

# Gagan Shivanna

Los Angeles, CA | (213) 571-7139 | [gshivann@usc.edu](mailto:gshivann@usc.edu) | [LinkedIn](#) | [Github](#)

## Education

### University of Southern California

*Master of Science, Computer Science*

• **Coursework:** Analysis of Algorithms, Database Systems, Deep Learning and Its Applications, Web Technologies, Applied NLP, Agentic AI

**Aug 2025 - Present**

*Los Angeles, USA*

### Visvesvaraya Technological University

*Bachelor of Engineering, Computer Science and Engineering*

• **Coursework:** Data Structures, Distributed System, Operating Systems, Machine Learning, Big Data Analytics

**Oct 2021 - May 2025**

*Bangalore, IN*

## Professional Experience

### Supraja Technologies

*Software Engineer Intern*

- Designed and implemented a backend service for an Automatic Malicious Website Blocker Chrome extension, processing 200+ malicious URLs and reducing page load latency by 30% through optimized request handling
- Collaborated with a team of 4 engineers to design and implement RESTful APIs using FastAPI and Python scripts, utilizing SonarQube static analysis to identify and resolve 15+ critical vulnerabilities, improving system reliability
- Optimized network request processing through asynchronous workflows and request batching, improving throughput and reducing redundant API calls across the URL classification pipeline

**Feb 2024 - Jul 2024**

*Hyderabad, India*

### Cranes Varsity

*Software Engineer Intern*

- Engineered Python-based data ingestion pipelines for IoT sensor integrations, streaming 100+ data points/sec to AWS Kinesis with zero downtime and reliable real-time processing
- Developed and deployed a leak detection system, reducing alert response times by 70%; implemented CI/CD pipelines to automate testing and deployment while resolving 15+ cloud synchronization issues and improving system reliability
- Spearheaded the development of an internal analytics dashboard that transformed real-time IoT sensor streams into operational KPIs, enabling engineers to detect anomalies and respond to incidents faster

**Mar 2023 - Jul 2023**

*Bangalore, India*

## Academic Projects

### Fault-Tolerant Distributed Storage System [Github](#)

**Jun 2026**

- Built a distributed peer-to-peer file system in Go with content-addressable objects, 3× replication, and automatic repair, ensuring high availability across a multi-node cluster
- Redesigned file transfers using io.Pipe/io.Copy streaming, reducing memory complexity from O(N) to O(1) and enabling ~1 GiB uploads on resource-constrained environments
- Validated system resilience through chaos testing on a 3-node Kubernetes cluster and monitored node health and throughput using Prometheus and Grafana, sustaining 22 MiB/s across 10 concurrent streams

### Hallucination-Aware Financial Search Engine [Github](#)

**May 2026**

- Engineered a financial document analysis pipeline over SEC 10-K filings (AAPL, MSFT, GOOGL, AMZN, NVDA), enabling accurate retrieval across 22K+ indexed vectors with a DeBERTa NLI hallucination detection layer that classifies financial claims as supported, contradicted, or unverifiable, reducing misinformation risk on high-stakes earnings data
- Leveraged Redis to serve frequently accessed document contexts, reducing repeated retrieval operations and improving system efficiency under concurrent query traffic

### Real-Time Session-Aware Recommendation System [Github](#)

**May 2026**

- Architected a real-time session-aware recommendation pipeline that fuses live user choices with historical embeddings, using a deep Two-Tower neural retrieval model (InfoNCE loss + hard-negative mining) and a feature-based MLP reranking layer
- Scaled candidate generation across a 62,000+ item catalog by integrating FAISS for Approximate Nearest Neighbor (ANN) search, serving the top-200 candidates to the ranking layer in under 200ms
- Reduced end-to-end recommendation latency from ~200ms to sub-50ms through Redis-backed embedding caching and retrieval optimization, enabling responsive real-time recommendations

## Technical Skills

- **Programming Languages / Databases:** Python, Java, C++, Go, JavaScript, SQL, Pytest, PostgreSQL, MongoDB, Pinecone
- **Backend & Frameworks:** FastAPI, Django, Node, React, GraphQL, Restful API
- **AI & Machine Learning:** LangChain, LangGraph, PyTorch, TensorFlow, HuggingFace TRL, scikit-learn, sentence-transformers, NLTK, OpenCV, Recommender Systems
- **Distributed Systems & Infra:** gRPC, Kubernetes, Docker, AWS (Kinesis, S3, Lambda), GCP, Redis, Kafka, CI/CD, OAuth

## Leadership & Achievements

- Winner, Origin Weekend Impact Hackathon (Google-sponsored), selected among top teams for high-impact solution
- Smart India Hackathon Grand Finale Winner: #1 in Blockchain domain (500 finalists; 50,000+ teams nationwide)
- Top 25 (of 400 teams), FKCCI Manthan Business Plan 2023 – drone-based blood delivery solution