PURETENSOR
INTELLIGENCE DISTILLED

Past the Peak, Replicated

The Top Rung of the Effort Ladder Loses Again on Two More Builds, and the Peak Below It Is Softer Than It

Paper: PT-R-2026-051 v1.0

Author: PureTensor Ltd

Date: 12 September 2026

Status: Published

Abstract

On 19 August we reported that the top reasoning-effort setting on one dense model scored lower than the middle setting (flat 84.6 against 93.4) while spending 81 percent more completion tokens. This paper asks whether that was one seat's quirk; on the evidence we hold, it was not. On a 4-bit build of Qwen3.8-Flash-Next the flat composite on our 94-item benchmark falls from 97.7 at low effort to 85.8 at the top rung (xhigh) while completion tokens rise 3.36 times. On a second model family, Agnes-3.0-Flash, it falls from 94.3 to 84.4, but that run used smaller generation budgets than the Qwen ladders and 14 of its 94 xhigh items were truncated, so the direction is established and the size of the drop is not. The FP8 build, published on 30 August, shows 96.8 against 84.8. Paired-item sign tests put the top-rung loss at p of 0.0018 or lower on the two Qwen builds, and below 0.0001 on Agnes with the budget confound attached. The damage falls mostly on tool selection, abstention, multi-step execution and grounding.

Two things we expected are softer than we thought. DeepSeek-V4-Flash does not invert between its two top rungs; high and max tie (84.9 against 84.3 hosted, paired $p = 0.55$). And low beating medium holds on the 4-bit build (7 items against 0, $p = 0.016$) but on the cleanest FP8 pair is 5 against 1 ($p = 0.22$), and on Agnes is absent. The peak sits at low or medium.

The larger part of the paper is about our own tables. Items that ran out of generation budget were excluded from scores rather than counted as failures; on Agnes at xhigh that was 14 of 94 items, and the printed 84.4 becomes 71.2 when they count. The harness reported "100 percent coverage" for that run, reported zero reasoning tokens for a model that spent 83 percent of its output at low effort on reasoning, and in our serving configuration one "effort none" arm on a GLM build had no effect at all. We corrected the harness on 12 September.

1. What We Asked After the First Result

The August result came from one model on one chat template. The dial might be worse at maximum because of that template, that seat's serving stack (a model deployed to serve one role), or the benchmark. So: does the top of the dial lose on other models and builds? Where is the peak, and how sure can one run of the 94-item bank (our fixed set of test items) be about it? And which of our own numbers are fragile because of truncation, bank version or harness accounting?

We have already published the FP8 effort ladder, its reliability comparison and a workspace-scored battery that tied low and medium; the FP8 ladder appears here as a reference row. New here are a 4-bit Qwen build, Agnes-3.0-Flash, a DeepSeek plateau, paired-item evidence, a re-scoring that counts truncation as failure, and the harness defects we found along the way.

The short answers: the top rung loses to low on every build we laddered through low, medium and xhigh (the FP8 and 4-bit builds of one Qwen model, and Agnes-3.0-Flash, whose size of loss is confounded by budget). DeepSeek-V4-Flash, laddered at none, high and max, plateaus instead. The peak is "low or medium", with low ahead by 0.7 to 5.4 points in single runs but significantly on only some builds.

2. The Instrument and the Runs

The instrument is our internal frontier benchmark: 94 original items across 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, D10 calibration). Every verdict is produced by code, with no language-model judge. The headline composite is flat, the unweighted mean of the ten dimension scores. An item that finishes on the length limit is marked unscorable rather than failed, and a dimension score is computed over its scored items only. Section 6 shows why that matters.

Runs are single passes at temperature 0 (k=1) unless stated; reliability runs (pass^5) use five rollouts at temperature 0.7 and count an item only if all five pass. The Qwen and Agnes runs used bank version 1.1, after our August audit; the DeepSeek arms ran on 31 July against version 1.0 and cannot be ranked against them.

Qwen3.8-Flash-Next FP8 ran under vLLM on two RTX PRO 6000 Blackwell GPUs; "none" is sent as thinking disabled. The 4-bit (NVFP4) build used the same placement on a day-zero engine build and is a probe build. Agnes-3.0-Flash ran in BF16 under a nightly SGLang build, one k=1 run per rung at older budgets, so our findings concern our run of it and not the model in general. DeepSeek-V4-Flash ran hosted and as a local GGUF build under llama.cpp, at none, high and max. One GLM-5.3-Flash arm appears only in Section 6.

For Qwen the "fair ladder" gave every rung 12,288 tokens for text and 8,192 for tool calls, after an earlier ladder at 6,144 text lost four items to truncation at the top rung only. Agnes ran at the older defaults of 6,144 and 4,096, a confound we address in Section 6.1.

