

Feedback on Nebius Token Factory, Serverless AI and AI Cloud

Kept from day 1 of the build. Every item has the same four parts: **problem**, **measured** (a real number from our runs), **workaround**, and **suggested fix**. Items marked *to measure* are known issues we haven't put a number on yet.

1. Cosmos is available only as a dedicated endpoint

- **Problem:** `nvidia/Cosmos3-Super-Reasoner` is offered on Token Factory as a dedicated endpoint only, sized at 4× H200. Physical AI perception is bursty (in our design one call per new hazard), so paying for a whole always-on endpoint is a poor fit. Earlier research couldn't find Cosmos on Token Factory at all, so it's also hard to discover.
- **Measured:** \$0.37 per minute (about \$22.20 per hour) while running, us-central1, from the console. Preflight on Oct 7 got HTTP 429 "model is starting, please retry" from the endpoint, so a stopped endpoint can't serve a demo on demand. *To measure:* cold-start time from stopped to Ready, and calls per demo run.
- **Workaround:** run the endpoint only during test and recording sessions, with min replicas at 0 or the endpoint stopped otherwise, and batch perception requests.
- **Suggested fix:** offer Cosmos Reasoner per token (serverless), or a smaller single-GPU configuration with scale-to-zero, and list it in the model catalog next to Nemotron.

2. Nemotron models are not fine-tunable on Token Factory

- **Problem:** the fine-tunable list doesn't include Nemotron, so the hackathon's headline model family can't be post-trained on Token Factory.
- **Measured:** *to measure:* engineering time to build our own LoRA container, and the job wall time on Serverless Jobs.
- **Workaround:** LoRA in our own container on Serverless Jobs (H200), served by vLLM on a Serverless Endpoint.
- **Suggested fix:** add Nemotron Nano to Token Factory LoRA fine-tuning, with serving on the same platform.

3. Batch inference only promises results within 24 hours

- **Problem:** `/v1/batches` only commits to a 24 h completion window, which is too slow to iterate on during a 3-week build or to show live.
- **Measured:** *to measure:* turnaround of our week-1 test batch and of each Night Shift batch.

- **Workaround:** real overnight batch, plus concurrent per-token calls for the demo, labeled as such.
- **Suggested fix:** a faster batch tier (for example, under 1 h) or a published turnaround estimate per model.

4. Custom weights on dedicated endpoints are beta and available only on request

- **Problem:** the documented way to serve our own fine-tuned weights on Token Factory needs access approval, with no stated turnaround.
- **Measured:** *to measure:* days from request to access (requested on day 1).
- **Workaround:** a vLLM Serverless Endpoint.
- **Suggested fix:** self-serve custom weights (or LoRA adapters) for open NVIDIA models.

5. Serverless Jobs: cold start and GPU quota

- **Problem:** *to measure:* cold-start time for our container on CPU and GPU, time to get H200 quota, and how logs behave across a 20–50 job fan-out.
- **Measured:** *to measure.*
- **Workaround:** *to be determined.*
- **Suggested fix:** *to be determined;* possibly a first-class fan-out or array-job primitive.