PURETENSOR
INTELLIGENCE DISTILLED

The Reliability Champion Is Not the Single-Shot Champion

pass^5 inverts a leaderboard

Paper: PT-R-2026-015 v1.0

Author: PureTensor Ltd

Date: 30 August 2026

Status: Published

Abstract

Single-shot evaluation, one rollout per item at temperature zero, remains the standard protocol for model leaderboards. We report two experiments that challenge the assumption that single-shot rankings transfer to production reliability. In the first experiment, conducted on 21 August 2026, we swept five models through our internal frontier benchmark (94 original items, ten dimensions, all verdicts produced by code with no language-model judge). At k=1 the single-shot champion was Nemotron 3 Ultra 550B with a flat composite of 92.7; the reliability champion under pass^5 (five independent rollouts at temperature 0.7, passing only if all five pass) was a local 27B open-weight model with a flat composite of 87.2. The 550B model shed 8.6 points when required to be right five times consecutively; the 27B model shed 1.6 points. The leaderboard inverted.

In the second experiment, conducted on 28 August 2026, we observed the same inversion within a single model's reasoning-effort ladder. Our on-premises seat, a sparse 125B-parameter model with roughly 6B parameters activated per token, accepts three reasoning-effort settings plus a switch that disables reasoning. At k=1 the low setting scored 96.8 flat and the shipped default (medium) scored 92.9, a gap of 3.9 points. Under pass^5 the gap widened to 8.7 points (90.0 versus 81.3), and the multi-step tool-loop dimension collapsed from 9 of 12 items surviving at low to 4 of 12 at medium. We changed the production default to low.

The pattern that emerges is consistent across both experiments: configurations that appear nearly equivalent at k=1 diverge sharply under repeated sampling, and the divergence concentrates in the dimensions that matter most for agentic deployment (tool selection, multi-step execution). We are not aware of a published single-shot-versus-pass^k rank inversion on a code-scored benchmark; if none exists, these results supply one.

Threats to validity are substantial. Most configurations were evaluated in a single run; per-dimension confidence intervals on six to twelve items are 30 to 50 points wide; a hosted-lane confound remains unresolved; and the two experiments used different bank versions that cannot be pooled. The composite run-to-run noise on 94 items is about two points, so composite-level inversions survive, but per-dimension claims do not transfer beyond the specific items tested.

1. Motivation and Definitions

A model that can produce a correct answer is not the same as a model that will produce a correct answer. The distinction matters in agentic systems where a single failure in a multi-step tool loop can cascade. Leaderboards optimise for the former; production deployments need the latter.

Our internal frontier benchmark comprises 94 original items spanning ten dimensions: D1 tool selection, D2 abstention, D3 tool arguments, D4 multi-step execution, D5 reasoning, D6 false-premise detection, D7 instruction compliance, D8 structured output, D9 grounding, and D10 calibration. Every verdict is produced by code; no language-model judge is involved. Three composite scores aggregate the dimensions: flat (unweighted), scorer-weighted, and agentic-weighted. The weightings are fixed in the harness; the agentic profile emphasises tool-loop behaviour.

We define two evaluation protocols. The first, k=1 at temperature 0, is the standard leaderboard setting: one deterministic rollout per item. The second, pass^5 at temperature 0.7, requires five independent rollouts per item; an item passes only if all five rollouts pass. Pass^5 does not measure whether a model can be right; it measures whether a model is right reliably.

Wilson 95 per cent confidence intervals on per-dimension scores drawn from six to twelve items are 30 to 50 points wide. Deltas smaller than that interval are not findings. Composite run-to-run noise on 94 items is approximately two points, so composite-level differences exceeding four points are likely real.

2. Experiment 1: A Sweep Triggered by a Stealth Model

On 20 August 2026 an anonymous model appeared on a hosted routing service. It was free to query, offered a one-million-token context, and was described as frontier-class. The provider label was withheld and the originating laboratory was unknown. We refer to it as "the stealth model." We submitted it to the benchmark, and the session expanded into a sweep of five models. The sweep used bank version 1.0 (pre-audit). All hosted models were accessed through the routing service; local models ran on our own hardware. Total API spend for the sweep was approximately $2.99.

Model	Agentic	Scorer	Flat	Note
Author self-baseline (not blind)	99.2	98.8	98.5	upper anchor only; author knew the items
Nemotron 3 Ultra 550B (hosted, paid lane), first run	90.7	93.5	91.5	16 items unscorable
Nemotron 3 Ultra 550B, re-run	90.9	95.4	92.7	record confirmed; 4 unscorable
GLM-5.3 (hosted)	87.1	94.8	90.2	first GLM on the bank
Qwen3.8-27B FP8 (local)	84.2	95.3	88.8	prior record
The stealth model	84.6	93.5	88.6	ties the local 27B
Qwen3.8 2.4T flagship (hosted)	85.8	91.0	88.1	same as its own 27B distillation

The bank saturates near 88 to 89 flat at k=1. The 2.4-trillion-parameter flagship scored 88.1, indistinguishable from its own 27B distillation at 88.8. Differences inside that band are noise. Only Nemotron 3 Ultra broke above the saturation ceiling.

