PURETENSOR
INTELLIGENCE DISTILLED

Same Answers, Different Confidence

4-bit weights preserve accuracy and destroy calibration

Paper: PT-R-2026-018 v1.0

Author: PureTensor Ltd

Date: 8 September 2026

Status: Published

Abstract

Quantised checkpoints are evaluated almost exclusively on accuracy: the fraction of items a model answers correctly. When a 4-bit checkpoint matches its full-precision parent on that metric, it is declared equivalent and deployed. We report a case in which this framing misses the one dimension that did not tie. On 2 September 2026 we ran GLM-5.3-Flash, a mixture-of-experts model with multi-head latent attention, in two configurations on the same eight-GPU Hopper node: the original BF16 weights (328 GB) and an NVFP4 quantisation (198 GB) in which only the MLP experts are stored at 4 bits. Both checkpoints completed our internal frontier benchmark (94 original items, ten dimensions, every verdict produced by code, no language-model judge) at temperature 0 with 100 per cent coverage. Every per-dimension accuracy delta was one item, inside Wilson 95 per cent intervals that span 30 to 50 points on samples of 6 to 12 items. On accuracy the quantisation is a tie.

The exception is calibration. Dimension 10 of our benchmark asks the model to state both an answer and a probability that the answer is correct. Both checkpoints achieved 85.0 per cent accuracy on these items. The BF16 checkpoint returned a Brier score of 0.070; the NVFP4 checkpoint returned 0.170, within reach of the 0.1875 random floor. The 4-bit weights preserved the model's ability to choose the right answer and destroyed its ability to know when it had done so. Wrong answers now arrive with the same stated confidence as right ones.

This result has a structural explanation. The gap between a model's best and second-best answer is usually wide enough to survive coarse weight storage; the fine probability differences that carry calibration are exactly what 4-bit quantisation discards. Temperature 0 removes sampling variance but does not restore the signal that, in production, distinguishes the confident-and-correct from the confident-and-wrong. For agentic lanes whose outputs are verified downstream, the accuracy tie is sufficient. For any lane that consumes the model's own confidence (a judge, a scorer, an abstention gate), the calibration loss is the cost of the smaller checkpoint.

The measurement is a single run at k=1, t=0, on 12 calibration items, with one checkpoint from one model. We state these limits plainly. The follow-up experiment, not yet run, is both checkpoints at k=5, t=0.7, where a calibration gap would be expected to surface as an accuracy gap. What we can say now is that a benchmark reporting only accuracy would have declared these checkpoints equivalent, and that declaration would have been incomplete.

1. Motivation

Production systems that delegate to language models do not always have an answer key. A remediation agent deciding whether to retry or escalate, a scorer assigning a grade, an abstention gate deciding whether to answer at all: each of these lanes consumes the model's stated confidence as an input to a downstream decision. If the model's confidence stops tracking correctness, the downstream decision inherits the error.

Quantisation benchmarks rarely measure this. The standard protocol is to run a held-out set, count correct answers, and compare. When the count matches, the checkpoint is approved. We wanted to know what happens to the dimensions that do not reduce to a count.

2. Experimental Setup

We used GLM-5.3-Flash, a mixture-of-experts model with multi-head latent attention, in two checkpoints from the same publisher lineage. The first is the original BF16 weights, occupying 328 GB. The second is an NVFP4 quantisation occupying 198 GB, in which only the MLP experts are stored at 4 bits; the attention layers, shared experts, embeddings, and dense MLP remain in higher precision. The NVFP4 checkpoint was produced by a third party from the publisher's weights.

Hardware was a rented node with eight H100 SXM5 GPUs (80 GB each, connected via NVSwitch). This node was chosen because it is the only configuration in which both checkpoints fit side by side under identical conditions. On Hopper there is no native FP4 arithmetic, so the NVFP4 checkpoint runs through a weight-only FP4 kernel (Marlin): the 4-bit weights are dequantised on the fly and the arithmetic proceeds at the same precision as BF16. This isolation is important. Any difference between the two runs is a property of the stored weights, not of the arithmetic path.

