PURETENSOR
INTELLIGENCE DISTILLED

What does a 763B model lose at three bits?

From the Lab | PT-R-2026-079

PT-R-2026-079 · PureTensor Research · v1.0

Research completed 28 November 2026 · Published 28 November 2026

Result

PureTensor evaluated whether the diffbot calibrated EXL3 3-bit-per-weight pack of DeepSeek-V4.1-Flash-1017, a 763B-parameter model at 1,048,576-token context, preserves output quality when served on two workstation cards instead of four. On 28 November 2026 we compared two production deployments: the 4-GPU seat, running native FP8 and FP4 weights across 4 NVIDIA RTX PRO 6000 Max-Q cards, and the 2-GPU seat, running the EXL3 pack across 2 NVIDIA RTX PRO 6000 cards with host RAM offloading. Across teacher-forced fidelity, accuracy suites, competition maths, writing, and our private benchmark, the 2-GPU seat showed no measurable quality loss. One nominally significant long-context result favoured the 3-bit seat and disappeared on repetition. The cost is throughput: between a third and two fifths of single-stream speed, with scaling stopping at 4 concurrent streams.

Setup

Both seats serve DeepSeek-V4.1-Flash-1017, a 763B-parameter model, with a 1,048,576-token context window.

The 4-GPU seat, listed as Seat A in the tables below, runs the source checkpoint with its native FP8 and FP4 weights under vLLM, tensor parallel across 4 NVIDIA RTX PRO 6000 Max-Q cards. All 100% of the model weights reside in GPU memory. Serving uses a fused decoder and adaptive DSpark speculative decoding (K7).

The 2-GPU seat, listed as Seat F in the tables below, runs the diffbot calibrated EXL3 3-bit-per-weight pack of the same model under vLLM, tensor parallel across 2 NVIDIA RTX PRO 6000 cards. About two thirds of the weights sit in GPU memory; about one third of the routed experts and the Engram tables sit in pinned host RAM, backed by a 512-slot GPU expert cache. Speculative decoding uses DSpark with k=5. This seat serves at most 8 concurrent requests.

These two seats therefore differ in weights (3-bit quantised against the source checkpoint), in GPU count (2 against 4), and in serving engine configuration. The evaluation compares seat against seat, as an operator experiences them in production. It does not isolate quantisation from the serving engine, nor does it separate weight precision from host-memory offloading, decoder implementation, or speculative decoding depth. Every measurement in this paper belongs to the complete seat rather than to EXL3 alone.

All runs took place on 28 November 2026 with both seats in production configuration and the evaluation clients running on a third machine. We used the same evaluation harness as our 31 October 2026 NVFP4 study.

Fidelity Results

Token-level fidelity was measured by teacher-forced evaluation over 207,735 tokens spanning code, notes, and reports. Before comparing seats, we measured the run-to-run floor of the 4-GPU seat against itself. Across repeated passes on the same seat, perplexity read 5.0110 and 5.0038, top-1 token agreement was 93.71%, and KL divergence over the top 5 tokens was 0.01531.

Comparing the 4-GPU seat against the 2-GPU seat over the same 207,735 tokens gave a perplexity of 5.0110 against 5.0490, a change of +0.76%. Top-1 agreement between seats was 91.83%, and KL divergence over the top 5 tokens was 0.02390. By corpus, perplexity on the 4-GPU seat against the 2-GPU seat measured 2.2812 against 2.2935 on code, 10.9503 against 11.0432 on notes, and 6.2177 against 6.2792 on reports. For reference, NVIDIA's NVFP4 build of the same model measured +1.9% perplexity in our October study.

Greedy generation was checked by exact match of whole outputs. The 4-GPU seat matched itself on 36% of prompts, while the 2-GPU seat matched itself on 42%. Between the two seats, whole-output exact match was 28%. Because repeated greedy runs on a single seat already diverge on most whole outputs, the cross-seat figure of 28% sits below, but not far from, the 36% that identical hardware and weights achieve against themselves.

Accuracy Results

Unless stated otherwise, accuracy suites drew a single sample per item at temperature 1.0, and items that hit the token budget were re-run once at double budget. Table 1 summarises the paired binary suites.

