

Shikhar Saxena

saxena.shikhar1808@gmail.com | +91 87550 43412

linkedin.com/in/shikhar-saxena-90a484296 | github.com/Shikhar1808 | leetcode.com/u/ShikharSaxena6251

Education

Thapar Institute of Engineering and Technology, Patiala

Aug 2023 – Jun 2027 (Expected)

- B.E. Computer Science and Engineering | CGPA: 8.59/10

Army Public School, Bareilly

Class of 2023

- 12th: 92.4%, 10th: 93%

Experience

Voxomos Systems Pvt Ltd, Software Intern

Jun 2026 – Jul 2026

- Developed and benchmarked CPU, **CUDA**, and **HIP** implementations of cosine similarity and vector search workloads using NVIDIA Nsight Systems and Nsight Compute, exploring cross-platform GPU portability via **AMD ROCm**, **HIP**, and **HIPIFY**
- Evaluated inference performance across **vLLM**, **SGLang**, and **TensorRT-LLM**, isolating KV-cache memory limits under load; sustained **14–15 ms/token decode latency** and mitigated Time-to-First-Token (**TTFT**) spikes by optimizing request concurrency.

TruckHai, Backend Intern

Jan 2026 – Apr 2026

- Built Go (Gin, GORM) **REST microservices** for fleet dispatch, modelling the trip lifecycle as a **finite-state machine** with **ACID transactions** so that invalid state transitions are rejected and concurrent updates stay consistent.
- Removed race conditions and row-lock contention in **PostgreSQL** through **composite indexing** and query-plan tuning, and added **RBAC middleware** with audit-logged admin impersonation across **5+ user roles**.

Projects

CivicSentinel | *MERN Stack, Solidity, Pinata, TypeScript, LangChain, AWS*

github.com/Shikhar1808/HackMol-6.0

- Built an intake pipeline (Twilio, GPT-4, LangChainJS) that converts unstructured audio/text into validated **FIR** records with automated severity triage; anchored evidence hashes on **Ethereum** smart contracts and **IPFS** for a tamper-proof audit trail.
- Added OTP-based citizen authentication and background task queues; monitored API latency, uptime, and health with **Prometheus** and **Grafana**.

ZaiGuard — Surveillance System | *FastAPI, PostgreSQL, Qdrant*

github.com/Shikhar1808/Zaiguard-Connected

- Built a 5-layer **FastAPI** middleware (Redis TTL sliding windows, PostgreSQL B-Trees, Qdrant HNSW vector suppression) that reduced false-positive alarms by **65%–90%** in testing at sub-second latency.
- Used a **Transactional Outbox** worker to prevent PostgreSQL/Qdrant dual-write inconsistencies; coordinated cross-device **Flower federated learning** across LAN SuperNodes.
- Built edge video ingestion with motion-gated frame sampling and an RTSP/HTTP/USB fallback chain, feeding an **ONNX**-accelerated **YOLO** detector (**96.5% precision**) with **ByteTrack**.

Zenith Website | *NodeJS, ExpressJS, MongoDB, JWT Auth, NGINX, AWS*

github.com/Shikhar1808/ZenithCryptic

- Built a competition platform serving **1000+ concurrent users** at **sub-100 ms** response times; designed real-time leaderboard calculation and stateless **JWT** auth to remove database bottlenecks during submission spikes.
- Secured the deployment with **Cloudflare** DDoS mitigation and **NGINX** reverse-proxy rate limiting and load balancing.

Technologies and Skills

Languages: C++, C, Go, Python, JavaScript, TypeScript, SQL, Solidity

Backend & Web: Microservices, REST APIs, Node.js, Express.js, FastAPI, Gin, GORM, React.js, Tailwind CSS, Socket.io, JWT, RBAC

AI & GPU Computing: PyTorch, CUDA, AMD ROCm/HIP, vLLM, LangChain, LangGraph, MCP, OpenCV, Qdrant, Flower

Databases & DevOps: PostgreSQL, MongoDB, MySQL, Redis, Docker, Linux, GitHub Actions (CI/CD), NGINX, Prometheus, Grafana, Git, AWS

Awards and Certifications

SPHINX'24 Hackathon – 1st Runner Up, MNIT Jaipur

Nov 2024

NVIDIA Certified: Accelerated Computing (Modern CUDA C++, CUDA Python), AI on Jetson Nano, Generative AI & Deep Learning

2025 – 2026

LeetCode: 417+ problems solved, 1551 Contest Rating (DSA: arrays, graphs, DP, trees, stacks, queues, heap)