Bentaus

BENTAUS NEOCLOUD · TECHNICAL EVALUATION · OCTOBER 2026

AMD Instinct™ MI325X vs MI350X

Which Instinct GPU fits the workload? Bentaus ran the MI325X and the MI350X, the air-cooled 1,000 W member of the MI350 Series, through the same production workloads: throughput, concurrency, agents, long context, large models, precision and power. Running the Ziani Power Asset Orchestrator.

1.84–1.91×
MI350X wins: agent tasks per hour within the performance target

1.90×
MI350X wins: Llama 405B as four 2-GPU copies

~24% lower cost
MI325X wins: GPU rent for equivalent Llama 70B FP8 output

+33%
MI325X wins: agent tasks per MW under Ziani power caps

Independent study by Bentaus. All data points measured solely by Bentaus.

SUMMARY

Two Instinct GPUs. Different optimization points.

The MI350X is faster on most tested workloads and substantially faster on agents and certain large-model layouts. The MI325X remains highly competitive where the MI350X's performance gain is smaller than its price premium, and it provides significantly more power flexibility under Ziani.

Both 8-GPU nodes ran at the same site on the same software (vLLM 0.23.0, ROCm 7.14.1), with the same model files, prompts and ramps.[B8–B10]

PERFORMANCE
14 of 21
matched comparisons led by the MI350X; 4 at parity; 3 led by the MI325X

ECONOMICS
1.72×
at current NeoCloud market rates ($2.00 vs ~$3.44 per GPU-hour), the MI350X must exceed this multiple of MI325X output to deliver more output per rental dollar

ENERGY
Two profiles
MI350X more efficient uncapped; MI325X has far more curtailment headroom under Ziani

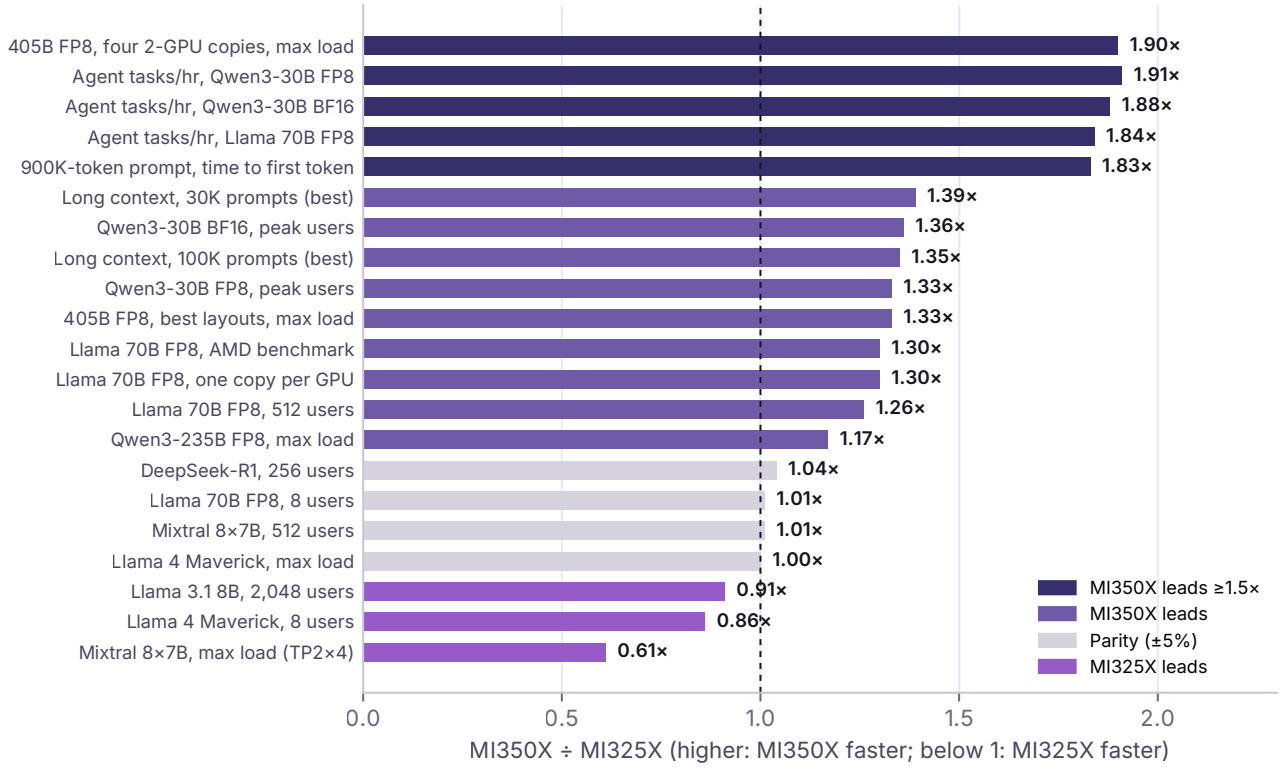

Figure 1. MI350X ÷ MI325X on every matched comparison in this study. Agent rows count tasks at the highest load where ≥90% of tasks met the performance target. Sources: B8–B10.

Independent study by Bentaus. All data points measured solely by Bentaus.

01 · WHAT CHANGED

From CDNA 3 to CDNA 4

	MI325X	MI350X (tested)	MI355X (reference)
Architecture	CDNA 3	CDNA 4	CDNA 4
HBM per GPU	256 GB HBM3E	288 GB HBM3E	288 GB HBM3E
Memory bandwidth	6.0 TB/s	8.0 TB/s	8.0 TB/s
FP8 dense compute	2.6 PFLOPS	4.6 PFLOPS	5.0 PFLOPS
Native FP4 / FP6	No	Yes · 9.2 PFLOPS	Yes · 10.1 PFLOPS
Board power · cooling	1,000 W · air	1,000 W · air	1,400 W · liquid
Peak engine clock	2.1 GHz	2.2 GHz	2.4 GHz

The MI350X and MI355X share the same chip and memory. The MI355X runs at higher power and clocks with liquid cooling. All MI350 Series results in this paper are measured on the MI350X at 1,000 W, the same power class as the MI325X, so the comparison isolates the architecture rather than the power budget.[A2, A3]

02 · THROUGHPUT

About 1.3× on dense models, growing with load

On AMD's maximum-throughput benchmark (Llama 3.1 70B FP8, 128 in / 2,048 out), the MI350X produced **24,794 tok/s against 19,086**, 1.30×, using 29% less energy per token (0.28 vs 0.39 J). With one copy per GPU it reached **38,749 tok/s against 29,807**.[B8, B10]

