Subhajit Paul

DEVOPS | SYSTEM ARCHITECT | SRE
KOLKATA, WEST BENGAL, INDIA
Ph- 9836069467

paulsubhojit730@gmail.com
https://subh0jit.me

ABOUT

A A CS student with a passion for **DevOps, Network security**, and a growing interest in **LLMs** and **Agentic auth** technologies. Actively building side projects while continuously sharpening my skills and problem solving mindset. Currently exploring modern tech stacks, writing functional code, and growing through hands on experience in web, mcp & llm architecture.

RECENT PROJECTS

I) **Supply Chain Optimizer using Federated Learning**
3rd year Project | GitHub – private repo

- Deployed on a **self-hosted home server (RTX 3060 12GB / 16GB RAM)** as a private inference + aggregation server; containerized all services via **Docker Compose** with GPU passthrough (--gpus all) for Ollama-based LLM inference.
- Ran **GGUF Q4_K_M quantized** Llama3.2/Gemma2 locally, cutting VRAM usage from ~8GB to ~4.2GB — freeing headroom for simultaneous FedAvg aggregation
- Secured aggregation endpoint with **JWT auth** — rejecting unauthenticated weight submissions to prevent model poisoning
- Implemented **quorum-based fault tolerance** (2-of-3 threshold, 60s timeout) — rounds complete even if a client node dies mid-cycle
- **Nginx** reverse proxy with rate limiting + **Cloudflare Tunnel** for secure public exposure; **CI/CD** via GitHub Actions → Docker Hub → Watchtower
- Monitored VRAM, round latency, and inference time with **Prometheus + Grafana**; stress-tested with **Locust** (50 concurrent clients), found and fixed a weight buffer race condition using threading.Lock()

EXPERIENCE

- **AI Model Architect** at azmth (2026)
  - Fine-tuned a **base LLM**, reducing hallucination through SFT dataset curation and preference alignment
  - Improved factual accuracy across iterative training runs with quantitative eval benchmarks
  - Integrated **real-time TTS**, building end-to-end pipeline where model generates and speaks responses
- **Research intern** at Adamas University (2025)
  - Co-authored a paper on multi-agent legal reasoning — presented at **40th Indian Engineering Congress, NIT Durgapur**
  - Built adversarial **multi-agent LLM framework** (LangChain + Ollama + ChromaDB) simulating lawyer-judge deliberation
  - Implemented **RAG pipeline** with nomic-embed-text for citation-grounded legal decisions over local vector store
- **Tech Consultant** at Skywaave llp (2024)
  - Administered a **self-hosted GPU inference server (Radeon RX 9060 XT)** — uptime, drivers, and secured remote access.
  - Built a **scraping pipeline** (BeautifulSoup + Scrapy) to collect and structure domain-specific training data
  - Fine-tuned **Gemma** on proprietary scraped data for internal startup tooling

SKILLS

- **Languages**: C, JS, **Python**, **Go**
- **Tools**: Git, **Docker**, AWS, Vercel, Supabase, **Postman**, **BurpSuite, n8n**
- OS: Windows, Linux Mint

LANGUAGE

- English (Fluent)
- Hindi (Fluent)
- Bengali (Native)

EDUCATION

CSE, **Adamas University**