

Questions to ask a GPU vendor

Get capacity, billing and exit terms in writing.

Updated October 29, 2026

Send a workload brief with your request. State the device requirement, runtime pattern, data sensitivity and required region. Ask the vendor to separate confirmed capacity from capacity that is only listed in the catalogue.

Allocation and recovery

Ask whether the quoted GPU is dedicated, shared or partitioned. Ask what happens if a host fails. Can the vendor move the job, or must your scheduler restart it? Find out how you receive maintenance and interruption notices. Ask where checkpoints must live to survive the loss of a host.

Billing and support

Request a sample invoice based on your workload brief. It should show compute, storage, transfer and support. Ask what is billed while a machine is stopped. Confirm the minimum rental term, cancellation rule and renewal process. Separate service credits from cash refunds.

Exit and evidence

Ask how you export volumes, logs and model artifacts. Check whether API credentials can be scoped and revoked. Request a support escalation path for a failed production job. Keep the answer with the quote and its expiry date.

AWS describes Spot capacity as interruptible. Treat interruption as a design requirement when evaluating similar offers, but do not assume every vendor uses AWS's rules. Confirm each vendor's notice and recovery terms directly. Modal and Lambda publish different billing structures. Use the linked pages to prepare questions, then obtain terms for your actual order.

Sources

- [AWS EC2 Spot Instances](#) — fetched 2026-10-29.
- [Modal pricing](#) — fetched 2026-10-29.
- [Lambda GPU cloud pricing](#) — fetched 2026-10-29.

Frequently asked questions

Is this a vendor quote?

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