

[Overview](#)[Projects](#)[Models](#)[Utilization & Efficiency](#)[Cost & Spend](#)[Policies & Quotas](#)[Incidents](#)[Reports](#)[Projects](#)

4

[Models](#)

12

[GPUs](#)

28

[Utilization](#)

76%

[Spend \(MTD\)](#)

\$18.4k

[SLA Compliance](#)

99.3 %

[Risk](#)

2

Usage by Project

Chatbot LLM  3 GPUsSearch API  6 GPUsRAG Service  4 GPUsBatch Job  3 GPUs

Busy

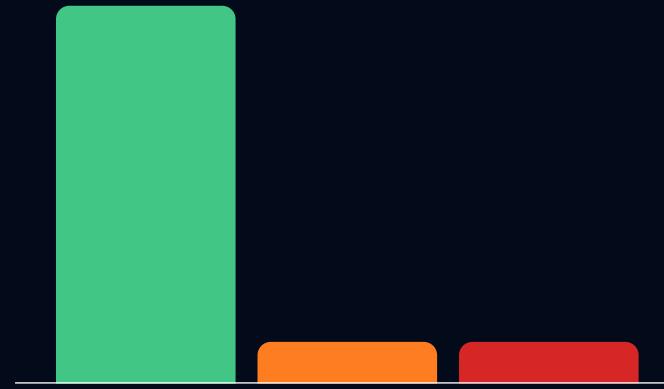
Under-utilized

Idle

Spend Trend



Model health & SLA



10 Healthy

1 Degrading

1 Falling

Risk and Violations

Policy Violations

3

Cost Overages

2

Under-utilized

4

Preemption Risk

1

Tenants Table

Project	Models	GPUs	Tier	Util %	Spend (MTD)	SLA	Risks	Status
Search	4	10	Reserved	88	\$6.2k	99.9	0	Healthy
Chatbot	3	8	Elastic	61	\$4.8k	98.7	2	⚠️
RAG	3	6	BestEffort	42	\$3.1k	97.8	3	⚠️
Batch	2	4	BestEffort	30	\$1.9k	-	0	Idle