Noise reference. Five k=1 runs of the FP8 low configuration on 28 August (three explicit-effort runs and two server-default runs after the switch to low, with budgets differing) gave flat scores of 98.5, 97.7, 96.8, 96.8 and 95.6, a range of 2.9 points. Two medium runs gave 90.8 and 92.9 and two xhigh runs 87.3 and 84.8. A single composite therefore carries 2 to 3 points of spread, and differences under 4 points need paired evidence.

3. The Top Rung Loses Again

Build	none	low	medium	xhigh
Qwen3.8-Flash-Next FP8 (published)	79.2	96.8	92.9	84.8
Qwen3.8-Flash-Next 4-bit	not run	97.7	92.3	85.8
Agnes-3.0-Flash	not run	94.3	93.6	84.4

Table 1. Flat composite by effort rung, k=1, temperature 0, bank version 1.1. The Agnes medium figure is disputed (Section 4.1), and the Agnes xhigh figure is a score on the 80 items

that finished at smaller budgets (Section 6.1).

The 4-bit build repeats the FP8 pattern, slightly sharper, and cost rises with rung on every build much faster than the score falls. Completion tokens per full run at low and xhigh were 41,742 and 104,389 on FP8 (2.50 times), 40,334 and 135,444 on the 4-bit build (3.36 times) and 57,894 and 161,314 on Agnes (2.79 times). On the 4-bit build reasoning is 74 percent of output at low and 94 percent at xhigh: 127,288 reasoning tokens at xhigh against 29,811 at low, for a score 11.9 flat points lower.

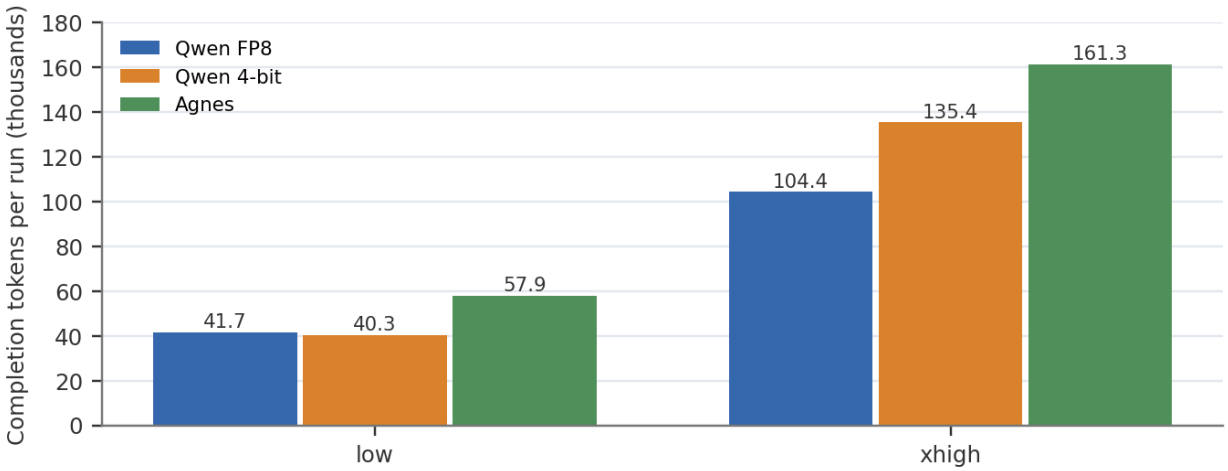


Figure 1. Completion tokens per run at low and xhigh, in thousands, for the three builds. Exact values are in the text above.

3.1 Where the damage falls

Cells hold 6 to 12 items, so one item moves a cell by 8 to 17 points; only the repeated pattern is evidence.

Dimension	FP8	4-bit	Agnes
D1 tool selection	25.0	12.5	12.5
D2 abstention	25.0	33.3	36.4
D4 multi-step	25.0	25.0	8.3
D5 reasoning	0.0	0.0	0.0
D6 false premise	0.0	11.1	20.0
D7 instruction	8.3	8.3	0.0
D8 structured output	0.0	0.0	0.0
D9 grounding	37.5	28.6	16.7
D10 calibration	0.0	0.0	5.0

Table 2. Score lost between low and xhigh, in points, scored items only. D3 tool arguments lost nothing on any build. Agnes D5 and D6 at xhigh are scored on 4 of 8 and 5 of 10 items, because the rest were truncated.

The top rung loses mostly on tool selection, abstention, multi-step execution and grounding, while structured output sits at ceiling at both rungs. Smaller losses appear on false-premise detection and instruction compliance on two builds each. The Agnes reasoning cell is not evidence of ceiling, since half its xhigh items were dropped. On the 4-bit build, abstention goes from 100.0 to 66.7, grounding from 100.0 to 71.4 and multi-step from 91.7 to 66.7. This fits the August reading: the dimensions that suffer are those where the right behaviour is to commit to

a plain action, to decline, or to stay with the provided evidence, though we have clean support for it on the Qwen builds only.