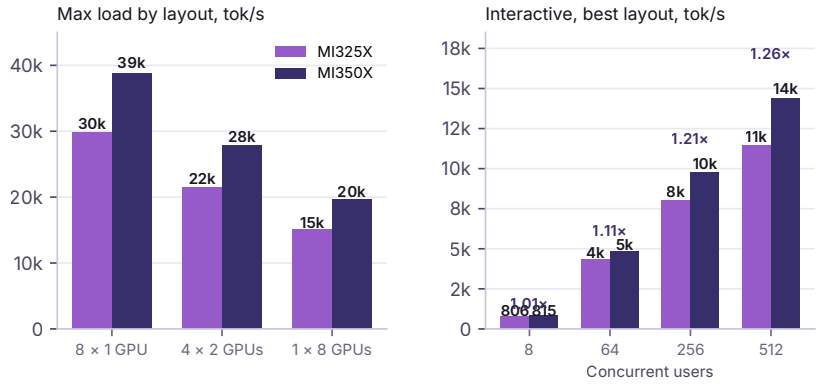

Figure 2. Llama 3.1 70B FP8. Left: maximum load by layout (128 in / 2,048 out). Right: interactive serving (1,024 in / 1,024 out), best layout on each GPU; labels show MI350X ÷ MI325X. Source: B8, B9.

Independent study by Bentaus. All data points measured solely by Bentaus.

> The MI350X delivered 30% more throughput on this workload: 24,794 output tokens/s against 19,086 on the MI325X. At current NeoCloud rates ($3.44 vs $2.00 per GPU-hour), **producing the same number of Llama 70B output tokens costs about 24% less on the MI325X**. The MI350X provides greater performance density per GPU; the MI325X provides lower rental cost for the same aggregate output on this workload.

At light interactive load the two are nearly identical (815 vs 806 tok/s at 8 users). The MI350X's lead grows with concurrency, reaching 1.26× at 512 users, where its extra compute and bandwidth carry larger batches.

03 · CONCURRENCY

Both scale to 1,024 users per GPU; the MI350X holds latency longer

We ramped Qwen3-30B-A3B from 1 to 1,024 users per GPU, one copy per GPU, until throughput stopped growing. Both GPUs saturated compute at 1,024 users per GPU with KV cache still available, so memory wasn't the limit on either.[B8, B9]

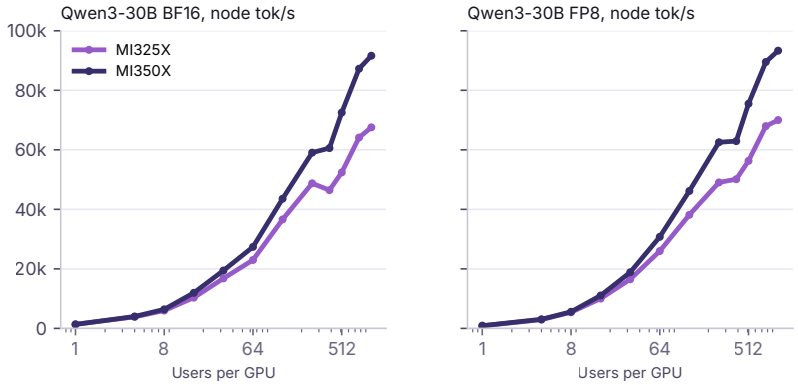

Figure 3. Qwen3-30B-A3B user ramp, eight copies per node, 1,024 in / 512 out. Source: B8, B9.

Peak throughput was **91,616 vs 67,551 tok/s at BF16** (1.36×) and 93,311 vs 70,002 at FP8 (1.33×). The bigger difference is latency. The MI350X stayed within the performance target (first token ≤2 s, inter-token ≤200 ms, P99) at every step up to **768 users per GPU**. The MI325X first missed it at 128 users per GPU, passed again at 384, and missed from 512.

04 · AGENTS

Close to 2× the completed work

Each agent task is 15 model calls with context growing to ~60,000 tokens. A load level counts only if at least 90% of its tasks meet the performance target, and every run uses fresh tasks.[B8, B9]

Independent study by Bentaus. All data points measured solely by Bentaus.

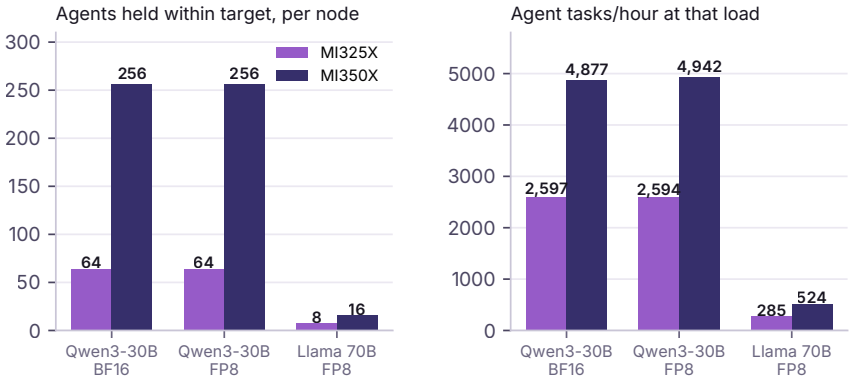

Figure 4. Highest agent load per node where ≥90% of tasks met the performance target, and tasks per hour at that load. Both GPUs used AMD's AITER attention. Source: B8, B9.

On Qwen3-30B the MI350X kept **all 512 tasks within target at 32 agents per GPU**, four times the MI325X's passing load, and completed 4,877 tasks per hour at BF16 against 2,597 (1.88×), and 4,942 against 2,594 at FP8 (1.91×). On Llama 70B it held two agents per GPU against one, at 524 vs 285 tasks per hour (1.84×).

05 · LONG CONTEXT AND LARGE MODELS

Where the extra 32 GB matters

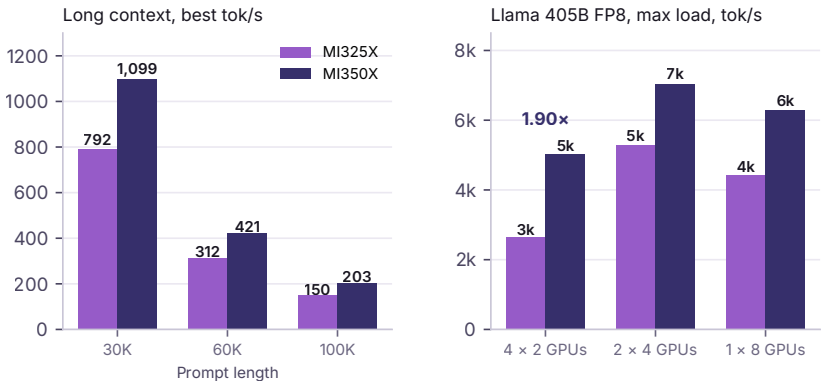

Figure 5. Left: Llama 3.1 70B FP8 long-context throughput, best session count on each GPU. Right: Llama 3.1 405B FP8 at maximum load by layout. Source: B8, B9.

