

# Qure.ai Locks-in Predictable GPU Costs and a Zero-Downtime Path to Production AI

Started from a month-long GPU benchmark to 10x infrastructure scale-up plan, Qure.ai is building its LLM training and diagnostic AI stack on AceCloud.

## Key Outcomes

### Zero

Downtime during UAT/  
Dev migration

### 1-Month

POC to validated  
production readiness

### 5 GPU

Types benchmarked (L4, L40S,  
H200, RTX PRO 6000, A30)

## Customer Background

**Industry:** Health-Tech - AI/ML | **Location:** Mumbai, Maharashtra, India  
| **Use Case:** GPU Compute for LLM Training & Diagnostic AI Workloads

Qure.ai needed a GPU cloud partner that could prove price-to-performance economics and platform stability before it would commit production healthcare workloads to the migration.

## Use Case

- **AI-Powered Diagnostic Imaging**  
Builds AI software that helps radiologists and clinicians detect diseases from X-rays, CT scans, and other medical imaging.
- **LLM Training & Diagnostic Inference**  
Runs GPU-intensive workloads for both diagnostic model inference and LLM training.
- **Multi-Provider Infrastructure**  
Operates across multiple providers. UAT/ Dev earlier on E2E, production on AWS.

## Challenges

- 🚩 **Cost Unpredictability Hampered Scaling**  
Unpredictable cloud billing made it hard to plan the growth that [Qure.ai](#) was expecting. Cost stalled committing further spend to any single provider.
- 🚩 **GPU Performance Had to Hold Up Under Clinical-Grade Stakes**  
They needed consistent GPU performance, not a one-time benchmark. Because these systems touch clinical workflows, any migration had to happen without service disruption.



"Moving our LLM training to AceCloud was one of the better infrastructure decisions we made. The stability has directly improved our delivery timelines."

—Suchit Kotiyan, Infra Head. Qure.ai

## Solutions (After AceCloud)

AceCloud delivered a predictable infra billing structure along with an on-demand access to a broad GPU fleet.

- ✓ **Broad GPU Fleet, On-Demand**  
On-demand access to a broad GPU Fleet (L4, L40S, H200, RTX PRO 6000, and A30 GPUs).
- ✓ **Zero-Downtime Migration**  
UAT and Dev environments moved from E2E with zero downtime.
- ✓ **Predictable, Price-Performance-led Cost Model**  
A predictable infrastructure cost structure replaced the billing uncertainty.
- ✓ **Validated in a One-Month POC**  
A month-long proof of concept benchmarked seamless performance across multiple GPU types alongside platform stability.

## Impact Highlight

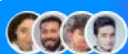


# Zero

Downtime during UAT/Dev migration

Qure.ai validated AceCloud's GPU performance and platform stability in a month-long POC. This improved delivery timelines with full production migration from AWS planned next, contingent on sustained 99.95% uptime.

AceCloud is a cloud infrastructure provider under RTDS Group, delivering secure and scalable cloud solutions to global businesses. Backed by 600+ experts, 20,000+ customers and 10+ data centers, AceCloud helps organizations build reliable, compliant and performance-ready IT environments.



Trusted by 20,000+ Businesses



Call us  
+91-789-789-0752



Website  
acecloud.ai



Mail to  
sales@acecloud.ai