Suite	Items	Seat A correct	Seat F correct	A-only right / F-only right	McNemar p
Knowledge (SimpleQA subset, thinking off, blind graded)	300	82	77	26 / 21	0.56
Reasoning (GPQA, AIME 2026, HMMT Feb 2026; thinking high)	261	239	235	8 / 4	0.39
Coding	95	94	94	1 / 1	1
Vision (charts, documents, tables, UI)	116	108	110	2 / 4	0.69
Long context, 32k to 240k tokens	180	158	168	1 / 11	0.0063
Long context, 222k tokens	45	34	36	1 / 3	0.62

On the 300-item Knowledge suite (a SimpleQA subset run with thinking off, blind graded, and hand checked across every grade), the 4-GPU seat recorded 82 correct, 106 incorrect, and 112 not attempted. The 2-GPU seat recorded 77 correct, 116 incorrect, and 107 not attempted. Discordant items split 26 to 21 (McNemar p = 0.56).

On Reasoning (261 items from GPQA, AIME 2026, and HMMT Feb 2026 at thinking high), the 4-GPU seat answered 239 correctly against 235 for the 2-GPU seat, with 8 against 4 discordant items (McNemar p = 0.39). Before the double-budget re-runs, the scores stood at 220 against 216. To reduce variance on competition maths (AIME 2026 and HMMT February 2026), we also

ran four samples per problem, totalling 252 samples per seat. The 4-GPU seat reached 95.2% (95% interval 90.5 to 98.8), and the 2-GPU seat reached 92.9% (95% interval 87.3 to 97.6); the intervals overlap.

On Coding (95 items), both seats scored 94, with 1 item solved only by each seat (McNemar p = 1). Before the double-budget re-runs, the counts were 94 for the 4-GPU seat and 93 for the 2-GPU seat. On Vision (116 items across charts, documents, tables, and UI), the 4-GPU seat scored 108 against 110 for the 2-GPU seat, with 2 against 4 discordant items (McNemar p = 0.69).

On the 222k-token long-context suite (45 items), the 4-GPU seat scored 34 against 36 for the 2-GPU seat, with 1 against 3 discordant items (McNemar p = 0.62). On the 32k to 240k-token suite (180 items), the 4-GPU seat scored 158 against 168 for the 2-GPU seat, with 1 against 11 discordant items and McNemar p = 0.0063. Before double-budget re-runs, that suite read 158 against 162. The section on the anomaly traces this result and explains why it did not survive repetition.

Two further evaluations sit outside Table 1. On Writing (30 prompts, blind pairwise judging in both presentation orders), the 2-GPU seat won 16 comparisons, the 4-GPU seat won 10, and 4 ended in ties, with mean scores of 7.60 against 7.45 out of 10. However, the two judging orders agreed on the winner for only 12 of 30 prompts, so the writing evaluation is a tie. On our private frontier-model benchmark (low reasoning effort, temperature 0, three runs), the 4-GPU seat scored 82.5 and the 2-GPU seat scored 81.7.

The Anomaly

In Table 1, a single entry produces a low p-value: on the 32k to 240k-token long-context suite, the 2-GPU seat scored 168 of 180 against 158 of 180 for the 4-GPU seat, with 1 against 11 discordant items and McNemar p = 0.0063. Read without scrutiny, that statistic claims that a 3-bit quantisation outperformed the full-precision source weights on long-context processing. Because a 3-bit pack beating the source checkpoint is not a credible effect, we investigated the gap before reporting it.

Breaking the suite down by task family isolated the entire difference to one subset. Of the 180 items, 140 covered multi-hop, multi-key, variable tracking, and real question answering; both seats scored 100% on all 140. The remaining 40 items were counting tasks that ask how many times an item appears in a long document. Every discordant item sat inside those 40 tasks, where the 2-GPU seat scored 28 of 40 and the 4-GPU seat scored 18 of 40.

Inspection of the budget recovery step revealed how the disparity grew. Before double-budget re-runs, the 180-item suite stood at 158 against 162. Re-running budget-exhausted items is a second single draw performed exclusively on the least stable prompts. In that pass, the 4-GPU seat recovered 0 of 12 items, while the 2-GPU seat recovered 6 of 10. Two separate single-sample draws compounded on the same unstable subset.

We tested whether the gap held under replication by running all 40 counting tasks with 3 fresh samples per item (120 samples per seat) at the doubled budget of 16,384 tokens. The gap vanished. The 4-GPU seat scored 65 of 120, and the 2-GPU seat scored 69 of 120. Across context lengths, with 30 samples per cell, the 4-GPU seat against the 2-GPU seat scored 22 against 23 at 32k tokens, 23 against 25 at 64k tokens, 13 against 15 at 128k tokens, and 7 against 6 at 240k tokens.