Long context. The MI350X led by 1.35–1.39× at its best session counts, with 10.0M tokens of context per node against 8.5M. On a single 900,000-token Llama 4 Scout prompt, it returned the first token in about 5 minutes against 9 on the MI325X.

405B. With four 2-GPU copies, each GPU holds roughly 205 GB of weights. The MI350X's extra 32 GB per GPU leaves room for users' context, and it produced **5,016 tok/s against 2,645 (1.90×)**. Comparing each GPU's best layout (two 4-GPU copies), the lead is 1.33×. At 64 interactive users the MI350X was also 1.33× faster (1,022 vs 771 tok/s).

Independent study by Bentaus. All data points measured solely by Bentaus.

06 · WORKLOAD TRADEOFFS

Where the MI325X leads or reaches parity

Newer hardware doesn't run every workload faster. In the tested ROCm and vLLM configuration, the MI325X matched or beat the MI350X on several models.[B8, B9]

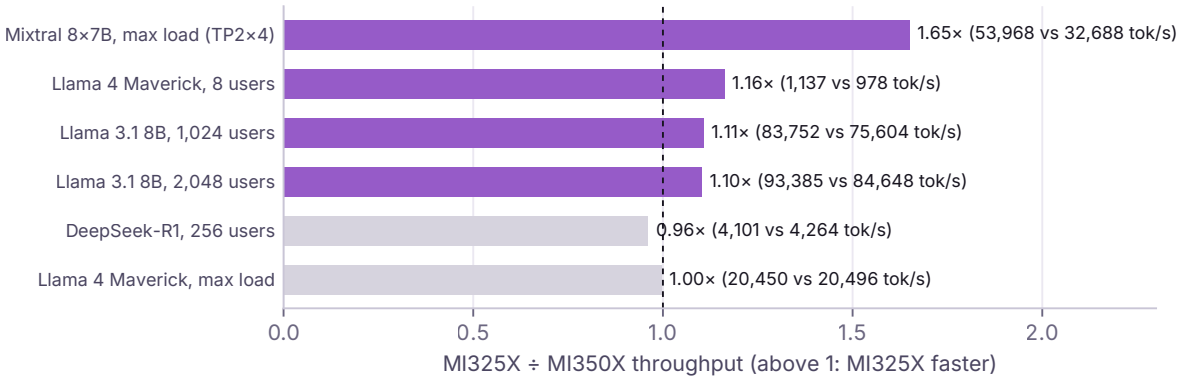

Figure 6. MI325X ÷ MI350X throughput on the tests where the MI325X led or matched. Source: B8, B9.

- **Mixtral 8×7B:** 53,968 vs 32,688 tok/s at maximum load as four 2-GPU copies, 1.65×. At 512 interactive users the two were at parity. Mixtral ran with AMD's AITER kernels on the MI325X and without them on the MI350X; further software optimization may change this result.
- **Llama 3.1 8B at very high concurrency:** 1.10–1.11× at 1,024–2,048 users, although the MI350X led at maximum load (1.26×).
- **Llama 4 Maverick:** parity at maximum load; the MI325X was 1.16× faster at 8 users.
- **DeepSeek-R1:** near parity at 256 users (4,264 vs 4,101 tok/s, MI350X +4%); the MI350X led by 16% on long reasoning outputs.

07 · PRECISION

MI325X for FP8 value. MI350X adds native FP4.

FP8 remains an important production inference format on both GPUs. The MI325X combines strong FP8 performance with its lower GPU-hour rate, making FP8 one of its strongest economic use cases. The MI350X advances the precision story by adding native FP4 hardware, so models validated at FP4 can cut their memory footprint while keeping much more of their FP8 throughput.

Independent study by Bentaus. All data points measured solely by Bentaus.

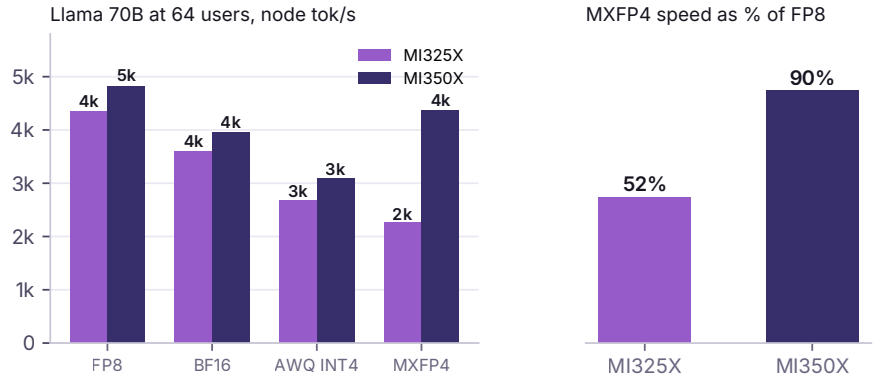

Figure 7. Llama 70B, one copy on 8 GPUs. Left: node throughput at 64 users by precision. Right: MXFP4 throughput as a share of FP8 on each GPU. Source: B8, B9.

On the MI350X, native MXFP4 ran at **90% of FP8 speed**; on the MI325X, which has no 4-bit math units, it ran at 52%. Quality was the same on both GPUs at every precision on our GSM8K, HumanEval and MMLU-Pro evaluations. FP8 remained the fastest option on both, so FP4's main benefit at this model size is memory: half the weight footprint at nearly FP8 speed. At maximum load the MI350X ran BF16 1.43× faster than the MI325X (15,775 vs 11,034 tok/s).

08 · ECONOMICS

Does the workload earn the price premium? ECONOMIC MODEL

This analysis uses current NeoCloud market rates as of October 2026: $2.00 per GPU-hour for the MI325X and ~$3.44 for the MI350 Series. Rental prices change; the measured performance ratios don't. If prices move, the break-even moves with the price ratio, so readers can substitute their own rates.

> **At these prices the MI350X costs 72% more per GPU-hour. To beat the MI325X on output per rental dollar, it must produce more than 1.72× the MI325X's useful output on the same workload.**

Independent study by Bentaus. All data points measured solely by Bentaus.

