# AI Factories Turn Output per Megawatt into a Competitive Metric

AI Leaders Briefing

2026-08-24

## AI Factories Turn Output per Megawatt into a Competitive Metric

*By AI Leaders Briefing • August 24, 2026*

NVIDIA's Vera Rubin rollout and MaxLPS claims point to a shift from acquiring AI capacity to measuring output per megawatt, reinforced by changes in model routing, production traffic, and pricing.

### Top Signals of the Week

**NVIDIA AI Infrastructure, OpenAI, and Microsoft — Vera Rubin reaches deployment while power efficiency becomes a product metric**

NVIDIA describes DSX MaxLPS as a suite for maximizing AI-factory throughput within a fixed power budget. Its three levers are dynamic power allocation, software techniques for performance per watt, and 45°C thermal/site design intended to convert less cooling overhead into compute. [1] The underlying problem is static rack provisioning: NVIDIA says an illustrative 540 kW site strands 170 kW, while dynamic provisioning can reclaim that headroom for an additional rack; the Dynamic Power Software used for this reallocation is currently in Developer Preview. [1]

NVIDIA projects up to 40% more Rubin GPU capacity within the same power budget, and a separate NVIDIA post attributes a measurement of 10x more tokens per second per megawatt on the Vera Rubin platform to CoreWeave. [1, 2] These should be treated as NVIDIA projections and reported customer measurements, not general independent benchmarks: the MaxLPS page labels the 40% figure as a projection, and its figure caption refers to GB300 even though the surrounding text refers to Vera Rubin. [1]

The hardware is moving beyond announcement status. OpenAI says its first Vera Rubin racks are running its training stack for next-generation frontier pre-

training; Microsoft CEO Satya Nadella says the first production Vera Rubins have arrived at Microsoft data centers; and NVIDIA says the platform is ramping into full production. [3, 4, 5]

**Why it matters:** The immediate infrastructure contest is shifting from securing megawatts to turning each megawatt into usable model work. For deployment decisions, GPU count is only one variable; power sharing, thermal design, workload throughput, and service-level behavior now matter alongside the silicon.

### Guillermo Rauch / Vercel — open-weight models take the majority of one gateway's tokens

Vercel AI Gateway reports that open-weight models accounted for 62% of its token volume on August 22, versus 28.4% on June 24; closed models fell from 71.6% to 38% over the same comparison. Rauch says enterprise adoption is still early and that harnesses, CLIs, IDEs, and SDKs will need to become model-agnostic. [6]

This is a gateway-level usage signal, not a measure of overall market share. Its importance is operational: model-agnostic routing and developer tooling can determine which models receive production traffic, making compatibility and deployment economics competitive variables alongside model quality.

### Jason Dong Jian / Shopee — an in-house model reaches production scale

An NVIDIA-hosted Shopee case study says Compass, the company's specialized model for Southeast Asian e-commerce, grew from 3 billion to 340 billion monthly API tokens in eight months and now handles the majority of Shopee's AI traffic. The case study lists search, recommendations, anti-fraud, parcel recovery, and multilingual customer service as production uses. [7]

The stack spans thousands of NVIDIA A100, H100, and RTX PRO 6000 Blackwell GPUs, with Megatron-Core for distributed pre-training, NeMo for post-training, and TensorRT-LLM for inference. [7] The same case study reports 50x efficiency over manual review for anti-fraud detection and 90% lower processing costs; those are attributed customer-case-study figures, not a controlled cross-company benchmark. [7]

**Why it matters:** The production unit is becoming a domain model plus its training, post-training, inference, and traffic loop—not a checkpoint evaluated in isolation. Token volume and operating cost provide a more decision-relevant test of enterprise adoption than model size alone.

## Research & Engineering

### Clem Delangue / Hugging Face — agentic coding harnesses show a path to specialized optimization, but public-set scores need a boundary

Clem Delangue reports that NVIDIA built a coding harness to optimize CUDA GPU kernels and achieved a 100% score on ARC-AGI-3's 25 public games, solving all 183 levels. He frames the result as evidence that agents could make running, optimizing, and post-training models and kernels accessible to a much larger builder population. [8]

François Chollet says the approach uses deep-learning-guided, on-the-fly synthesis of symbolic world models, but cautions that a perfect score on the public demonstration set is not the same as a perfect score on the ARC-AGI-3 benchmark. He also asks for the cost per run. [9, 10]

**Why it matters:** The engineering signal is credible as a demonstration of harness-assisted low-level optimization; the evaluation signal is narrower than the headline. Public-set saturation does not establish general autonomous engineering or favorable economics across unseen workloads.

## Strategy & Industry

### OpenAI — model pricing becomes a short-term competitive lever

OpenAI says it is reducing API and credit pricing for GPT-5.6 Sol by more than 20% for three months. The change applies to the API and eligible ChatGPT Work and Codex credits; Pro, Plus, and Business subscription usage is unchanged. [11, 12]

The limited duration and unchanged subscription terms make this a tactical price move rather than a broad list-price reset. It nonetheless puts inference cost directly on the product surface, alongside capability, latency, and infrastructure efficiency.

### CoreWeave, NVIDIA, and Hudson River Trading — Vera Rubin is being positioned for specialized research workloads

CoreWeave says Hudson River Trading chose its AI cloud for scale, while NVIDIA says HRT will use Vera Rubin NVL72 with Spectrum-X Ethernet networking on CoreWeave Cloud for its next generation of model development and research. [13, 14] HRT's quantitative-trading focus makes this a distinct deployment profile from frontier-lab pre-training: the vendor announcements position Rubin as infrastructure for high-performance, specialized research as well as large lab runs.

## Worth Watching

### Superwhisper — on-device open weights move into a user product

Superwhisper introduced S1-mini, its first open-weights language model: a 0.6B-parameter system that processes transcripts entirely on the device. [15] It is a small but concrete edge-inference signal: local execution is being packaged as the product experience, not only pursued as a systems optimization.

## Editorial outlook

The new signals put deployment economics at the center: output per megawatt, gateway token mix, production traffic, and API price are becoming as legible as benchmark scores. [1, 6, 7, 11] The next useful discriminator is independent, workload-specific measurement—particularly for vendor-reported power and cost claims and for agents evaluated on public demonstration sets. [9]

---

### Sources


1. Maximizing AI Factory Performance per Watt with NVIDIA DSX MaxLPS
2. X post by @NVIDIAAIInfra
3. X post by @udayruddarraju
4. X post by @satyanadella
5. X post by @nvidia
6. X post by @rauchg
7. NVIDIA Powers Shopee's Frontier LLMs at Scale
8. X post by @ClementDelangue
9. X post by @fchollet
10. X post by @fchollet
11. X post by @OpenAI
12. X post by @OpenAI
13. X post by @CoreWeave
14. X post by @NVIDIAAIInfra
15. X post by @superwhisper