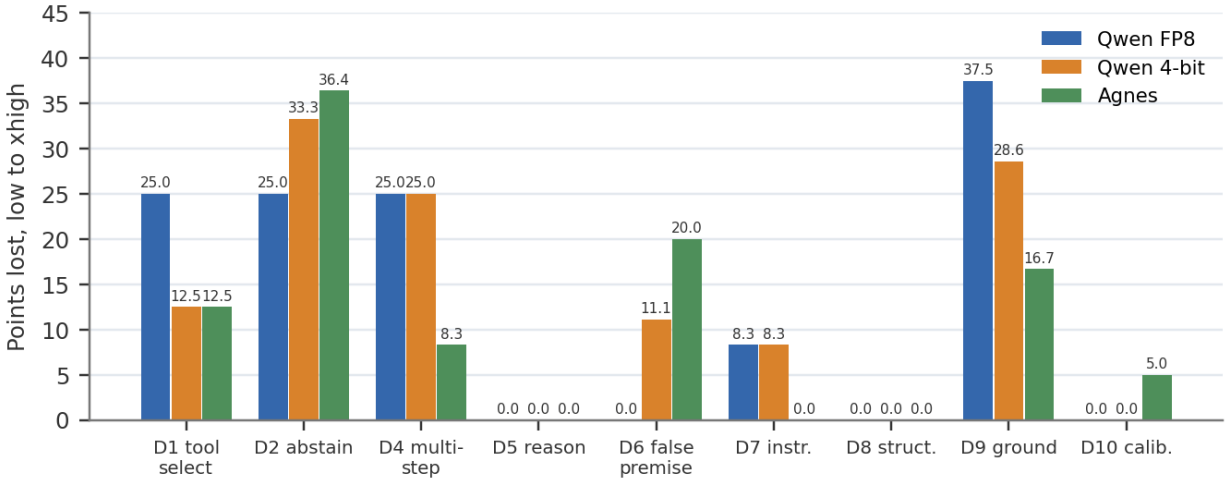


Figure 2. Points lost between low and xhigh by dimension and build. Values are in Table 2.

3.2 Paired-item evidence

The same 94 items run at every rung, so we compare runs item by item: for each pair we count items passed in one run and not the other (an unscorable item counts as not passed) and apply an exact two-sided sign test to those discordant items. The p-values are unadjusted for the number of comparisons.

Build, comparison	Passed only at first	Passed only at second	Exact p
FP8, low vs xhigh	13	1	0.0018
FP8, medium vs xhigh	10	2	0.039
FP8, low vs medium (fair ladder)	5	1	0.22
FP8, 6,144 budgets, low vs medium	9	1	0.022
FP8, 6,144 budgets, repeat low vs medium	8	1	0.039
4-bit, low vs xhigh	15	0	0.0001
4-bit, low vs medium	7	0	0.016
Agnes, low vs xhigh	25	0	below 0.0001
Agnes, low vs medium	5	3	0.73

Table 3. Exact paired sign tests on discordant items, 94 items per run. The Agnes low-versus-medium row uses the saved medium summary, the second execution described in Section 4.1.

The top-rung loss is solid on the two Qwen builds. On Agnes the direction is clear but its size is confounded by the smaller budgets (Section 6.1). On the FP8 build, xhigh is not distinguishable from no reasoning at all (8 items against 14, p = 0.29, flat 84.8 against 79.2).

4. The Peak Below the Top Is Soft

Our August paper said medium, then low. The data support "low or medium" and little more.

On the 4-bit build low beats medium by 7 items to 0 (p = 0.016). On FP8, all three low-versus-medium pairs favour low. Two are significant on their own at the smaller budgets, but both use the same medium run set against two different low runs, so they are one medium run and not two replications. At the fair budgets it is 5 against 1 and p = 0.22; taken alone, that pair is a tie. On Agnes, low and medium are indistinguishable.

4.1 One sweep, two executions

The saved Agnes medium summary reads flat 93.6; the console log of the same sweep printed 92.6. The raw transcripts for medium and xhigh each hold two responses per item (188 and 185 lines for 94 items), where low holds 94.

The two halves of the medium transcript sum to 72,451 completion tokens and 70,771. The saved medium total of 70,844 equals the second half plus the 73-token tool-parser preflight; the saved xhigh total (161,314) matches the first half (161,239) plus preflight. We infer that two executions of the sweep ran against one server and wrote to the same files; the retained logs do not establish why. We quote Agnes medium as 92.6 or 93.6, within noise of low, and claim no medium-versus-low difference on that model.