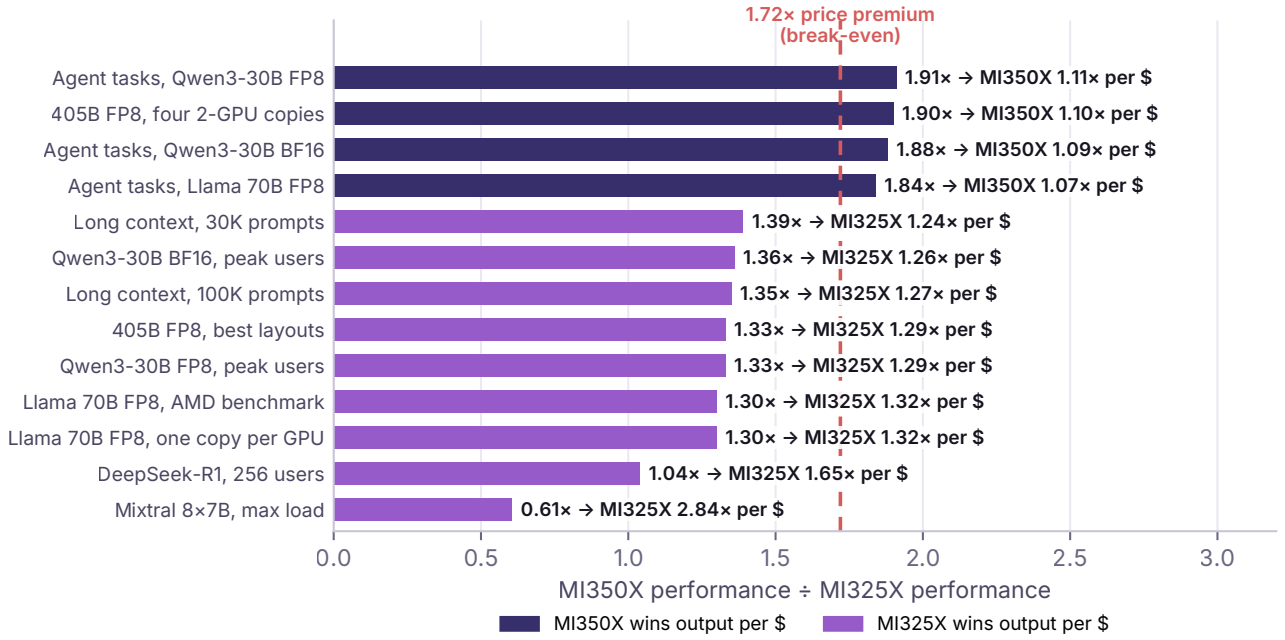

Figure 8. Measured MI350X ÷ MI325X performance against the 1.72× break-even. Right of the line, the MI350X delivers more output per dollar; left of it, the MI325X does. Per-dollar advantage = performance ratio ÷ 1.72. Source: B8–B10.

Cost to produce the same output

The simplest way to read the economics: normalize each workload to the same useful output, and compare what each GPU costs to produce it.

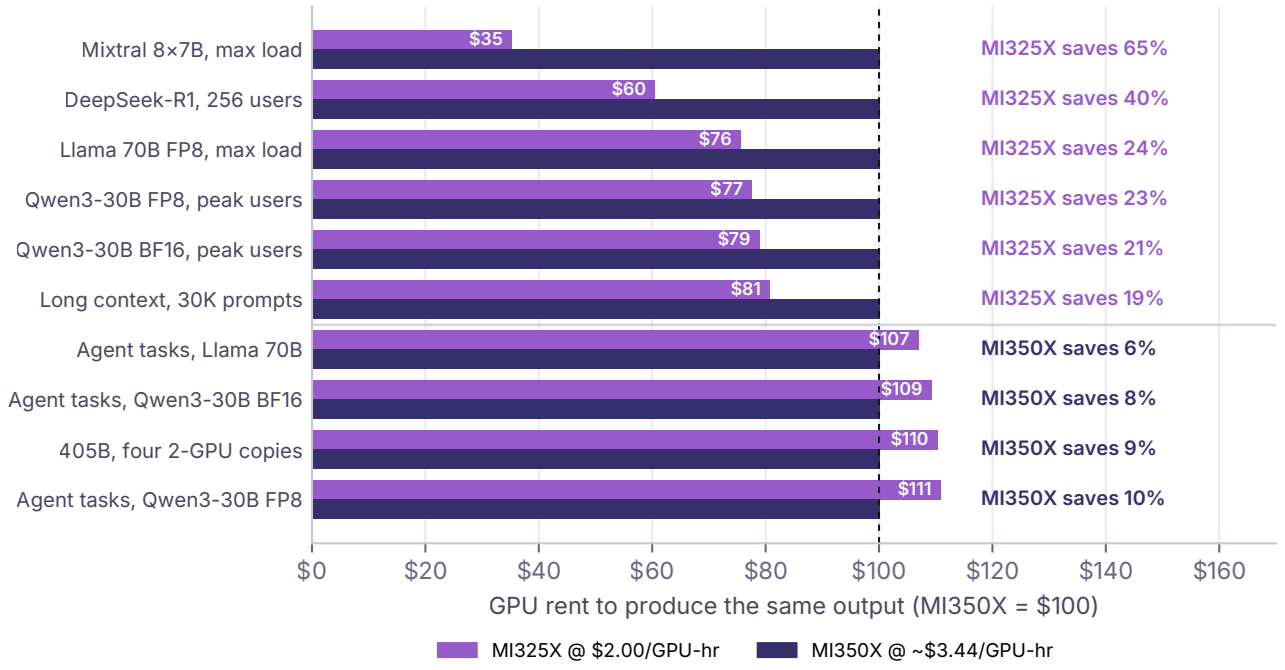

Figure 9. GPU rent to produce the same output on each workload, with the MI350X set to $100. Above the line, the MI325X is the lower-cost choice; below it, the MI350X. Agent rows use tasks at the highest load meeting the performance target. Source: B8–B10.

Independent study by Bentaus. All data points measured solely by Bentaus.

For the same Llama 70B output, the MI325X costs about 24% less. On reasoning it's 40% less, and on Mixtral 65% less. On agents and the four-copy 405B layout, the MI350X's performance gain outweighs its price, and it's 6–10% cheaper for the same work.

