PURETENSOR
INTELLIGENCE DISTILLED

The Blocker Was Twelve Lines

A 4-bit checkpoint rejected on Thursday and serving the following Wednesday

Paper: PT-R-2026-020 v1.0

Author: PureTensor Ltd

Date: 15 September 2026

Status: Published

Abstract

On 4 September 2026 we rejected a 4-bit quantised checkpoint of our resident model after it decoded at 8.69 tokens per second, roughly one-tenth of the incumbent. The rejection was logged as final: the blocker appeared to be the absence of any serving image that carried both a named upstream commit and the ability to offload a 47.68 GiB embedding table to host memory. Five days later a re-read of the engine's source showed the diagnosis was wrong. The upstream commit named in the checkpoint's README did not add offload capability; it extended an existing offload path to recognise mixed-precision configurations. The image we needed had shipped with the model itself, a week before our evaluation. A twelve-line backport restored the path, and the same checkpoint then served at 110.1 tokens per second single-stream, 3,881 tokens per second aggregate at 64 concurrent requests, and 2.25 times the KV-cache capacity of the FP8 resident.

We report both days in full because the failure mode is instructive. The 4-bit experts were never the obstacle; a fixed-size lookup structure that does not fit is a fit problem, not a quantisation problem. Following a README's stated requirement led us to the one image that satisfied it rather than to the image that had the capability the seat actually needed. The twelve lines that unblocked the probe were a single-file copy over a pinned image digest.

Benchmark results on 94 original items (ten dimensions, every verdict produced by code, no language-model judge) show no accuracy gap outside noise at any of three effort settings. Two cells warrant re-measurement before any swap: the multi-step dimension under a five-of-five pass criterion (66.7 versus 75.0, n = 12, intervals overlap) and the scorer composite under the same criterion (86.7 versus 93.0). The resident remains FP8; the probe image, launcher, and weights are retained.

During the probe an autonomous remediation agent, the fleet's self-healing loop, intervened three times, crash-looping the service and ultimately stopping the probe container. We describe the hold recipe now used before deliberate seat downtime and a second trap involving PATH resolution in chained runners.

1. The Resident Seat and the Question

The fleet's resident seat, the model every agent calls, is Qwen3.8-Flash-Next in FP8. It is served with vLLM at tensor-parallel 2 on two workstation Blackwell GPUs (96 GB each), configured for

262,144-token context, 256 sequences, and 90 percent GPU memory utilisation. On 3 September 2026 the GPU vendor published an NVFP4 quantisation of the same model: 4-bit routed experts with mixed precision elsewhere. The question was whether this checkpoint would fit the resident shape and, if so, what it would buy.

Both checkpoints share one unusual component: a fixed 47.68 GiB per-layer embedding table (PLE), a static lookup structure the model consults per token. The FP8 resident already keeps this table in host memory rather than GPU memory, via a private patch in a locally built serving image. The published checkpoint's table is byte-identical to the FP8 one; the publisher's own card confirms this.

2. The Rejection on 4 September

We evaluated the checkpoint on the engine's stock nightly image. The checkpoint's README named an upstream commit, merged five minutes before the checkpoint was published, as required. The nightly was the only image carrying that commit.

The stock nightly has no host-offload path for the table. Its allocation was 51,201,966,080 bytes on every attempt, byte-identical: a fixed structure, not a batch buffer. The arithmetic is unforgiving. At 90 percent utilisation each card has roughly 85.5 GiB usable. Weights occupy 62.56 GiB; the table occupies 47.68 GiB. The sum is 110.24 GiB. About 25 GiB per card must leave the GPU before anything fits, and the only stock lever offloads weights, which is fatal to throughput.