We then ran pass^5 on three models: the local 27B, the stealth model, and Nemotron 3 Ultra.

Model	Agentic	Scorer	Flat	Change in flat	D1 tool selection	D4 multi-step
Qwen3.8-27B FP8 (local)	85.0	89.9	87.2	−1.6	held	held
The stealth model	74.2	87.6	80.2	−8.4	50.0	58.3
Nemotron 3 Ultra 550B (hosted)	79.4	89.6	84.1	−8.6	62.5	58.3

The single-shot champion (Nemotron Ultra, 92.7 flat) and the reliability champion (the local 27B, 87.2 flat and 85.0 agentic under pass^5) are different models. Under pass^5 the local seat leads every model measured.

The stealth model's "frontier" claim is not supported by the bank. It ties a 27B open-weight model at k=1 and sheds 8.4 points when required to be right five times consecutively. High run-to-run variance is exactly what frontier models do not exhibit. Our verdict: a strong open-weights-class model with a flashy single-shot ceiling.

The degradation pattern (−8.4 and −8.6 points) now looks like a class trait of large hosted mixture-of-experts models sampled at temperature 0.7, rather than a defect specific to the stealth model. What distinguished the stealth model is that it degraded from a tied baseline rather than a leading one. Two Nemotron Ultra items were unscorable in all five rollouts, indicating a deterministic provider-side issue on the paid lane.

We attempted a control to separate model variance from lane variance. The 27B was served through the hosted lane rather than locally. Its k=1 leg matched the local run (88.1 versus 88.8), so the hosted lane does not distort single-shot scores. Its pass^5 leg was cancelled after three unrelated failures: a wedged connection, a session restart, and a checker crash that produced a harness fix. The reliability-champion claim is therefore "lane-inclusive": we cannot yet separate "hosted models are less reliable" from "hosted lanes are less reliable."

One operational note: a hosted run hung mid-item in a chunked TLS read. The provider's keepalive drips reset the 300-second read timeout indefinitely, so the harness hung despite its timeout. A stack dump showed the blocked read, and killing the single TCP connection un-wedged the run at the cost of one unscorable item. Per-item wall-clock deadlines are the fix.

The full 94-item bank was, by this sweep, disclosed to the unknown stealth laboratory and to routing-service providers generally. Checkers and gold answers never left our machines. If contamination is ever suspected, a held-out refresh is the remedy.

3. Experiment 2: The Same Inversion Inside One Model's Effort Ladder

On 28 August 2026 we examined our on-premises resident seat after a model change. The new model, Qwen3.8-Flash-Next in FP8, is a sparse architecture: 125B total parameters with approximately 6B activated per token, plus approximately 51B of n-gram embedding parameters. It runs tensor-parallel across two workstation-class Blackwell GPUs with a 262,144-token context. The bank had been updated to version 1.1 (post-audit); results are not comparable to Experiment 1's version 1.0 numbers.

The chat template accepts three reasoning-effort settings (low, medium, xhigh) and a separate switch that disables reasoning entirely. The seat had been serving "medium" as its default. Settings labelled "high" and "max" are rejected with HTTP 400; a run that passed "high" returned 0 per cent coverage and the harness reported it "not comparable" rather than scoring it zero.

At the stock 6,144-token text budget the first ladder was incomparable because xhigh silently lost four items to truncation while low lost none. We raised the budget to 12,288 text tokens and 8,192 tool tokens, achieving zero truncation on none, low, and medium, then re-ran both sides.

Effort	Agentic	Scorer	Flat	Completion tokens
none (reasoning off)	77.2	78.8	79.2	17,248
low	96.2	98.1	96.8	41,742
medium (shipped default)	91.7	94.3	92.9	47,688
xhigh	81.2	89.1	84.8	104,389

At k=1 the low setting leads by 3.9 flat points over medium. The gap is real but modest. Under pass^5 the gap widens.

Effort	Agentic	Scorer	Flat	D4 multi-step
low	87.3	93.0	90.0	75.0
medium	74.8	85.8	81.3	33.3

The gap that is 4.5 agentic points at k=1 becomes 12.5 at pass^5. D4 multi-step is the tool-loop dimension (12 items). At medium, only 4 of 12 multi-step items survive five rollouts. At low, 9 of 12 do. A setting that looks nearly as good single-shot is unreliable exactly where an agentic tool loop lives.

The first low run returned 99.2 with nine of ten dimensions at exactly 100.0, the shape of a scoring artefact. We confirmed it was real before changing anything. The checker gate had passed (83 gold pass, 30 wrong fail). A repeat k=1 reproduced (97.8 / 98.5 / 97.7). Pass^5 held the lead. Transcript inspection showed low genuinely answering the items medium failed: choosing the correct storage-health tool, rejecting a non-existent node, flagging an unsatisfiable constraint, computing the correct figure 7521.5.

