CASE STUDY

How Scatter Lab powers 1,000+ inference requests per second with Runpod

Runpod helped Scatter Lab scale real-time LLM serving for Zeta.

2.1M
cumulative active users

1,000+
requests per second

~50%
customer-reported infrastructure cost reduction

The challenge

Scatter Lab needed hundreds of GPUs for live LLM serving, but scarcity, quota limits, and high instance costs threatened scaling economics.

The Runpod approach

The team used Runpod APIs as part of a multi-region deployment strategy, dynamically scaling compute according to live service load.

The result

Runpod became part of Scatter Lab's operating model for flexible real-time inference, helping the team serve demand without over-provisioning.