Metric (same harness, 262,144 context)	FP8 resident	NVFP4 on stock nightly
Single-stream tok/s	85.99	8.69
Aggregate at 8 concurrent	328.11	22.76
GPU KV cache, tokens	1,592,310	537,617
Concurrency at full context	6.07×	2.05×

With 56 GiB of weights offloaded to host memory, only 6.07 GiB stayed resident per card and the seat decoded at 8.69 tokens per second. A faster configuration exists only at 32k context and 16 sequences; it was not measured. An autonomous remediation loop reverted the drop-in mid-experiment. We logged the gap as unmeasured rather than estimated.

Our internal frontier benchmark (94 original items, ten dimensions, every verdict produced by code, no language-model judge) was deliberately not run on the 8.69-token configuration. The live co-pilot budget is 1.5 seconds; the number is disqualifying on its own.

The day did establish one useful fact. The NVFP4 mixture-of-experts kernels work on this workstation GPU class under the stock engine. The engine selected a CUTLASS-based NVFP4 backend and the weights loaded clean at 62.56 GiB per card. No vendor-qualified image is needed for the arithmetic, correcting an expectation we had held since an earlier cross-node experiment.

Two capability limits also surfaced. The FP8 KV-cache option is unavailable for this architecture; the engine requires a BF16 main KV cache for its attention variant, and this constraint applies to the FP8 incumbent too. Both checkpoints have a native 262,144 position limit with no rope-scaling block, so the 4-bit checkpoint could never buy a longer window, only KV headroom.

The verdict recorded that day: rejected as unservable at the resident shape, with the blocker named as "an image carrying both the upstream commit and table offload".

3. The Diagnosis Was Wrong

Five days later we re-read the upstream branch with fresh eyes. The engine's day-zero image for this model, published the day the FP8 checkpoint appeared, already carried the host-offload environment variable for the table and the mixed-precision quantisation config. Its table-embedding quantisation selector only recognised the plain FP8 config class, so a mixed-precision checkpoint never received the FP8 embedding method. The upstream commit named in the README is exactly that branch: "if the config is mixed-precision and this layer's algorithm resolves to FP8, use the FP8 embedding method".

The 4 September evaluation had gone straight to the nightly (which lacks offload) because the README said the commit was required. We never tried the image that had the offload.

The fix was a twelve-line backport of that branch into the day-zero image: one file copied over a pinned image digest. Build and boot took about three minutes. The offload matched 132 tensors. The NVFP4 experts ran on the CUTLASS backend. Upstream pull requests for table offload remain open and unrebased, so the day-zero image is the only offload-capable base and is pinned.

4. The Probe on 9 September

We measured the NVFP4 checkpoint in the same placement as the FP8 resident: tensor-parallel 2, 262,144-token context, 256 sequences, 90 percent GPU memory utilisation, effort set to low.

Metric	FP8 resident (28 Aug and 2 Sep)	NVFP4 probe (9 Sep)
Weights per card	62.6 GiB	38.0 GiB
GPU KV cache, tokens	1,592,310 (6.07× at full context)	3,578,226 (13.65×)
Single-stream tok/s	124.7	110.1
Aggregate at 64 concurrent / p95 latency	3,368 tok/s / 4.84 s	3,881 tok/s / 4.19 s

The KV cache is 2.25 times larger. Aggregate throughput at 64 concurrent requests is 15 percent higher. The p95 latency is 13 percent lower. Single-stream throughput is 12 percent lower.

5. Benchmark Results

Our internal frontier benchmark comprises 94 original items across ten dimensions. Every verdict is produced by code; there is no language-model judge. We report k = 1 (single attempt per item) and pass^5 (five independent runs must all pass an item). Effort is the model's reasoning-effort setting.

Benchmark measure	FP8 resident	NVFP4 probe
k = 1, effort low (agentic / flat)	96.2 / 96.8	97.8 / 97.7
k = 1, effort medium	91.7 / 92.9	89.1 / 92.3
k = 1, effort xhigh	81.2 / 84.8	83.2 / 85.8
pass^5, effort low (agentic / scorer / flat)	87.3 / 93.0 / 90.0	88.8 / 86.7 / 88.0
pass^5 low, multi-step dimension (n = 12)	75.0	66.7

Benchmark measure	FP8 resident	NVFP4 probe
Completion tokens, low / medium / xhigh	41.7k / 47.7k / 104k	40.3k / 51.0k / 135k

The benchmark ties at every effort rung; every delta is inside the Wilson 95 percent intervals. Both checkpoints show the same non-monotonic ladder (low best, medium worse, xhigh worst) that we reported for the FP8 seat earlier.

