

Shivang Bhatnagar

Distributed Systems & Applied ML

✉ shivangb237@gmail.com



shivangb237



Shivb237



Portfolio

WORK EXPERIENCE

KFin Technologies Ltd.

06/2024 – Present

Software Engineer

- Led an AI proof of concept: **MCP** server for LLM access to MSSQL data plus a 3D avatar; earned CEO's letter.
- Deployed an event-driven **S3/SQS/Fargate** pipeline on legacy MSSQL data to speed up endorsement validation.
- Led and mentored 2 juniors migrating **10** read APIs to a common API layer on MF Central (2.5M users).
- Moved unit calculations from SQL procedures to a Node.js/TypeScript microservice, cutting peak load **30%**.
- Hardened **Loan Against Mutual Funds** APIs for VAPT security testing: request validation, rate limiting.
- Built Group SIP, a payroll-based investment platform: 30+ APIs, 95% coverage; MongoDB cut latency **40%**.

KFin Technologies Ltd.

02/2024 – 06/2024

Software Engineer Intern

- Re-engineered backend by replacing 8+ SQL procedures with Node REST APIs, reducing response time by **90%**.
- Established quality assurance by writing Jest unit tests, which reduced post-release defects by **55%**.
- Delivered real-time **Loan Against Mutual Funds** dashboard in React.js for clients and cross-team visibility.
- Authored OpenAPI (Swagger) documentation for **13+** API endpoints, streamlining developer onboarding.

Scogo Networks

05/2022 – 10/2022

Frontend Web Development Intern

- Engineered dynamic, data-driven email templates using **Handlebars.js** for personalized communications.
- Developed a Node.js automation tool converting CSV data to PDF reports, reducing processing time by **60%**.
- Streamlined asset tracking for CRIS (Centre for Railway Information Systems) via **automated tagging**.
- Integrated **modular** React components into an Electron-based desktop application, working with other teams.

PROJECTS

LLM Caching Gateway for RAG Pipelines

09/2026 – 10/2026

- Built a 4-layer LLM cache (exact, semantic, retrieval, KV) for RAG; semantic layer cut Claude API cost **39%**.
- Evaluated recall@k, F1, p50-p99 latency and wrong-answer rate on Qwen and Claude via a harness, **40+ runs**.
- Eliminated **stale** answers after data edits via data-version invalidation (29 of 30 stale before, 0 after).
- Counted local KV-cache reuse: **25-34%** of prompt tokens vs 2% cold; Claude prompt caching cost 18-25% more.
- Deployed a 1.62 GB arm64 Docker image to **AWS ECS Fargate** with API-key auth, spend cap and health check.
- **Technology Stack:** TypeScript, Node.js, SQLite, Docker, AWS (ECS Fargate), Claude API, Ollama, Vitest

Lucida: Explainable Fashion Search

10/2026 – Present

- Dropped the 36 s AI-image step: text search takes **0.3 s** and ranked products 2.6x better (20 queries).
- Added failover (0 errors, one server down) and load shedding: 4 req/s at 32 parallel requests, was 0.
- Engineered **idempotent** SQS ingestion with retries & DLQ replay; systemd restarts failed services in <8 s.
- Defined **4** AWS CDK stacks (IaC) with CloudWatch request tracing and 10 alarms; clean teardown, 0 left.
- Shipped on **Cloud Run** (Firestore) with rate limits, SSRF allow-list and a safety guard; 360 CI tests.
- **Technology Stack:** Python, TypeScript, AWS (EC2, SQS, CDK), Google Cloud Run, SigLIP, Qwen3, Ollama

SKILLS

- **Front-End:** HTML, CSS, JavaScript, React.js, Next.js, Redux, React Three Fiber
- **Back-End:** Node.js, Express.js, NestJS, FastAPI, TypeScript, Python, REST APIs, Microservices, OpenAPI
- **AI/ML & LLM Systems:** RAG, LLM Evaluation, LLMops, Embeddings, Semantic Caching, MCP, Claude API
- **Local LLMs:** Ollama, LM Studio, PyTorch, Hugging Face Transformers
- **Databases:** SQL (Microsoft SQL Server, MySQL), MongoDB, DynamoDB, Firestore, SQLite
- **Cloud & DevOps:** AWS (S3, SQS, ECS Fargate, EC2, CloudWatch, CDK), GCP (Cloud Run, Cloud Build), Docker
- **Testing & CI/CD:** Jest, Vitest, pytest, GitHub Actions

ACHIEVEMENTS

- **Letter of Appreciation** from Sreekanth Nadella, MD & CEO, KFin Technologies Ltd for AI POCs initiatives
- **Smart India Hackathon 2022** – Grand Finale (NIT Silchar), Team lead
- **CodeBite 2023** – **Winner**, Web Development Hackathon (Bhopal) competed against 100+ teams
- **Design Showcase** – **4th Place**, Sheryians Coding School, created responsive UI/UX prototype (80+ devs)
- **Hacktoberfest 2022 & 2023** – Open-Source contributor, successfully got 10+ feature pull requests merged

EDUCATION

B.Tech Computer Science Engineering

Lakshmi Narain College of Technology & Science

06/2024 | Bhopal, India

GPA: 3.4/4