

Prepare your first GPU cloud run

Make a small run reproducible, bounded and easy to remove.

Updated October 29, 2026

Choose a small workload with a known result. Use public or harmless data for the trial. Write down the expected output before you start so that a fast but incorrect run cannot pass.

Prepare the run

Pin your container image and dependencies. Record the framework and driver requirements. Use the provider's documented method to attach the GPU. PyTorch's official site links its installation guidance; match the installation to the target environment rather than copying a command from an old notebook.

Keep secrets out of images and source files. Use the service's credential mechanism. Confirm a spending limit or alert before you create resources. Estimate the trial bill from the quoted rate and expected billed time, then include storage and network costs.

Execute and recover

Save logs and checkpoints to durable storage. Measure startup, data loading, compute and shutdown separately. Repeat the run with the same input. Test resume from a checkpoint in a fresh worker before you schedule a long job.

Close the trial

Delete unused workers, endpoints and volumes. Check the account for remaining billable resources. Compare the invoice with the estimate and explain any difference before scaling up.

Keep the run command, image digest, input revision, result and source links together. SkyPilot documentation describes job management across infrastructure. If you use it, record the underlying provider and region as well as the orchestration settings. A successful tool launch does not prove your job finished correctly.

Sources

- [PyTorch](#) — fetched 2026-10-29.
- [SkyPilot 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.