

SARTHAK KALPASI

+91 99689 52941 | srtk.in | skalpasi@protonmail.com

WORK EXPERIENCE

Ameeba Labs LLC

DevOps Engineer

(Remote) Sunnyvale, CA

Oct 2025 – Present

- Provisioned a production **EKS** cluster from scratch: node groups, **ALB** ingress with IP target-type and session stickiness for stateful **Socket.IO**, autoscaling (**HPA**, **cluster-autoscaler**), and isolated namespaces.
- Automated deployment with **GitHub Actions** (containerize, push to **ECR**, roll out to **EKS**) and cut monthly **AWS** spend ~**45%** by consolidating duplicate **EC2** instances, decommissioning idle **EBS/EFS/RDS**, and sharing **ALBs** via **IngressGroup**.
- Centralized metrics and logs with **Prometheus**, **Grafana**, **Loki**, and **Alertmanager** (via **Helm**), replacing ad-hoc SSH debugging with alerts on errors and crashes.
- Root-caused two production incidents: recurring **API** pod restarts from slow queries blocking the Node.js event loop, which timed out liveness probes (missing database indexes), and **HPA** silently failing during a usage spike (a broken **IAM** permission path).
- Engineered an on-device AI agent that translates natural language into workspace operations on quantized LLMs via **llama.cpp**; no user data leaves the device. Raised tool-call accuracy from **70% to 98%** on a 7B model via **GBNF** grammar-constrained decoding and schema validation.
- Designed a **zero-knowledge relay**: web and mobile clients receive inference from the user's own desktop, while the backend queues ciphertext only and holds no key. Integrated with **E2EE** chat by delegating encrypt-and-send to the key holder; writes require explicit confirmation.
- Built a **54-fixture evaluation harness** scoring tool accuracy, argument correctness, and **false-action rate** per model size; it made reliability regression-tested and disproved three accepted root-cause diagnoses.
- Root-caused a failure class to **context starvation**: a hardcoded 4096-token context, split across sequences, left a 32k-capable 7B with ~2k tokens for a 3.8k-token prompt. Sized from **GGUF trainContextSize** instead, with breadcrumb backoff for an allocation failure that **kills the process rather than throwing**.

Selfless Sewa NGO

Backend Developer Intern

Delhi, India

Sep 2024 – Nov 2024

- Integrated the **PhonePe** gateway, automating reconciliation of **1k+ monthly transactions**; cut CI release time **3x** with parallel stages and caching.

Networking and Security Lab, IIIT Delhi

Summer Research Intern

Delhi, India

Apr 2023 – Jul 2023

- Deployed a **30-node Ethereum test network** with **Geth**; built a **Bash** RPC benchmarking tool and tuned node config for **6x** throughput.

EDUCATION

Indraprastha Institute of Information Technology, Delhi

Bachelor of Technology in Electronics and Communication Engineering

Nov 2021 – Jun 2025

Research Thesis: Deep Learning based User Behavior Prediction by Exploiting Android Side-channel Vulnerabilities

SKILLS

Programming Languages: Python, Bash, C, C++, SQL

Cloud & Infrastructure: AWS (EKS/EC2/EFS/IAM/S3), Kubernetes, Docker, Helm, GitHub Actions

Systems & Observability: Linux, TCP/IP, POSIX, Socket Programming, Nginx, Prometheus, Grafana, Loki

ACHIEVEMENTS AND PROJECTS

- Discovered a **critical RCE-class vulnerability** in the Programiz online compiler. Writeup: srtk.in/programiz.
- Top 5** in CSAW CTF India Region (2023, 2024); **2nd place** in WreckCTF 2022 as a [d4rk0de](#) team member.
- Self-hosted a home server running Immich, Gitea, Cloudreve, Bitwarden and more. Writeup: srtk.in/lab.
- Projects: [BitChess](#), a multithreaded chess engine and [JustTalk](#), a real-time video conferencing client/server