Shubh Saxena

Machine Learning Engineer
shubh.saxena2012@gmail.com | +91 7703901127 | LinkedIn

EDUCATION

IIT Roorkee 2023 – 2028
Integrated M.Tech in Geophysical Technology | Minor in Economics

EXPERIENCE

Decompute | *Remote MLE Intern* Apr 2026 – June 2026 | *Cupertino, CA*

- Reduced routing errors by **50%** and streamlined cross-service operations by engineering a high-performance Python/FastAPI smart router with LLM-based intent classification and a unified task-frame system for seamless state-preserved transitions.
- Increased system resilience by **60%** by configuring reverse proxy security, rate limiting, and circuit breakers, while integrating message queues to automate crash-reporting pipelines.
- Secured subscription revenue by implementing a fail-closed Stripe verification pipeline that synchronizes webhooks with signed workbook entitlements, eliminating access leakage and ensuring accurate tier-gating for **100%** of users.

Department of Management Studies, IIT Roorkee | *DevOps Intern* Feb 2026 – Mar 2026

- **Achieved** 100% operational risk mitigation by architecting a production-grade verification platform featuring a structured pan-India vendor database for a Government IPR initiative.
- **Reduced** manual compliance lookups by **80%** by building a dual-sided natural-language search framework enabling consumers to verify real-time product authenticity and sellers to query plain-language IPR rules.
- **Automated** 70% of manual steps while maintaining **99% accuracy** on vendor compliance checks by deploying a containerized, end-to-end workflow CI/CD verification pipeline.

HCLTech | *DevOps Intern* May 2025 – July 2025

- **Decreased** release cycle times by **40%** with zero manual intervention by engineering a GitOps-driven Jenkins → ArgoCD pipeline enforcing shift-left security via Trivy and SonarQube quality gates (**80% coverage threshold**) for containerized Kubernetes microservices.
- **Saved** $55K+ annually (**96% infrastructure cost reduction**) by consolidating 100 individual NAT Gateways into 2 via a Terraform-automated Transit Gateway hub-and-spoke cloud architecture.
- **Accelerated** enterprise multi-account onboarding to **less than 30 minutes** by writing automated Terraform infrastructure-as-code provisioning playbooks for 50+ isolated AWS target environments.

PROJECTS

Autonomous Swarm Intelligence & SecOps Pipeline | GitHub

- **Enabled** zero-lag drone telemetry anomaly detection across **4 health dimensions** by implementing a stateful PyFlink data pipeline over Redpanda streams integrated with **DVC** and **MLflow** for absolute ML reproducibility.
- **Maintained** stable, sub-100ms multi-agent swarm orchestration by deploying a **Ray Serve** GPU inference cluster (YOLOv26 + FP16 + SAHI + BoT-SORT) tied to a **LangGraph StateGraph** hierarchy, and cut incident root-cause time by **65%** via **OpenTelemetry** tracing exported to **Jaeger**, centralized **ELK** logs, and Prometheus/Grafana on Kubernetes.
- **Hardened** the LLM-to-actuator command path against unsafe maneuvers by enforcing a deterministic policy gate exposed through a typed **MCP** tool boundary and validated by a deterministic **agentic evaluation harness** that replays mission-safety scenarios in CI before any command reaches the drone.

Intelligent Document AI Pipeline | GitHub | Certificate

- **Secured** ≥93% processing accuracy at a localized cost threshold of **$0.01/document** by constructing an automated 3-tier adjudication extraction ladder targeting unstructured financial invoice fields.
- **Minimized** absolute production compute overhead by **30%** by containerizing parsing engines into isolated, horizontally scalable Docker microservices governed by intelligent selective SLM routing.

Autonomous Agent Evaluation & Orchestration Framework | GitHub

- **Captured** a **4× reduction** in manual evaluation overhead alongside a **3× throughput gain** by developing a FastAPI + WebSocket orchestration server that models Terminal-Bench-2 tasks as an OpenEnv-compliant reinforcement learning environment via GRPO step reward shaping.
- **Realized** a **60% drop** in model evaluation latency by deploying distributed, multi-pod agent training loops inside isolated Docker environments managed by a decoupled Kubernetes runner infrastructure.

PUBLICATIONS

MACE-RL: Meta-Adaptive Curiosity-Driven Exploration with Episodic Memory in RL Environments *(Zenodo)* 2026

TECHNICAL SKILLS AND CERTIFICATIONS

DevOps & Cloud: AWS, Kubernetes, Docker, Terraform, ArgoCD, Jenkins, GitHub Actions, Prometheus, Grafana, OpenTelemetry, Jaeger, ELK Stack
MLOps & Data: MLflow, DVC, Ray Serve, Redpanda, Apache Flink, Langgraph, Langsmith, MCP, Agentic Harness
Security & Lang: Trivy, OWASP, SonarQube, Falco, ZAP, Python, Bash, YAML, Pytorch
Certifications: Introduction to FinOps (FinOps Foundation)