MLPerf Training v6.1 and MLPerf Endpoints: what NVIDIA reports for Vera Rubin and throughput per watt

25 November 2026 · 8 min read · blogs.iictp.org/mlperf-training-v6-1-endpoints-vera-rubin-throughput-per-watt-explained

Short answer: on 16 November 2026 NVIDIA said it swept every benchmark in MLPerf Training v6.1 and, in its first MLPerf Endpoints submission, delivered more than 10x generational throughput per watt for agentic AI on its Vera Rubin platform. These are NVIDIA's summaries of results published by MLCommons. Throughput per watt matters economically because data centres are power-limited, so tokens per watt sets how much agent work a site can serve.

TL;DR The short version

- NVIDIA says it had results on every benchmark in MLPerf Training v6.1 and, in its first MLPerf Endpoints submission, more than 10x generational throughput per watt on Vera Rubin.
- Throughput per watt matters because data centres are power-limited, so tokens per watt sets how much agent work a site can serve.
- These are a vendor's summaries and the Endpoints figure is provisional, so read the MLCommons entries and normalise by power and response-time targets.

Read online, listen to it, or change the text size: https://blogs.iictp.org/mlperf-training-v6-1-endpoints-vera-rubin-throughput-per-watt-explained

What are MLPerf Training and MLPerf Endpoints?

Per NVIDIA's description, MLPerf Training measures how fast a system trains a range of models to a target accuracy. MLPerf Endpoints is a newer benchmark that tests systems on agentic coding and enterprise workflows and checks that agents still complete their tasks correctly. Both are run by MLCommons; this post relies on NVIDIA's blog, not on the MLCommons tables.

How the results are framed

- **Training coverage.** NVIDIA says its platform was the only one with results on every benchmark in the round, spanning LLMs, image generation, recommenders and reinforcement-learning post-training.
- **Generation gain.** In its debut, Vera Rubin NVL72 delivered up to 2x more performance per GPU than GB300 NVL72, according to NVIDIA.
- **Scaling efficiency.** On the DeepSeek-V3 benchmark, Vera Rubin NVL72 scaled from 256 to 1,024 GPUs with about 97% efficiency, cutting training time from 16 minutes to about four.
- **Endpoints.** In a provisional Kimi K3 result, NVIDIA reports more than 10x the throughput per watt of GB300 NVL72 at the same response speed.

Why throughput per watt is the number to watch

NVIDIA argues that agents reason, call tools and check their work many times per task, each step generating tokens. Because AI data centres are limited by power, tokens generated per second per watt determines how many agents a facility can serve. That is a capacity and cost argument, not just a speed argument.

Other figures in the post

- Largest training scale in the round: NVIDIA ran 10,240 GPUs on GB300 systems, and Microsoft Azure matched with 10,240 GPUs on GB200 systems.
- Partner submissions came from 20 organisations, listed in the post, including Dell, HPE, Google Cloud, CoreWeave, Oracle and Supermicro.
- The post says Vera Rubin performance will keep climbing through software optimisation.

Limitations and open questions

- This is a vendor’s account; check the MLCommons results for exact configurations.
- The Endpoints figure is described as provisional, and the NVIDIA footnote shows different accelerator counts for the two systems compared, so read the configuration before drawing conclusions.
- Benchmark workloads are not your workload. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to use benchmark claims in planning

1. Open the MLCommons entry IDs cited in the vendor footnotes and read the system configurations.
2. Normalise by power and by response-time targets, not only throughput.
3. Test your own model and traffic pattern on candidate hardware or cloud instances.
4. Include power, cooling and networking in cost, not only the accelerator price.

What it means for IT buyers and infrastructure planners

The framing from tokens per second to tokens per watt reflects how capacity is now constrained. For planners, power availability may limit growth more than chip supply. See NVIDIA’s Q3 fiscal 2027 results and NVIDIA’s supply chain commitments.

Frequently asked questions

- **What did NVIDIA report?** A sweep of MLPerf Training v6.1 and more than 10x throughput per watt in a provisional MLPerf Endpoints result, versus GB300 NVL72.
- **What is MLPerf Endpoints?** A benchmark testing agentic coding and enterprise workflows while checking tasks are completed correctly.
- **Why does power matter?** Data centres are power-limited, so tokens per watt sets serving capacity.
- **Are these independent results?** They come from MLCommons submissions, summarised by NVIDIA.

Related reading

- NVIDIA Q3 fiscal 2027 results
- NVIDIA supply chain commitments
- Cloud Architecture Practitioner certification

Related credential: Cloud Architecture Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. NVIDIA. NVIDIA Sweeps MLPerf Training and Endpoints as Vera Rubin Debuts With Leading Performance. 16 November 2026. blogs.nvidia.com/blog/mlperf-training-endpoints-vera-rubin

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/mlperf-training-v6-1-endpoints-vera-rubin-throughput-per-watt-explained. More guides: blogs.iictp.org.