Faster vs. more economical

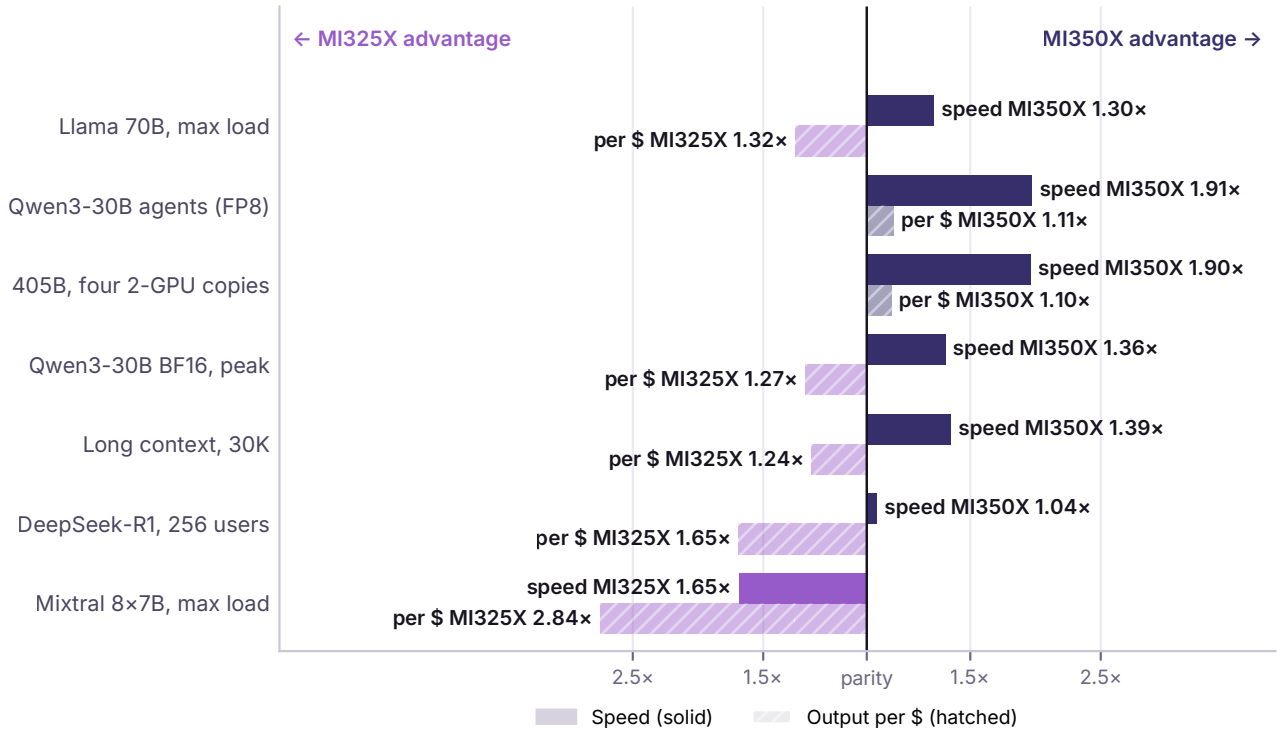

Figure 10. For each workload, which GPU is faster (solid) and which delivers more output per dollar (hatched). Bars to the right favor the MI350X, to the left the MI325X. Source: B8–B10.

> **The fastest GPU and the most economical GPU are not always the same GPU.**

09 · ZIANI

Two very different power profiles MEASURED BY BENTAUS

Under Ziani power caps, the two GPUs behaved in opposite ways. The MI325X draws close to its full 1,000 W budget and gives up little output when capped. The MI350X draws less power out of the box (6.4 kW vs 7.4 kW per node at maximum load) and is already near its efficient operating point, so caps reduce output almost in proportion to power.[B8, B9]

Independent study by Bentaus. All data points measured solely by Bentaus.

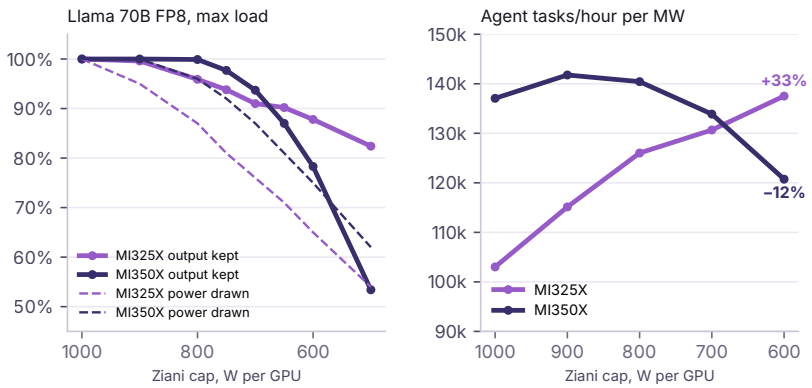

Figure 11. Left: Llama 70B FP8 at maximum load (served), output kept and power drawn at each cap, 3 runs per cap. Right: Llama 70B agent tasks per hour per MW of GPU power, 8 agents per GPU, counting all completed tasks. Source: B8, B9.

MI325X: greater power flexibility

At 600 W: 88% of maximum-load output on 65% of the power, and **+33% agent tasks per MW**. Capped, it matches the uncapped MI350X on agent tasks per MW (137.5K vs 137.1K).

MI350X: higher efficiency at its natural operating point

Lower energy per token uncapped (0.28 vs 0.39 J on AMD's benchmark). Caps help little: at most +3% agent tasks per MW (900 W), and −12% at 600 W.

Ziani brings the lower-cost MI325X to the same agent work per megawatt as an uncapped MI350X: 137.5K vs 137.1K tasks per hour per MW, while keeping a flexible load profile that can shed a third of its power with a modest loss of output.

Independent study by Bentaus. All data points measured solely by Bentaus.

10 · ENTERPRISE SCALE

How much infrastructure buys the same output? ECONOMIC MODEL

Enterprises size deployments to an output target. Using the measured Llama 70B FP8 result (19,086 vs 24,794 output tok/s per 8-GPU node), here's the MI325X fleet needed to match an MI350X fleet's output, and what each costs to rent.

