

Ishan Sharma

SENIOR SOFTWARE ENGINEER · SURREY, BRITISH COLUMBIA, CANADA

ishansharma1320@gmail.com · LinkedIn · GitHub

Ishan Sharma is a Senior Software Engineer based in Surrey, British Columbia, Canada, specializing in production AI systems: low-latency LLM inference, retrieval-augmented generation, real-time speech pipelines, and privacy-preserving on-prem deployment. His operating principle: the model is the smallest part of the system – the deterministic execution, curated context, and trust architecture around it are what make AI dependable.

EXPERIENCE

2025 →

Senior Software Development Engineer

YACTRAQ ONLINE · SURREY

Own production agent-assist systems for contact centres: on-prem LLM summarization, real-time PII redaction, and a Ruby indexing pipeline reindexing under live traffic. Delivered sub-500ms inference at 200+ concurrent sessions on 2×H100 by tuning batching and concurrency, and built GPU instrumentation to catch degradation before it lands.

Inference is tiered by latency class: H100s where the deadline is half a second, a single consumer GPU where it is thirty minutes – the batch tier clears 3,500+ transcripts a day on one card, with VRAM budgeted across co-tenant workloads. Hardened that pipeline from a 30% error rate to zero by building its observability – journey logs, metrics, alerts – from scratch, and cut agent-assist evaluation latency from 7.5s to 1s with right-sized models. Model upgrades are gated by golden-set evaluation against human QA. Also ships a supervisor-facing analytics agent: one planning call compiled to a validated, deterministic execution plan, delivered as a React component inside the existing Angular dashboard. 10+ on-prem enterprise go-lives in 18 months, engineering live within a day.

2025 →

Founding Software Engineer

MAXCORE CLOUD · PART TIME

Led technical design of a decentralized container execution platform running workloads on GPU-backed DePIN networks. Defined execution flows, isolation boundaries and deployment primitives, and built orchestration to schedule and monitor across heterogeneous GPU environments.

MaxCore makes fragmented compute markets usable for reliable AI workloads: an opportunistic execution layer that evaluates available infrastructure against user-defined preferences – cost, geography, performance, reliability – and places each workload on the best option across containers, VMs and serverless, without provider lock-in. He designs the orchestration and resilience model: when low-cost capacity disappears, workloads fail over to a ready alternative while a replacement is provisioned in parallel, then rebalance back as preferred capacity returns. The work spans provider abstraction, workload portability, placement strategy, failover, recovery and state reconciliation – plus the Next.js platform, the React company site, and a 20-page GPU-infrastructure RFP with compliance matrix and bill of quantities.

2024

Information Technology Developer

CANADA REVENUE AGENCY · HALIFAX

Built citizen-facing Angular features including “Authorize My Representative,” and worked across Angular and Java services on the “File a Formal Dispute” workflow – accessibility, responsive behaviour and regression coverage for services taxpayers depend on.

Scoped business requirements directly with stakeholders, then implemented them end to end across the Angular front end and Java services.

2022 – 2024

Full Stack Software Engineer

YACTRAQ ONLINE · PART TIME

Built real-time ASR pipelines on Kaldi and NVIDIA NeMo with high-availability architecture on AWS and distroless Docker images. Led the Angular 14 migration of the BI dashboard, and shipped the PII redaction system, a conversational SQL query assistant, and the desktop audio capture framework.

Multilingual ASR pipelines with Python and MongoDB services behind them, run as highly available speech-processing deployments on AWS. The capture layer reads speaker attribution from physical device topology – agent on the microphone, customer on the sound card – so no diarization model sits in the path, and silence is signalled explicitly per channel rather than inferred from absence; roughly a third of customers run on this path today. The role grew from machine-learning and backend work into ownership across architecture, infrastructure and customer-facing applications – the foundation for leading the real-time Agent Assist and GenAI initiatives that followed.

2021

Data Engineer

DATAFACT.IO · BANGALORE

Server-side clickstream tracking for Shopify merchants on GCP with Tag Manager, Pub/Sub and containerized ETL – multi-tenant behavioural collection, plus async pipelines and CI/CD that cut manual setup per client.

SKILLS

LANGUAGES	Python · JavaScript/TypeScript · Ruby · Java
AI / ML	LLM inference & serving · RAG · agents & MCP · Kaldi · NVIDIA NeMo · NER/PII redaction · vLLM serving & quantization (AWQ, FP8 KV cache) · eval harnesses & golden sets
DATA & INFRA	Elasticsearch · MongoDB · Redis · Docker · Kubernetes · AWS · GCP · Node.js-orchestrated microservices
FRONTEND	React · Angular · Electron · Next.js

EDUCATION & CERTIFICATIONS

Master of Applied Computer Science – Dalhousie University, Halifax, Canada
B.Tech, Computer Science & Engineering – Manipal University Jaipur, India
Microsoft Azure Fundamentals (AZ-900)
Canadian permanent resident