Serving used vLLM (a nightly build with native support for this model) inside Docker, with tensor parallelism across all eight GPUs, a 131,072-token context window, and 90 per cent GPU memory utilisation. Both checkpoints pulled at approximately 2.7 GB/s (about 3 minutes); the interval from download to final verdict was about 35 minutes; the node was held for approximately 1.5 hours in total.

The benchmark was version 1.1 of our internal frontier benchmark: 94 original items across ten dimensions, every verdict produced by code, no language-model judge. Both runs were single-shot at temperature 0 (k=1, t=0) in the agentic seat. Both achieved 100 per cent coverage and were flagged comparable by the harness.

Dimension 10, calibration, deserves separate description. Items in this dimension ask the model for an answer and a stated probability that the answer is correct. Accuracy on D10 counts whether the letter chosen is right. The Brier score is the mean squared difference between the stated probability and the binary outcome (1 if correct, 0 otherwise). A Brier score of 0 is perfect calibration; for this item mix the random floor is 0.1875.

3. Accuracy Results

Table 1 presents per-dimension accuracy for both checkpoints on the same node, alongside a third column: the same NVFP4 checkpoint served on our own premises four days earlier (29 August 2026), using four Blackwell GPUs across two nodes in a cross-node tensor-parallel-4 configuration.

Dimension	BF16	NVFP4 (same node)	NVFP4 (on-premises, cross-node TP4)
D1 tool selection	75.0	75.0	62.5
D2 abstention	83.3	91.7	91.7
D3 tool arguments	100	100	100
D4 multi-step	83.3	83.3	83.3
D5 reasoning	100	100	100
D6 false premise	87.5	100	100
D7 instruction	91.7	83.3	91.7
D8 structured output	100	100	100
D9 grounding	100	100	100
D10 calibration (accuracy)	85.0	85.0	85.0
Agentic composite	88.2	89.6	88.3
Scorer composite	93.1	94.1	95.4
Flat composite	90.6	91.8	91.4

Every per-dimension delta between BF16 and NVFP4 is one item. Wilson 95 per cent confidence intervals on samples of 6 to 12 items span 30 to 50 points; none of these deltas escapes the noise. The composites agree within a point or two. The third column, from different hardware four days earlier, reproduces the NVFP4 accuracy result within noise, which gives us some confidence that the accuracy tie is not an artefact of the particular node.

On accuracy, the quantisation is a tie.

4. The Calibration Exception

Table 2 isolates the calibration dimension.

Metric	BF16	NVFP4
D10 accuracy	85.0	85.0
D10 Brier score	0.070	0.170
Random-floor Brier	0.1875	0.1875

Both checkpoints answered the same fraction of calibration items correctly. The difference is in the confidence attached to those answers. The BF16 checkpoint's Brier score of 0.070 indicates that its stated probabilities tracked outcomes reasonably well. The NVFP4 checkpoint's Brier score of 0.170 is 2.4 times worse and sits near the random floor. The 4-bit weights kept the answers and flattened the model's ability to distinguish confident-and-correct from confident-and-wrong.

This is not a case of the quantised model being less sure. It is a case of the quantised model's stated confidence no longer tracking correctness. Wrong answers arrive with the same tone as right ones.

5. Throughput

Table 3 presents throughput measurements for 512-token chat completions.

Metric	BF16	NVFP4
Single-stream decode (tok/s)	155–159	149–159
Aggregate at 4 concurrent	460	474
Aggregate at 8 concurrent (repeated run)	not recorded	915
Aggregate at 16 concurrent	510 (first hit)	542–598
Benchmark run aggregate at 16	1,246	1,396
KV cache at 90% utilisation (tokens)	2.59M	4.01M
Engine init (s)	326	160

First-hit numbers at a new batch size include CUDA-graph warm-up; repeated runs are the truth. The single-stream decode rates overlap. On Hopper the weight-only dequantisation path does not make NVFP4 faster per token. What it buys is approximately 12 per cent higher aggregate throughput and 55 per cent more KV cache (the weights are 130 GB smaller). Engine initialisation is roughly half as long.

