

Run inference yourself or buy a managed service

Compare control, operations work and cost per accepted response.

Updated October 29, 2026

A managed inference service can remove deployment and scaling work. A rented GPU gives you more control over the runtime. Neither choice removes your responsibility for application correctness, data handling and spending limits.

List the work you own

For self-managed serving, include image builds, model loading, queueing, scaling, patching, metrics and incident response. Plan how you recover from a failed host. Identify who receives alerts and who can roll back a model.

For a managed service, check supported models, custom weights, request limits and data retention. Ask whether you can pin a model revision. Test the export path for logs and artifacts. Check the failure behaviour when the service is busy.

Measure a comparable result

Use the same input set and output requirements. Record successful responses, errors and latency under load. Measure idle time as well as active work. Compare the full bill and the engineering effort needed to keep the system available.

NVIDIA Triton documents a serving system for model inference. Ray documents tools for distributed applications and serving. Read their deployment and operations requirements before treating software availability as a ready service.

Choose the option that meets your acceptance rule with an operations burden your team can support. Revisit the choice when traffic becomes more predictable or model requirements change. Keep the benchmark so that a later migration uses the same evidence.

Sources

- [NVIDIA Triton documentation](#) — fetched 2026-10-29.
- [Ray documentation](#) — fetched 2026-10-29.

Frequently asked questions

Is this a vendor quote?

No. Confirm current terms and capacity with the provider before you rent.