We changed the server-side default effort to low. Verification through the production alias with no client flags returned 96.2 / 98.1 / 96.8, identical to the ladder's low rung.

4. A Counterweight from a Separate Battery

A separate workspace-scored agentic battery (12 coding probes, three repetitions each, yielding 36 units per configuration, scored by inspecting the resulting workspace rather than the model's own report) did not reproduce low's quality lead. Low scored 34 of 36 in 25.3 minutes; medium scored 34 of 36 in 50.5 minutes. Identical score, half the wall clock. The quality claim therefore rests on the frontier benchmark only; the cost and latency claim rests on both.

The battery also supplied a cautionary result about small samples. A three-repetition read suggested low was worse on an impossible-specification probe (1 of 3 versus 2 of 3). We re-ran at n=64 per arm. Low scored 14 of 64 (21.9 per cent, 95 per cent confidence interval 13.5–33.4); medium scored 15 of 64 (23.4 per cent, 95 per cent confidence interval 14.7–35.1). Fisher two-tailed p = 1.00. The n=3 signal was noise. Even an intermediate n=16 read (12.5 per cent versus 37.5 per cent, p=0.22) pointed the wrong way.

5. The Shape of the Inversion

Both experiments share a structure. At k=1 the configurations under comparison appear nearly equivalent: the stealth model ties the local 27B; medium trails low by fewer than four points. Under pass^5 the gap widens, and the widening concentrates in the multi-step and tool-selection dimensions.

The transferable lesson is that the default is never a neutral baseline. A prior paper on this site (PT-R-2026-011, "Past the Peak") reported an effort-ladder inversion on the previous 27B seat where the middle setting won. On this seat the lowest setting wins. The reliability axis makes the wrong default much more expensive than single-shot scores suggest.

We are not aware of a published single-shot-versus-pass^k rank inversion on a code-scored benchmark. If none exists, these results supply one. The claim is narrow: we have demonstrated that such inversions occur, not that they are universal.

6. What These Numbers Will Not Carry

Sample sizes are small. Most configurations were evaluated in a single run. Per-dimension confidence intervals on six to twelve items are 30 to 50 points wide; per-dimension deltas smaller than that interval are not findings. The composite run-to-run noise on 94 items is approximately two points, so composite-level inversions exceeding four points are likely real, but per-dimension claims do not transfer beyond the specific items tested.

The hosted-lane confound is unresolved. The control run that would have separated model variance from lane variance was cancelled after three unrelated failures. The reliability-champion claim from Experiment 1 is therefore lane-inclusive: we cannot distinguish whether hosted models are less reliable or whether hosted lanes are less reliable.

Experiment 1 and Experiment 2 used different bank versions (1.0 and 1.1 respectively). The versions are not comparable and must not be pooled. Any cross-experiment comparison of absolute scores is invalid.

The workspace-scored agentic battery did not reproduce low's quality lead over medium. The quality claim rests on the frontier benchmark only. If the frontier benchmark is unrepresentative of production workloads, the quality claim does not generalise.

The pass^5 protocol is arbitrary. We chose k=5 because it is small enough to be affordable and large enough to expose variance. A different k might produce different rankings. We have not characterised the sensitivity of the inversion to k.

Disclosure effects are possible. The full 94-item bank was disclosed to the unknown stealth laboratory and to routing-service providers. Checkers and gold answers never left our machines, but if contamination is ever suspected, a held-out refresh is the remedy.

The claims that survive are these: (1) a rank inversion between single-shot and pass^5 occurred in both experiments; (2) the inversion concentrated in the multi-step and tool-selection dimensions; (3) changing the production default to low improved both single-shot and pass^5 scores on the frontier benchmark while halving wall-clock time on the workspace battery. Claims about generality, about the mechanism of the inversion, and about per-dimension effects beyond the specific items tested do not survive.

7. What We Changed and What Comes Next

We changed the production default for our on-premises seat from medium to low. The change was verified through the production alias.

The next experiment is the completion of the hosted-lane control. We will re-run the 27B through the hosted lane under pass^5 to determine whether the reliability gap is a property of the models or a property of the lanes. If the hosted lane reproduces the local 27B's reliability,

the gap is model-specific. If it does not, the gap is lane-specific, and the operational implication is that local inference is preferable for reliability-sensitive workloads regardless of model quality.

The workspace-scored battery's failure to reproduce low's quality lead may reflect a ceiling effect (34 of 36 leaves little room for differentiation) or a genuine difference between the benchmark and production workloads. A held-out refresh of the bank, which the disclosure in Experiment 1 already argues for, is the other open item.

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 for this paper were taken on 21 August and 28 August 2026; raw artefacts are retained. This paper is PT-R-2026-015.