Two cells are the ones to re-measure before any swap. The multi-step dimension under pass^5 is 66.7 versus 75.0, with n = 12 and intervals that overlap. The scorer composite under pass^5 is 86.7 versus 93.0. The pass^5 run at effort medium was cut at run 4 of 5; two further legs never ran.

Canary checks passed: exact arithmetic, coherent prose, well-formed tool calls, clean code.

At effort xhigh the 4-bit checkpoint emitted 30 percent more completion tokens than FP8. This is irrelevant at the effort the seat runs.

The seat decision: the resident stays FP8 until the two open cells are re-measured. The NVFP4 probe image, launcher, and weights are retained.

6. Remediation Loop Interference

During the probe an autonomous remediation agent, the fleet's self-healing loop, intervened three times. First it repointed two consumers from the resident seat to a different model on another node; this produced a wave of 400 errors from a sampling-parameter mismatch until reverted. Then it restarted the resident service against the probe, crash-looping on out-of-memory. Finally it stopped the probe container and restarted the resident.

This is the same class of event as the 4 September incident, when the remediation loop reverted the drop-in mid-experiment.

The hold recipe now used before any deliberate seat downtime is as follows. Stop the two conformance timers on the node. Stop the fleet remediation cycle timer. Place an alert silence on the seat's endpoint; the remediation gatherer honours silences. Restore all three afterwards.

A second trap surfaced the same day. A wave runner resolved a tool by name on a non-login PATH that did not include the user's local binaries. Every unit returned exit code 127 in 0.0 seconds and the wave reported "complete". The fix: export PATH explicitly in chained runners.

7. What These Numbers Will Not Carry

The sample sizes are small. We ran one probe per configuration. The pass^5 criterion operates on 94 items with 6 to 12 items per dimension. The FP8 comparison figures were taken one to two weeks earlier on the same placement; they are not simultaneous measurements.

The two open cells (multi-step pass^5 and scorer pass^5) have overlapping confidence intervals, so neither a regression nor parity can be claimed. The pass^5 run at effort medium was incomplete; its results are partial.

The throughput and latency figures come from a single harness run at 64 concurrent requests. We did not sweep concurrency levels; the 15 percent aggregate gain and 13 percent latency reduction apply only at that operating point.

The benchmark is code-judged, which removes language-model judge variance but introduces its own brittleness: a formatting change in model output can flip a verdict even when the answer is correct. We have not measured inter-run variance on the code judges themselves.

All measurements were taken on a single two-GPU node. We have not tested the NVFP4 checkpoint at tensor-parallel 4 or on a different GPU class.

The claims that survive: the 4-bit checkpoint fits the resident shape with the twelve-line backport, achieves 2.25 times the KV cache, and shows no accuracy gap outside noise at effort low under k = 1. The claims that do not survive: any statement about multi-step or scorer dimensions under pass^5, any statement about effort medium, and any generalisation to other placements.

8. Next Steps

The immediate work is to re-measure the two open cells: multi-step pass^5 and scorer pass^5 at effort low. If both close inside noise, the seat swaps to NVFP4 for the KV headroom. If either shows a gap outside noise, the seat stays FP8 and we investigate whether the gap is quantisation-induced or sampling variance.

The hold recipe for deliberate downtime is now documented and will be enforced by a pre-flight check in the experiment launcher.

The pinned day-zero image is a maintenance liability. We will track the upstream pull requests for table offload; when one merges and appears in a release image, we will retire the backport.

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 4 September and 9 September 2026; raw artefacts are retained. This paper is PT-R-2026-020.