Yash Hirani

yashhirani.in | yashhirani.0055@gmail.com | linkedin | github | Leetcode

Education

Vellore Institute of Technology — Vellore, India
B.Tech in Computer Science and Engineering (Data Science) — CGPA: 8.98/10 — *2023 – Present*

Sardar Patel School — Bhopal, India
Class XII (CBSE Board) — Percentage: 88.4% — *2021 – 2023*

Experience

Research and Development Intern — July 2025 – Present
Samsung R&D Institute — *Remote*

- **Led the architecture of a video-based benchmarking system** for Android camera KPIs using **OpenCV** and **Scikit-image**, successfully improving measurement precision by **40%** across **25+ flagship devices**.
- **Orchestrated an automated computer vision pipeline** leveraging **Keras** for stage detection and **Scikit-learn** for classification, which effectively reduced manual QA overhead by **150+ hours weekly**.
- **Engineered a high-performance video processing engine** utilizing **MoviePy**, **ImageIO**, and **NumPy** to extract and analyze frame-level data for competitive benchmarking against Samsung and global rival OEMs.

Full-Stack Developer Intern — May 2025 – July 2025
Sstudize — *Remote*

- **Revamped backend architecture**, boosting API speed by **192%** via pagination and query caching.
- **Reengineered database schema**, cutting latency by **40%** and enhancing multi-module performance.
- **Programmed LaTeX workflows**, reducing export time by **50%** while ensuring document consistency.

Projects

System Design Reviewer | *LangGraph, FastAPI, Celery, Redis, Next.js, Python* — 2026

- Architected a **multi-agent LLM orchestration system** using LangGraph and Python, coordinating **5 parallel domain-expert agents** with a **supervisor-judge reflection loop** and RAG via Qdrant vector DB on **140+ engineering principles** to reduce hallucinations.
- Engineered an **event-driven backend** using FastAPI, Celery, and Redis Pub/Sub with **SSE streaming** and a **sliding-window rate limiter**, decoupling LLM workloads from HTTP requests and ensuring **100% compliance** with third-party API rate limits **(15 RPM)**.
- Developed a **full-stack web application** using Next.js 14, React, and Tailwind CSS with real-time SSE streaming, **Mermaid.js architecture diagrams**, and **140+ Pytest/Jest tests** with AsyncMock for distributed Redis pipeline simulation.

ContextCapsule | *Next.js, Cloudflare Workers, Claude API, Chrome Extensions* — 2025

- Built **Context Capsule**, a Chrome Extension + Cloudflare Workers backend that compresses AI conversations into structured summaries, scaling to **2,000+ active users in under 2 months** with **70%+ token reduction** and zero paid acquisition.
- Devised a local-first **NLP summarization pipeline** (TF-IDF scoring, positional bias, cosine similarity deduplication) replacing LLM API cutting per-request cost to **$0** while maintaining **<5s p95 latency** at scale.
- Designed a **stateless, privacy-first backend** with SHA-256 IP hashing, rate limiting, 100K character payload validation, and word-entropy heuristics for abuse detection, ensuring zero data retention across all user requests.

Technical Skills

Languages: Python, TypeScript, JavaScript, C++, HTML, CSS, SQL
AI & ML: LangGraph, LangChain, RAG, Qdrant, FastEmbed, OpenCV, Keras, Scikit-image, Scikit-learn, NumPy
Frameworks: Next.js, React, FastAPI, Spring Boot, Tailwind CSS, Prisma, Astro,
Data & Cloud: PostgreSQL, Firebase, Redis, Google Cloud (GCP)
DevOps & Tools: Docker, Celery, ImageIO, MoviePy, REST APIs, LaTeX Automation

Achievements and Positions of Leadership

Chairperson – ISA-VIT: Led 50+ members, driving 20% growth in chapter engagement via 10+ events.
Security Researcher – Netflix: Reported session concurrency logic error; acknowledged by Netflix Security.