

Shubh Saxena

RL Engineer

shubh.saxena2012@gmail.com | +91 7703901127 | [LinkedIn](#)

EDUCATION

IIT Roorkee

2023 – 2028

Integrated M.Tech in Geophysical Technology | Minor in Economics

EXPERIENCE

Polymath (YC) | RL Engineer

Jul 2026 – Present

- **Architected** the task distribution for **4 long-horizon coding-agent RL environments**, specifying tool-use observation/action spaces, state-transition and reset semantics, and parameterized scenarios across checkout, release, failover, and voice recovery.
- **Designed** verifier-derived reward functions with shaped intermediate signals and terminal invariant checks, improving long-horizon credit assignment while rejecting reward hacking, no-op policies, and shallow state edits.
- **Ran distributed GRPO policy optimization** and ablations over reward normalization, KL control, rollout sampling, and curriculum difficulty; diagnosed reward distributions and failure trajectories to stabilize learning and guide environment iteration.

Decompute | Remote MLE Intern

Apr 2026 – June 2026 | Cupertino, CA

- **Enabled** date-aware recall across **3 memory classes** by architecting a memory-aware RAG layer over structured temporal records, semantic user profiles, and episodic chat history for meetings, reminders, and profile-aware conversations.
- **Grounded** profile-aware responses in retrieved user evidence by fusing metadata-filtered BM25 and dense search with reciprocal-rank fusion, and specifying cross-encoder reranking with CRAG-style relevance correction for ambiguous memory queries.
- **Improved** resilience by **60%** and protected subscription revenue by combining reverse-proxy security, rate limiting, circuit breakers, and message queues for crash reporting with fail-closed Stripe webhook verification, maintaining accurate tier-gating for **100%** of users.

Department of Management Studies, IIT Roorkee | DevOps Intern

Feb 2026 – Mar 2026

- **Supported** the government beta's million-user capacity target by independently autoscaling RAG retrieval and CNN inference on EKS using CPU demand and Redpanda consumer lag.
- **Reached 91.3% cache hits** and held shard-load variance to **8.2%** through three-tier Redis caching, consistent-hash routing, and PostgreSQL read scaling.
- **Eliminated** single-AZ event-loss risk by engineering multi-AZ Redpanda replication, idempotent producers, ordered processing, and dead-letter recovery for asynchronous verification workloads.

HCLTech | DevOps Intern

May 2025 – July 2025

- **Decreased** release cycles by **40%** with zero manual intervention through Jenkins → ArgoCD GitOps, Trivy/SonarQube gates, and an **80% coverage threshold** for Kubernetes microservices.
- **Saved \$55K+ annually (96% infrastructure cost reduction)** by consolidating 100 NAT Gateways into 2 with a Terraform Transit Gateway hub-and-spoke architecture.
- **Accelerated** multi-account onboarding to **under 30 minutes** with Terraform modules for 50+ isolated AWS environments.

PROJECTS

Contrastive Latent-State Policy Learning | [GitHub](#)

- **Reduced** held-out world-model prediction error by **28%** by feeding cosmetic-invariant contrastive features into matched deterministic and stochastic GRU-RSSMs over seven public-state channels under partial observability and delayed actions.
- **Cut** live-trajectory requirements by **35%** at matched support-restricted policy quality by combining component-masked recurrent BC/IQL with two-step pessimistic imagination using MOPO penalties, MOREL OOD halting, and COMBO-style conservative targets.
- **Sustained 99.5% valid-action execution** at **sub-100ms p95 decision latency** by running fresh-batch recurrent PPO across multi-GPU AWS and enforcing four real-time gates for observation freshness, action legality, geometry integrity, and execution authority.

Autonomous Agent Evaluation & Orchestration Framework | [GitHub](#)

- **Cut** manual evaluation overhead by **4×** and increased throughput by **3×** by modeling Terminal-Bench-2 tasks as an OpenEnv-compatible FastAPI/WebSocket RL environment with GRPO reward shaping.
- **Reduced** model evaluation latency by **60%** through distributed Docker/Kubernetes runners.

Autonomous Swarm Intelligence & SecOps Pipeline | [GitHub](#)

- **Enabled** telemetry anomaly detection across **4 health dimensions** with stateful PyFlink over Redpanda, DVC, and MLflow.
- **Maintained** sub-100ms multi-agent orchestration with Ray Serve/Kubernetes and cut incident root-cause time by **65%** using OpenTelemetry, ELK, Prometheus, and Grafana.

RESEARCH

MACE-RL: Meta-Adaptive Curiosity-Driven Exploration | Accepted, AAAI-26 Student Abstract Program | [Zenodo](#)

2026

TECHNICAL SKILLS AND CERTIFICATIONS

RL & World Models: GRPO, Recurrent PPO, Behavior Cloning, IQL, RSSMs, Model-Based RL, Contrastive Learning

RL Systems: PyTorch, DDP, Ray Serve, MLflow, DVC, AWS, Kubernetes, Docker, Redpanda

Agents & Languages: LangGraph, LangSmith, MCP, Agentic Evaluation, Python, Bash

Certifications: [Introduction to FinOps](#) (FinOps Foundation)