Harsh Verma

harsh376@gmail.com | linkedin.com/in/harsh376
github.com/harsh376 | harshverma.com

SUMMARY

Senior software engineer specializing in retrieval systems, AI evaluation, and production reliability. Owned evaluation for agentic retrieval services, implemented hybrid search, and established BenchSci's first SRE function. Sharing implementations and technical write-ups on backend systems, retrieval, and AI agents.

EXPERIENCE

BENCHSCI

Oct 2019 - Jun 2026

Senior Software Engineer — Dec 2022 - Jun 2026

- Owned evaluation for agentic retrieval services; curated **~600 questions** with domain experts and combined labeled data with LLM evaluators. Helped improve MAP@10 from **0.60 to 0.85** in six months.
- Parallelized evaluation with a producer-consumer pattern, cutting runtime **60%** (**90 to 36 minutes**) without increasing cloud costs.
- Implemented Elasticsearch hybrid search using keyword and kNN vector retrieval with reciprocal rank fusion (RRF) to improve relevance and recall.
- Upgraded self-managed Elasticsearch from v7 to v8 across local and GCP production environments with **zero production downtime**.

Senior Software Engineer / SRE Tech Lead — Oct 2021 - Dec 2022

- Established BenchSci's first SRE function, introducing incident response and partnering with application teams to define and track service-level indicators.
- Built synthetic checks and SLO-based alerts for critical, low-traffic services, reducing noise from sparse user traffic.

Senior Software Engineer — Dec 2020 - Oct 2021

- Built CDC and denormalization pipelines from PostgreSQL to Elasticsearch; designed parent-child queries to eliminate application-level joins and aggregations.
- Created a hermetic Bazel integration-test environment with PostgreSQL and Elasticsearch for isolated, reproducible tests.

Software Engineer — Oct 2019 - Dec 2020

- Migrated a Node.js API to Django and built reusable React components for a shared design library.

TOP HAT | Software Developer

Jul 2017 - Jun 2019

- Extended Auth0 authorization with granular custom roles; built an LTI tool provider integrating Blackboard, Moodle, and D2L.

EVENTMOBI | Software Developer Intern

May 2015 - Jul 2016

- Built event apps with Angular, React, Flask, and Node.js, including private chat for **~7,000 users** and live polling for **5,000 concurrent users**.

TECHNICAL SKILLS

Languages and applications: Python, JavaScript, TypeScript, Django, FastAPI, React

Search and AI evaluation: Elasticsearch, PostgreSQL, vector search, CDC, retrieval benchmarking, dataset curation, LLM-based evaluation

Infrastructure and reliability: Terraform, Kubernetes, Docker, GitHub Actions, Bazel, distributed tracing, synthetic monitoring, SLO-based alerting

EDUCATION

University of Toronto | Bachelor of Applied Science, Computer Engineering