

Yash Vardhan Sharma

Chicago, IL

yashvardhansharma2025@gmail.com | linkedin.com/in/ysharma6 | github.com/Parzivalart3mis | 312-776-9053 | [Portfolio](#)

Education

Illinois Institute of Technology

Master of Science in Computer Science, CGPA: 3.80/4.0

Chicago, IL

08/2024 - 05/2026

Vellore Institute of Technology

Bachelor of Technology in Computer Science Engineering, CGPA: 8.87/10.0

Chennai, India

07/2019 - 06/2023

Technical Skills

Languages: Go, Java, Python, TypeScript, JavaScript, SQL, Bash

AI/LLM: LangGraph, LangChain, RAG (hybrid retrieval), Qdrant, RAGAS, LangSmith

Frameworks/Libraries: FastAPI, Spring Boot, Node.js/Express, React/React Native, Next.js

Data: PostgreSQL, Redis, Kafka, MySQL, MongoDB

Cloud & DevOps Tools: AWS (ECS, Lambda, S3, VPC, ELB), Docker, Kubernetes, Terraform, ArgoCD, CI/CD

Professional Experience

Cashfree Payments

Bengaluru, India

Software Engineer

06/2023 - 08/2024

- Replaced **Jenkins** with **ArgoCD** for **GitOps-based** CD across **50+** microservices, cutting deploy time from **~25 min** to under **5 min** per service and shipping daily releases.
- Built a payment **webhook** delivery service in **Go** with **HMAC** signing, idempotency, and backoff retries + **DLQ**, raising delivery success from **97.2% to 99.8%** across hundreds of merchants.
- Designed risk blacklist/whitelist **APIs** in **Go/Postgres/Redis**, inline in the payment path at **18 ms p99**, blocking **12,000+** fraudulent transactions/month at under **0.4%** false positives.
- Reduced transaction latency from **450 to 290 ms** and scaled throughput to **10K RPS** by refactoring **Java** microservices with async multithreading on **Docker** and **Kubernetes**.

Cashfree Payments

Bengaluru, India

Software Engineer Intern

09/2022 - 05/2023

- Built a real-time reconciliation engine for payouts in **Java** with **Kafka**, **Redis**, and **PostgreSQL**, cutting failed payouts **28%** and reconciliation time **70%**.
- Provisioned **AWS** infrastructure (**VPC**, **ELB**, **S3**) as **Terraform IaC**, replacing manual setup with repeatable multi-environment deploys supporting **3x** traffic growth, on infra that held **99.9%** uptime.
- Built **CI** pipelines on **Bitbucket** with **Dockerized** builds and automated unit tests gating every PR, cutting post-release defects **40%** on new APIs.

Projects

GatewayLLM: Multi-Provider LLM Gateway with Semantic Caching | [Code](#)

07/2026

Stack: Go, Redis, Qdrant, PostgreSQL, OpenTelemetry, Prometheus, Grafana, Docker, Cloud Run

- Built an **OpenAI**-compatible **LLM** gateway in **Go** over **OpenAI**, **Groq**, and **Gemini** with policy routing, classified **retry/failover**, and a **Redis-Lua** circuit breaker shared across replicas.
- Engineered a two-tier cache (**Redis** exact-match, **Qdrant** vector at **0.95** cosine), serving exact hits at **0.42 ms p50 vs 128 ms** live calls and cutting spend **6x** on a **180-request** replay.

PayGuard: Agentic Compliance Co-Pilot for Payment Disputes | [Code](#) | [Demo](#)

04/2026

Stack: LangGraph, LangChain, Qdrant, FastAPI, PostgreSQL, Redis, LangSmith, RAGAS, AWS ECS

- Built a **LangGraph** multi-agent system over a **Qdrant** hybrid retrieval pipeline, hitting **0.91 RAGAS** faithfulness and cutting median draft time **38%** on a **200-case** benchmark.
- Deployed on **AWS ECS** with **PostgreSQL** checkpointing, human-in-the-loop gates, and RAGAS regression tests wired into **CI**.