Rishit Verma

mail.rishit.verma@gmail.com | linkedin.com/in/rish-kun | github.com/rish-kun

Education

Birla Institute of Technology and Science, Pilani — Pilani, India
B.E., Computer Science Engineering; CGPA: 8.04 — *2024 – 2028*

India International School in Japan (CBSE) — Tokyo, Japan
Class XII; Score: 96.4% — *2024*

National Public School (CBSE) — Chennai, India
Class X; Score: 96.6% — *2022*

Experience

Software Engineering Intern — May 2026 – Jul 2026
Vitalize Healthcare — AI-powered hospital management platform digitizing end-to-end hospital operations

- Built the backend within a **NestJS TypeScript monorepo**, storing data in a **FHIR-compliant schema**
- Integrated the app as a **Unified Health Interface** node on India's **ABDM** network (**1B+** records, **350K+** facilities)
- Containerized backend services with **Docker** on **AWS EC2** and deployed the frontend on **Cloudflare Workers**
- Shipped **20+ frontend screens** by orchestrating **coding agents** that enforce the design system via **skills + MCP**

Projects

Semantic LLM Response Cache | *C++, Systems Programming* — Jun 2026 – Jul 2026

- Built an **in-memory semantic cache** in C++ that embeds prompts locally and reuses responses for paraphrased queries
- Implemented **vector similarity search** over a flat, contiguous store with a **SIMD**-vectorized dot product at query time
- Enforced a memory budget with **cost-aware eviction**, persisting to a **binary snapshot** so restarts reload a warm cache

Hostel Allocation Portal | *Express.js, MongoDB, Algorithms* — Jun 2026 – Jul 2026

- Modeled the **official BITS** hostel allocation for **5,000+ students** as a capacity-constrained **multiple-knapsack** problem
- Developed **Express.js APIs** over **MongoDB** to manage wing preferences and batch-specific occupancy rules for allocation

LLM Fine-Tuning for Agentic Capabilities | *Machine Learning, LoRA, PEFT* — Oct 2025

- Fine-tuned **Qwen3-30B-A3B MoE** on 20K examples with **LoRA (PEFT)** for multi-step tool-calling on one **H200 GPU**
- **Tripled training throughput**: traced **50% GPU under-utilization** to **CPU-bound starvation**, rebuilt the pipeline
- Removed preprocessing bottlenecks with Rust-backed tokenizers, parallelizing tokenization across **22/24 CPU cores**

Production Backend for BITS Festivals (Oasis & Apogee) | *Django, Cloud Deployment* — Apr 2025 – May 2026

- Architected a **Django** registration and payment system handling **$240K+** in fest revenue and **6K+ participants**
- Scaled a **TOTP QR** system across **100+ events** serving 50K+ requests via **Dockerized** services behind **NGINX**
- Automated 10K+ transactional emails with **Amazon SES** and dynamic **FCM notifications** via **Celery** queues
- Built a **Firestore** food-ordering backend with real-time status updates, processing **50K+ orders** across fest vendors
- Hardened reliability with **Prometheus**, **Loki**, and **Grafana** for metrics and centralized logs, **Sentry** error tracking, rate limiting, **Cloudflare** DDoS protection, and **IP-restricted SSH-tunneled** access

Positions of Responsibility

CoSTAAn (Head of Fest & Department), Dept. of Visual Media — May 2026 – Present
Committee of Students for Academic Activities — apex body of APOGEE '27, BITS Pilani's flagship technical fest

- Elected **1 of 9** CoSTAA members from **1,100+ batchmates** to lead 70+ student bodies and host APOGEE's **45th edition**
- Spearheading planning and execution of the upcoming tech fest, expecting **6,000+ attendees** and a budget of **$190K+**
- Leading a **50-member** product & design team producing websites, portals & videos for BITS Pilani's cultural & tech fests

Senior Analyst — Sep 2024 – Present
The Wall Street Club, BITS Pilani — India's oldest undergraduate investment fund (AUM > $18K)

- Drove **30% YoY** returns by authoring **7+ investment** theses across equity research and fundamental valuation
- Integrated **5Swans**, a Singapore-based quant firm, via a **redis-streams** pipeline deploying trading strategies in real-time
- Designed the club's website & recruitment platform (**4K+ views**, **600+ applications**) with an internal investing tracker

Technical Skills

Languages: C++, Python, TypeScript, Java, SQL
Frameworks: Django, Express.js, NestJS, Node.js
Developer Tools: Docker, AWS, Cloudflare Workers, Git, NGINX, Celery, Redis
Concepts: Object Oriented Programming, Data Structures & Algorithms, Database Systems, Backend Web Development