Per-item stability across those 3 samples explains why the original p-value of 0.0063 was wrong. On each seat, 18 of the 40 counting tasks produced both right and wrong answers across their 3 samples. McNemar's test assumes that every item has a fixed binary outcome on each system, treating each discordant pair as a deterministic disagreement. At temperature 1.0, these counting tasks violate that assumption. Single-sample evaluation turns stochastic

draws on volatile items into apparent paired disagreements, and a conditional double-budget retry compounds the noise. The resulting p-value understates true variance. Because the false positive here pointed in an implausible direction, favouring the 2-GPU seat over the 4-GPU seat, it forced an audit. Had the same noise landed in the opposite direction, favouring full precision, a single-sample test would have invited a false claim of quantisation loss.

Speed

Decode throughput was measured as served, with no tuning window, at temperature 0, reporting streamed decode tokens per second. Table 2 lists the results across single-stream prompts and concurrent loads.

Test	Seat A (4 GPUs)	Seat F (2 GPUs)
Single stream, reasoning prompt, thinking off	283.8	91.5
Single stream, code prompt, thinking off	486.6	190.4
Single stream, long webhook prompt, thinking off	535.9	220.5
Single stream, reasoning prompt, thinking high	320.2	104.6
Single stream, code prompt, thinking high	399.8	166.8
2 concurrent streams, total	652.8	244.9
4 concurrent streams, total	953.6	318.1
8 concurrent streams, total	1,223.7	322.2
16 concurrent streams, total	1,736.9	not run (seat serves at most 8)

Across single-stream prompts, the 4-GPU seat is 2.4 to 3.1 times faster than the 2-GPU seat. Time to first token on those prompts is about 0.2 s on the 4-GPU seat, compared with 0.33 to 0.56 s on the 2-GPU seat.

Under concurrency, the gap widens. The 4-GPU seat scales from 652.8 tokens per second at 2 streams to 953.6 at 4 streams, 1,223.7 at 8 streams, and 1,736.9 at 16 streams. The 2-GPU seat rises from 244.9 tokens per second at 2 streams to 318.1 at 4 streams, then flattens at 322.2 tokens per second at 8 streams. Scaling stops at 4 concurrent streams, leaving the 4-GPU seat 3.8 times faster at 8 streams. The 16-stream test was not run on the 2-GPU seat because the seat serves at most 8 concurrent requests. As noted in the setup, these throughput numbers reflect the complete serving stacks rather than EXL3 dequantisation alone.

Method Disclosures

Five procedural points and threats to validity apply to this study.

1. The first pass of four suites (knowledge, vision, writing, and coding) on the 4-GPU seat, along with its first private-benchmark run, was lost when the launch of the second arm synchronised over the same client working directory. All affected runs were repeated from separate directories on the same day. No result was selected between passes; the lost pass was never scored.
2. The 2-GPU seat reasons by default, whereas the 4-GPU seat does not. The first pass of the two thinking-off suites on the 2-GPU seat, knowledge and vision, therefore ran with

thinking on. That left 197 of 300 knowledge answers cut off at the 512-token budget. Both suites were re-run on the 2-GPU seat with thinking explicitly off, and only the re-run is reported.

3. Four competition-maths answers that the automatic scorer could not decide were checked by exact arithmetic. All four were correct: three on the 4-GPU seat and one on the 2-GPU seat.

4. Writing comparisons were judged by language-model judges that saw neither the seat labels nor the key.

5. Accuracy suites other than competition maths use one sample per item (with three runs on the private benchmark at temperature 0 and three samples per item in the counting repeat test), so differences of a few items sit within sampling noise.

Conclusion

Within the limits of a seat-against-seat comparison, the EXL3 3-bit pack of DeepSeek-V4.1-Flash-1017 showed no measurable quality loss on any evaluated suite. Teacher-forced perplexity rose by +0.76%, every accuracy gap fell within sampling noise, and the single nominally significant difference favoured the 3-bit seat and vanished upon repeated sampling.

The price of running the pack on two cards is speed. The 2-GPU seat delivers between a third and two fifths of the single-stream throughput of the 4-GPU seat, adds latency to the first token, and stops scaling at 4 concurrent streams. For engineers choosing between these deployments, the evidence shows that a 763B model at 1,048,576-token context runs with intact quality on two 96 GB workstation cards when lower throughput and concurrency are acceptable.

PureTensor Research · PT-R-2026-079 · v1.0