For scale: the eight-H100 node delivers about 155 tokens per second single-stream either way, which is 1.5 times our on-premises cross-node NVFP4 single-stream and about twice its aggregate at 16 concurrent requests.

6. Why Calibration Matters When the Answers Tie

The structural argument is straightforward. Accuracy is coarse. A model's top answer is usually separated from its second-best answer by a margin wide enough to survive 4-bit weight storage. The fine probability differences that carry calibration are exactly what coarse storage discards.

Temperature 0 removes only the sampling cost: the model will not wander to its second-best answer. It does not restore the signal that, in production, identifies the wrong 10 per cent. There is no answer key in production. The system must rely on the model's stated confidence to decide when to trust itself, when to abstain, when to escalate.

This matters for any lane whose output is itself a confidence. A judge lane that scores another model's answer. A scorer lane that assigns a grade. An abstention gate that decides whether to answer at all. An escalate-or-fix decision in an autonomous remediation loop. Each of these lanes consumes the model's probability as an input. If that probability stops tracking correctness, the downstream decision inherits the error.

For agentic and tool lanes, where the output is an action that gets verified downstream, the accuracy tie is sufficient. The quantised checkpoint is, for these purposes, free. For any lane that consumes the model's confidence, the calibration loss is the cost.

Our own fleet practices already assume this separation. The autonomous remediation loop exits on independent re-verification, never on model certainty. Judgment stays with the orchestrator; delegated workers only mutate. Worker self-calibration, measured on a separate battery, sits around 70 per cent. The benchmark scores abstention and calibration as first-class dimensions precisely because accuracy alone would miss this failure mode.

7. What These Numbers Will Not Carry

The measurement is a single run per configuration at k=1, t=0. This setting cannot surface a calibration gap as an accuracy gap; the model's top answer is taken regardless of its confidence. The follow-up experiment, not yet run, is both checkpoints at k=5, t=0.7 (five

rollouts, an item passes only if all five pass). This is the production-like setting where a calibration gap would be expected to become an accuracy gap. Until that experiment is complete, we cannot say whether the calibration loss translates to an accuracy loss under sampling.

Dimension 10 contains 12 items. The Brier difference is computed on those 12 items. It is larger than any other signal in the table by a wide margin and is consistent in direction with the proposed mechanism, but it is one run on one dimension.

The NVFP4 checkpoint was produced by a third party from the publisher's weights. Any calibration loss could in principle be a property of that particular quantisation recipe rather than of 4-bit storage in general. We have tested one checkpoint from one model. Generalisation to other quantisation methods or other models is not established.

On this hardware the 4-bit weights are dequantised to full precision before arithmetic. This isolates the weight-storage effect from kernel-arithmetic effects. On hardware with native FP4 arithmetic, the two effects would compound, and the result could differ in either direction.

An earlier vLLM path was tried and abandoned on this node (a CUDA 13 build against a CUDA 12.8 driver, then a generic fallback without native support). The runs reported used the nightly image with native support for both checkpoints. We do not believe the earlier attempts affected the final measurements, but we note them for completeness.

The cross-hardware reproduction (the on-premises run four days earlier) did not record a Brier score for the NVFP4 checkpoint. The accuracy reproduction is therefore stronger than the calibration result, which rests on the single same-node comparison.

8. Next Steps

The immediate follow-up is the k=5, t=0.7 experiment: both checkpoints, five rollouts per item, an item passes only if all five pass. If the calibration gap surfaces as an accuracy gap under sampling, the deployment guidance becomes simpler. If it does not, the distinction between accuracy lanes and confidence-consuming lanes remains the operative one.

We also hold an FP8 checkpoint of the same model. Running it through the same protocol would show whether the calibration loss is a function of bit width (FP8 should then sit between BF16 and NVFP4) or specific to this quantisation recipe.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. Measurements were taken on 2 September 2026; raw artefacts are retained. This paper is PT-R-2026-018.