

Cost Component	Unit Cost	Monthly Usage	Monthly Estimate
GPU Compute (T4)	0.27	720	194.4
Storage	0.1	100	10
Network Transfer	0.09	170	15.3
Monitoring (W&B)	25	1	25
Total Monthly Cost			244.7

Metric	Value
Average Latency (seconds)	3.6
Theoretical Requests per Sec	0.277777778
Theoretical Requests per Hour	1000
Conservative Requests per Hour	800
Hourly GPU Cost	0.27
Cost per Request	0.0003375
Cost per 1,000 Requests	0.3375

Strategy	Monthly GPU Cost Savings vs Baseline	
Baseline (24/7)	194.4	
Auto-Scaling (12h active)	97.2	97.2
INT8 Quantization (25%)	145.8	48.6
Batching (18% reduction)	159.408	34.992
Spot Instances (40% reduction)	116.64	77.76