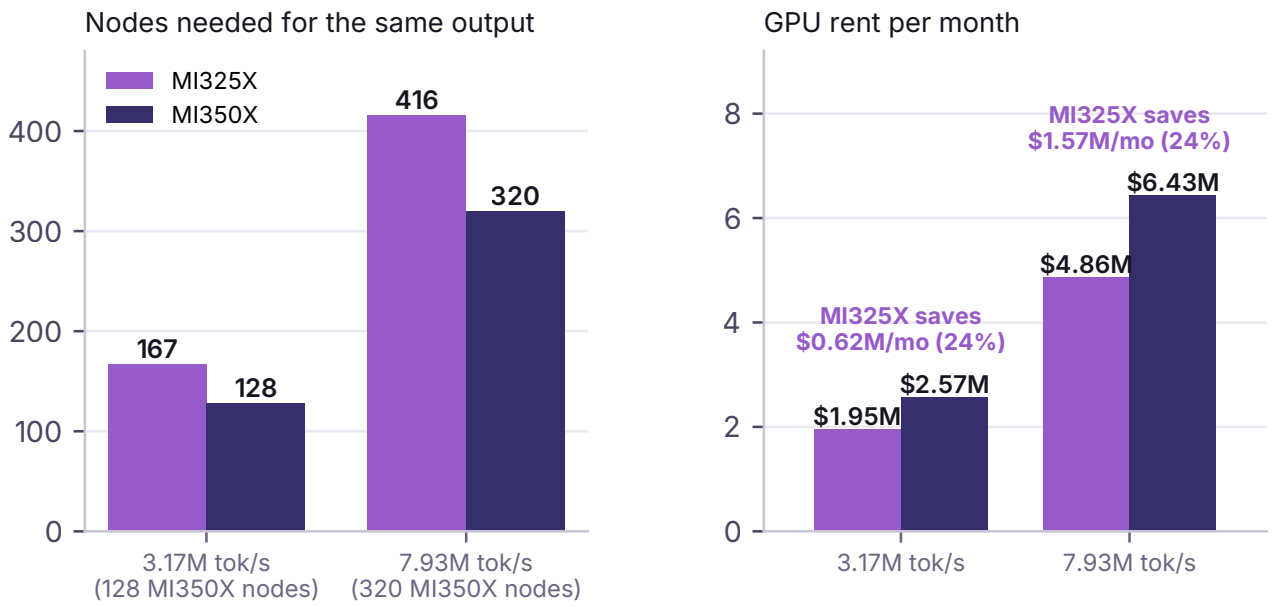

Figure 12. Same output, Llama 70B FP8 at maximum load. Rent at $16.00 vs $27.52 per node-hour, 730 hours per month.

MI350X fleet	Output	MI325X nodes for the same output	MI350X rent / month	MI325X rent / month	MI325X saves
64 nodes (512 GPUs)	1.59M tok/s	84 nodes (672 GPUs)	$1.29M	$0.98M	$0.30M (24%)
128 nodes (1,024 GPUs)	**3.17M tok/s**	**167 nodes (1,336 GPUs)**	**$2.57M**	**$1.95M**	**$0.62M (24%)**
256 nodes (2,048 GPUs)	6.35M tok/s	333 nodes (2,664 GPUs)	$5.14M	$3.89M	$1.25M (24%)
320 nodes (2,560 GPUs)	**7.93M tok/s**	**416 nodes (3,328 GPUs)**	**$6.43M**	**$4.86M**	**$1.57M (24%)**

128 MI350X nodes produce the same Llama 70B output as about 167 MI325X nodes, and the larger MI325X deployment still costs about 24% less in GPU rent: $0.62M a month. At 320 MI350X nodes, it takes about 416 MI325X nodes, saving about $1.57M a month.

Independent study by Bentaus. All data points measured solely by Bentaus.

The extra nodes bring extra memory

Matching the MI350X's output takes more MI325X nodes, 39 more at 128 MI350X nodes and 96 more at 320. Those nodes come with their own HBM. So the equal-output MI325X fleet also holds more memory and more resident context than the MI350X fleet it matches.

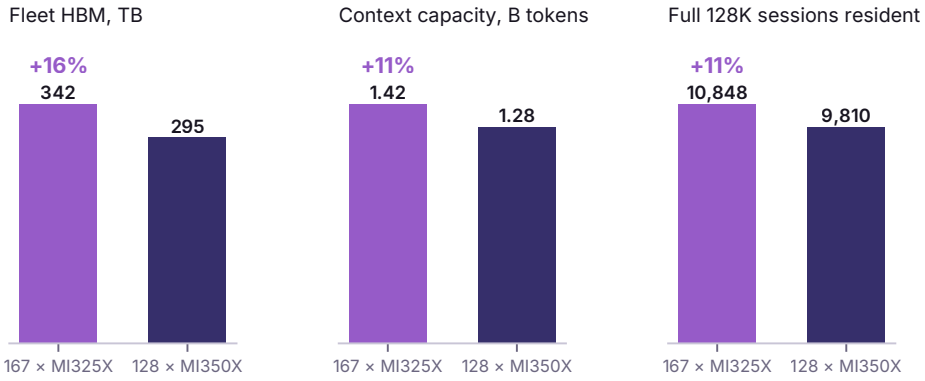

Figure 13. Equal Llama 70B output: 167 MI325X nodes vs 128 MI350X nodes. Context capacity and 128K sessions from vLLM's measured KV-cache size per node (8.51M vs 10.04M tokens).

Same output as	Fleet HBM	Context capacity	Full 128K sessions resident	GPU rent per month
128 MI350X nodes vs 167 MI325X	295 vs 342 TB (+16%)	1.29B vs 1.42B tokens (+11%)	9,810 vs 10,848 (+11%)	MI325X 24% lower
320 MI350X nodes vs 416 MI325X	737 vs 852 TB (+16%)	3.21B vs 3.54B tokens (+10%)	24,525 vs 27,023 (+10%)	MI325X 24% lower

For memory-bound work (long context, large resident session pools, RAG), the equal-output MI325X fleet carries about 10–16% more capacity at about 24% lower GPU rent. For agents, the gap doesn't close: in our tests, agents were limited by per-step latency, not memory, so the MI350X still held more agents within the performance target even against the larger MI325X fleet.

Same number of nodes

320 nodes, 2,560 GPUs	MI325X	MI350X
Output (Llama 70B FP8, max load)	6.11M tok/s	7.93M tok/s · **+30% output**
GPU rent per month	$3.74M · **42% lower cost**	$6.43M

The tradeoff is clean. **The MI350X wins compute density:** fewer nodes, racks and network ports for the same output. **The MI325X wins rental economics** by deploying more of them.

Linear fleet illustration based on measured single-node throughput. Excludes differences in networking, facility, rack, CPU, storage and other infrastructure costs associated with additional MI325X nodes. Savings are GPU rental savings under this model, not total cost of ownership.

Independent study by Bentaus. All data points measured solely by Bentaus.

11 · WHERE THE MI350X SHINES

When the MI350X earns its premium

MEASURED BY BENTAUS ECONOMIC MODEL

On some workloads the MI350X's gain is large enough that it needs fewer nodes *and* costs less for the same work. Agents are the clearest case.

