reproducible and shippable.

FOCUS

Research interests: agent infrastructure (MCP, tool-use orchestration), LLM evaluation & benchmarking, RAG / Text2SQL, and ML systems / distributed compute.

EXPERIENCE

Senior Software Engineer — Zebra Technologies

Nov 2022 – Present

Pune, India (Hybrid)

- Authored a Comparative Analysis of GenAI Strategies benchmarking leading LLMs across cost, latency, and output quality; the evaluation became the team's standard for model selection and prompt engineering.
- Built Model-Context-Protocol (MCP) integration over an enterprise DB2 schema, establishing secure groundwork for natural-language (Text2SQL) analytics over production data.
- Architected an AI Scrum-Master automation suite (sprint-health dashboards, auto-retrospectives, semantic ticket-deduplication), delivering enterprise-grade AI at \$0 incremental cost by ethically reusing existing authenticated LLM sessions.
- Engineered custom GitHub Copilot agents (quality standards codified as Markdown SOPs) that autonomously generated 33,586 lines of test code in 7 days – ~80x faster than manual – with zero production incidents.
- Re-architected a monolithic batch god-class into decoupled microservices (orchestrator pattern) and migrated legacy VM cron-jobs to Google Cloud Run under release pressure, with zero critical regressions across three releases.

Software Engineer II — Zebra Technologies

Sep 2020 – Nov 2022

Pune, India

- Built Jenkins-on-GCP CI/CD with GKE + Helm and self-service pipelines; led SVN→Git and MongoDB→Atlas migrations off legacy infrastructure.
- Delivered first-in-company Google Secret Manager integration (atomic rotation, least-privilege IAM) and a Python MongoDB query utility that cut manual effort ~95%.

Software Developer — Reflexis Systems (acquired by Zebra Technologies)

Jul 2019 – Sep 2020

Pune, India

- Built the backend for an AI performance-management product (Java, Spring Boot, MongoDB, REST, Spring State Machine), shipping production features managing thousands of records.
- Drove rollout quality across a 111-client deployment: resolved 100+ issues at a 45% defect-burndown rate, lifting client satisfaction.

PROJECTS

SmrutiCortex *TypeScript, Chrome MV3, Ollama (local AI)*

Privacy-first browser history-search extension with optional local LLM inference; engineered with experimentation rigor – 1,233 unit + 45 E2E tests and 9 scoring/eval functions. Published to the Chrome Web Store.

Samvada Studio *React 18, TypeScript, Vite; 6 LLM providers*

Multi-provider LLM chat UI unifying 6 providers behind one interface with 35+ features and in-memory key security; a hands-on testbed for cross-model evaluation. Launched on Product Hunt.

Agentskills Garden *AI agent infrastructure*

A hierarchy of 88 composable AI-agent skills exploring structured, reusable agent capabilities – part of 10+ public repos and an active engineering blog.

TECHNICAL SKILLS

Applied ML / GenAI: LLM evaluation & benchmarking, prompt engineering, agentic workflows, RAG, Text2SQL, Model

Context Protocol (MCP), GitHub Copilot agents, Vertex AI, Ollama (local inference)

ML Systems & Infra: Microservices (orchestrator pattern), Google Cloud Run, GKE, Helm, Jenkins CI/CD, readiness probes / zero-downtime rollouts, SLA alerting & self-healing pipelines

Math & CS Foundation: Number Theory, Parallel Computing, Graphs & Networks, Data Mining, Computer Architecture; First Rank in Mathematics (school batch)

Languages: Python, Java, TypeScript, JavaScript, SQL, Shell

Backend & Data: Spring Boot, REST, Spring State Machine, MongoDB / Atlas, DB2, MySQL, Cloud SQL, BigQuery, Angular, React 18 + Vite

Experimentation & Quality: Test harness design, scorer/eval frameworks, automated test generation, Trivy, observability, reproducible benchmarking

Certifications: Google Cloud: AI Agents (2025), Vertex AI Prompt Design, BigQuery, Core Infrastructure; Microsoft VS Code Agent Mode (2025); GCP Skills Boost Gold League (22,395 pts)

EDUCATION

B.E. (Hons.), Computer Science

2015 – 2019

BITS Pilani, K. K. Birla Goa Campus

Coursework: Data Mining, Parallel Computing, Number Theory, Graphs & Networks, Computer Architecture. First Rank in Mathematics (school batch).

CERTIFICATIONS & RECOGNITION

Certifications: Google Cloud – AI Agents Edition (2025); Google Cloud – OnBoard Edition (2026); Vertex AI Prompt Design; Derive Insights from BigQuery; GC Fundamentals: Core Infrastructure; Migrate MySQL to Cloud SQL; Microsoft – VS Code Agent Mode (2025)

Awards: 5-Year Service Award, Zebra Technologies (2024); Google Cloud Skills Boost – Gold League (22,395 pts) •

Languages: English (professional), Gujarati (native), Hindi