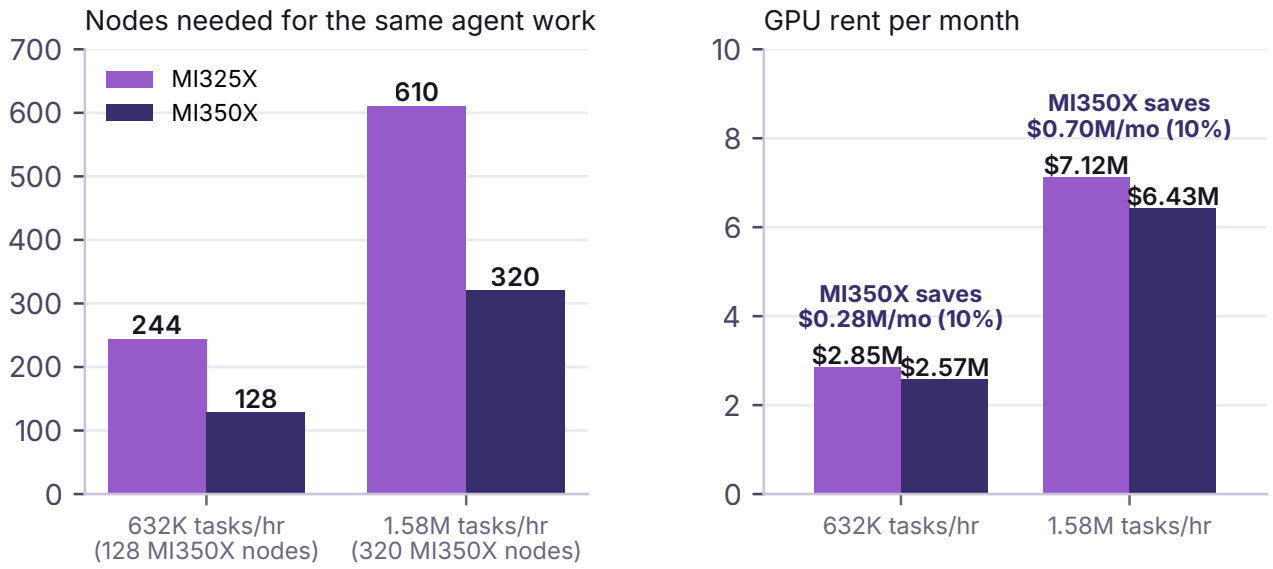

Figure 14. Same agent output, Qwen3-30B FP8 (4,942 vs 2,594 tasks per hour per node at the highest load meeting the performance target). Same rent assumptions as Figure 12; linear fleet illustration.

To match 128 MI350X nodes of agent work, the MI325X needs about 244 nodes, and costs about 11% more. At 320 MI350X nodes it takes about 610 MI325X nodes. The MI350X saves money, space and network at once.

Agentic AI

1.84–1.91× tasks per hour, 4× the agents held within target on Qwen3-30B, and 6–10% lower cost for the same agent work.

Large models, many copies

1.90× on Llama 405B as four 2-GPU copies, where its extra 32 GB per GPU leaves room for users' context. About 9% lower cost for the same output.

Latency at high concurrency

Stayed within the performance target to 768 users per GPU on Qwen3-30B, where the MI325X first missed it at 128.

Native FP4 and performance density

MXFP4 at 90% of FP8 speed, plus about 23% fewer nodes for equivalent Llama 70B output. On the measured Llama maximum-throughput workload, the MI350X also used 29% less GPU energy per output token.

Independent study by Bentaus. All data points measured solely by Bentaus.

12 · FIT

Where each GPU fits best

MEASURED BY BENTAUS

Performance and economics can point to different GPUs. The table shows both for each workload.

Workload	Performance	Economics (per $)
Llama 70B, maximum load	MI350X 1.30×	MI325X 1.32×
Qwen3-30B, high concurrency	MI350X 1.33–1.36×	MI325X 1.27–1.29×
Qwen3-30B agents	MI350X 1.88–1.91×	MI350X 1.09–1.11×
Llama 70B agents	MI350X 1.84×	MI350X 1.07×
405B, four 2-GPU copies	MI350X 1.90×	MI350X 1.10×
405B, best layouts	MI350X 1.33×	MI325X 1.30×
Long context	MI350X 1.35–1.39×	MI325X 1.24–1.27×
Mixtral 8×7B, maximum load	MI325X 1.65×	MI325X 2.84×
DeepSeek-R1	Near parity (MI350X +4%)	MI325X 1.65×
Native 4-bit inference	MI350X	Workload-dependent

Economics at current NeoCloud market rates as of October 2026 ($2.00 vs ~$3.44 per GPU-hour). Power profiles are compared in section 09.

Independent study by Bentaus. All data points measured solely by Bentaus.

13 · CUSTOMERS

Who should use each GPU?

MI325X: large-memory, cost-efficient FP8 inference with a highly flexible power profile

Customer / workload	Examples	Why the MI325X
Enterprise inference ●	Internal copilots, document intelligence, knowledge assistants, enterprise search	Strong FP8 inference; 1.32× tokens per $
High-volume API providers ●	Summarization, extraction, classification, translation, content generation	The MI350X's 1.3× gain on general inference doesn't clear its price premium
RAG and long context ●	Legal corpora, financial filings, medical records, engineering manuals, code repositories	256 GB per GPU; 1.24–1.27× tokens per $ on long context
Reasoning services ●	DeepSeek-R1 applications, research assistants, analytics	Near speed parity; 1.65× output per $
MoE inference ●	Mixtral-class serving	1.65× faster at maximum load on the tested software path
Power-constrained AI factories ●	Fixed MW allocations, demand response, flexible compute	+33% agent tasks per MW at 600 W with Ziani
Sovereign and private AI	Private enterprise LLMs, regulated workloads, government models	Large memory for self-hosted open models
Research and HPC	CFD, climate, physics, engineering simulation	256 GB HBM and strong FP64; AMD positions MI325X for AI and HPC

● Validated in this study. Other rows: architecture fit, not benchmarked in this study.[A1]

MI350X / MI350 Series: premium inference and agentic compute

Measured rows are MI350X results. The MI355X is the higher-power, liquid-cooled MI350 Series option; this study doesn't attribute MI350X measurements to it.

Independent study by Bentaus. All data points measured solely by Bentaus.

Customer / workload	Examples	Why the MI350 Series
Agentic AI platforms ●	Coding, research and customer-service agents, enterprise automation, tool-calling systems	1.84–1.91× tasks per hour; wins tasks per $
Frontier-model inference ●	405B-class and very large MoE models	1.90× on 405B as four 2-GPU copies; 288 GB HBM
High-concurrency inference ●	Large SaaS AI services, multi-tenant endpoints	Within the performance target to 768 users per GPU
FP4 inference ●	High-volume generation where FP4 quality has been validated	Native MXFP4 at 90% of FP8 speed vs 52%
Power-limited premium compute ●	AI factories maximizing useful compute per MW	29% less energy per token at full power
Model developers and AI labs	Fine-tuning, post-training, RL rollouts, distillation	More compute and bandwidth; AMD positions MI350 Series for training
Scientific and HPC organizations	Simulation, computational modeling, large-scale data processing	AMD-stated target markets for the MI350 Series

● Validated in this study. Other rows: architecture fit, not benchmarked in this study.[A3]

Frontier AI and research

Large-model inference, post-training, reasoning, RL rollouts, synthetic data, distillation and agent evaluation. University and national labs: CFD, climate, molecular simulation, genomics, seismic and scientific ML. Enterprise AI groups: domain models, RAG, coding agents and multimodal AI.

Independent study by Bentaus. All data points measured solely by Bentaus.

14 · PLACEMENT

Workload placement guide

Workload	MI325X	MI350X
General FP8 inference	FP8 value: ~24% lower modeled GPU rent for equivalent Llama 70B output	FP8 performance: 1.30× output per node
Agentic AI	Strong	Performance and tasks per $: 1.84–1.91×, 6–10% lower cost
High concurrency	Economics: ~21–23% lower cost	Performance and latency: within target to 768 users per GPU
Long context	Economics: ~19% lower cost	Performance: 1.35–1.39×
405B+ models	Strong	Performance and density: 1.90× as four 2-GPU copies
Native FP4	—	Native FP4: 90% of FP8 speed
Mixtral (tested stack)	Performance and economics: 1.65× faster	—
DeepSeek-R1	Economics: ~40% lower cost	Slight speed advantage (+4%)
Flexible-load inference	Ziani flexibility: +33% agent tasks per MW	Natural efficiency: 29% less energy per token

Additional workloads in the Bentaus characterization roadmap include video inference, multimodal agents, vision-language models, speech and voice inference, and video generation.

At 2,560 GPUs, the question isn't which chip has the biggest benchmark

Frontier AI lab

How many models, agents and experiments can I complete per MW and per dollar?

Enterprise

How many users and tasks can I serve inside my latency and budget targets?

NeoCloud

Which GPU maximizes revenue and utilization per GPU-hour?

Research lab

What produces the shortest time-to-solution within the power allocation?

Data-center operator

How much useful compute can I run, and how much of that load can Ziani make flexible?

Independent study by Bentaus. All data points measured solely by Bentaus.

15 · METHOD

Test setup

	AMD Instinct MI325X	AMD Instinct MI350X
GPUs	8 × MI325X OAM, 256 GB HBM3E, 1,000 W[A1]	8 × MI350X OAM, 288 GB HBM3E, 1,000 W[A2]
Location	Bentaus us-lab, San Jose	
Software	vLLM 0.23.0 on ROCm 7.14.1 (pinned rocm/vllm image); AMD's benchmark also on AMD's vLLM 0.11.2 image	
AMD AITER kernels	On for Llama and Mixtral; attention-only for Qwen agents	On for Llama FP8; off for Llama BF16, Mixtral and Maverick, where it produced invalid output on this stack; attention-only for Qwen agents
Models	Llama 3.1 8B, 70B and 405B FP8 (AMD); Llama 70B BF16, AWQ and MXFP4; Qwen3-30B-A3B BF16 and FP8; Qwen3-235B FP8; Llama 4 Maverick and Scout FP8; Mixtral 8×7B FP8; DeepSeek-R1-0528	
Performance target	Users: time to first token ≤2 s and inter-token latency ≤200 ms (P99). Agents: every call ≤3 s to first token and ≤200 ms per token, for ≥90% of tasks.	
Power	GPU board power, read every second. Default 1,000 W caps except in Ziani sweeps (1,000 → 400 W, 3 runs per cap).	

Technical notes. DeepSeek-V4-Flash isn't included: on vLLM 0.23 it didn't produce valid output on either GPU. The MI350X's 400 W maximum-load step ended in a GPU runtime resource error after 21 runs and is excluded. One node per GPU type; most points are single runs, power-cap steps are three runs. GPU board power, not wall power. Results apply to the configurations described. Prices are current NeoCloud market rates as of October 2026; the MI350 Series rate reflects available MI350 Series listings, which were MI355X.

BENTAUS PRIMARY RESEARCH

[S1] Bentaus, *Where the AMD Instinct™ MI325X Wins*, executive brief, October 2026.

[S2] Bentaus, *Where the AMD Instinct™ MI325X Wins*, technical evaluation, October 2026.

[S3] Bentaus, *NVIDIA H100: Where It Performs Best*, technical evaluation, October 2026.

[S4] Bentaus, *AMD Instinct™ MI325X vs MI350X*, technical evaluation, October 2026 (this paper).

[B8] Bentaus, "MI325X: test results," test report, 4 Oct 2026.

[B9] Bentaus, "MI350X: test results," test report, 4 Oct 2026.

[B10] Bentaus, "MI325X vs MI350X," comparison report, 4 Oct 2026.

[B11] Bentaus, Ziani GPU power-cap validation results, October 2026.

TECHNICAL AND VENDOR REFERENCES

[A1] AMD, Instinct MI325X data sheet. amd.com/en/products/accelerators/instinct

[A2] AMD, Instinct MI350X specifications. amd.com/en/support/downloads/drivers.html/accelerators/instinct/instinct-mi350-series/amd-instinct-mi350x.html

[A3] AMD, Instinct MI350 Series (MI350X, MI355X). amd.com/en/products/accelerators/instinct/mi350.html

[O1] vLLM project. github.com/vllm-project/vllm

AMD, AMD Instinct, CDNA and ROCm are trademarks of Advanced Micro Devices, Inc.

Disclaimer: Independent study conducted by Bentaus. All data points were measured solely by Bentaus. AMD did not sponsor, commission or fund this study.

BENTAUS NEOCLOUD · ENERGY-FIRST

Performance starts with the workload.

Different workloads reward different accelerator characteristics. Bentaus tests the model, precision, layout, concurrency, economics and power profile together to determine where the workload runs best.

Bring us the workload. We'll show you how it runs.

MI350X · built for agents
1.84–1.91× tasks per hour; lower cost per task

MI325X · built for FP8 value
Strong FP8 inference at lower cost for equivalent output

MI350X · built for large models
1.90× on 405B as four 2-GPU copies

MI325X · built for output economics
~24% lower cost for the same Llama 70B output

MI350X · built for performance per watt
29% less energy per token at full power

MI325X · built for memory economics
10–16% more fleet memory and context capacity in the equal-output deployment model

MI350X · built for native FP4
MXFP4 at 90% of FP8 speed

MI325X · built for power flexibility
+33% agent tasks per MW with Ziani

Bentaus