4.2 The reliability gap

The FP8 pass^5 headline (low 90.0 against medium 81.3 flat) was published on 30 August. The per-dimension file adds that the gap is not confined to multi-step (75.0 against 33.3): abstention and false-premise detection lose 17 to 20 points (83.3 to 66.7; 90.0 to 70.0), tool selection 12.5 (87.5 to 75.0) and instruction compliance 8.4, while tool arguments move the other way (87.5 to 100.0). This is one pair on one build, in cells of 10 to 12 items, with no significance test.

On the 4-bit build, pass^5 at low completed at 88.0 flat. Pass^5 at medium was attempted twice on 9 September and never completed (10 percent and 60 percent of scoring weight measured). We have not established why they stopped. The comparison does not exist for the 4-bit build, and we do not infer it from the FP8 one.

5. A Plateau on DeepSeek, Not an Inversion

DeepSeek-V4-Flash is where we expected the top rung to lose, and it does not.

Lane	none	high	max	Completion tokens, none / high / max
Hosted, flat	74.0	84.9	84.3	12,639 / 87,125 / 122,023
Local GGUF, flat	69.6	85.0	84.1	14,090 / 73,691 / 120,720

Table 4. DeepSeek-V4-Flash, bank version 1.0, 31 July. Not comparable with the version 1.1 rows elsewhere.

High minus max is 0.6 flat points hosted and 0.9 local, inside the noise. The paired sign tests give 7 against 4 hosted (p = 0.55) and 7 against 1 local (p = 0.070). This is a plateau with a cost penalty: max spends 40 percent more tokens hosted and 64 percent more local. Turning reasoning on is a large gain on the local lane (none against high, 3 items to 16, p = 0.0044, unadjusted) and a non-significant one hosted (4 to 11, p = 0.12).

Truncation points the other way, and we hold it loosely. Items lost to the length limit were 4 at high and 7 at max hosted, and 3 and 7 locally; counting them as failures puts max below high by 3.5 points hosted (floor 77.2 against 80.7) and 4.8 local (76.9 against 81.7), but the budgets were not recorded, so that gap may be a budget effect. We say only that high and max tie on the printed score and that max costs more.

6. What Was Wrong With Our Own Tables

6.1 Truncation was excluded, not penalised

A truncated item is dropped from its dimension, so a rung that reasons until it runs out of budget may shed items it would have failed. We re-scored every Qwen and Agnes run counting unscorable items as failures (the "floor" composite: the mean of dimension scores multiplied by item coverage).

Run	Printed flat	Unscorable items	Floor flat
FP8 xhigh (12,288 budget)	84.8	1	83.8
4-bit xhigh	85.8	2	84.0
Agnes low	94.3	1	93.3
Agnes medium (saved)	93.6	2	91.9
Agnes xhigh	84.4	14	71.2

Table 5. Printed and floor flat composites. The floor counts every unscorable item as a failure.

For the Qwen builds the correction is small, 1 to 2 points, and the ordering does not change. For Agnes at xhigh it is large: 14 of 94 items hit the length limit (D5 4 of 8, D6 5 of 10, D7 2, D9 2, D2 1), the printed 84.4 is a score on the 80 items that finished, and the whole-bank pass rate is 68.1 percent (64 of 94). In the raw first-execution transcript, 18 of 94 xhigh responses finished on length, against 3 at medium and 1 at low.

Agnes's top-rung collapse is partly a budget artefact. The August lesson (raise the budget and re-run every rung) was not applied to Agnes, and a budget-fair Agnes ladder has not been run. Agnes xhigh scores below low on the 80 items that finished (84.4 against 94.3), a comparison that is itself conditional on those items being short; whether the gap survives a budget-fair run is untested.

6.2 Harness defects found on 12 September

Coverage counted dimensions, not items. The comparability flag treated a dimension with any scored item as fully measured, so Agnes at xhigh was printed as "100 percent coverage, comparable". Earlier runs carry the same label. Scoring now weights each dimension by the share of its items scored, fixed the same day.

Reasoning tokens were read from one dialect only. The harness read the nested usage field used by OpenAI-style APIs; the SGLang build we ran reports the count at the top level. Every Agnes summary therefore shows zero reasoning tokens, while the per-response usage in the raw transcript shows reasoning in every response: at low, 48,139 of 57,819 completion tokens, or 83 percent. Our cost table had called the heaviest reasoner in the comparison a zero-reasoning model; this is also fixed.

One lane really reports nothing. The llama.cpp build we ran for the local DeepSeek lane put only prompt and completion tokens in its usage block, while the transcript shows reasoning text in all 94 responses at high and at max. On that lane, compare cost by completion tokens.

6.3 An effort setting that did nothing

On 27 August we started an effort-none arm on what we attribute to a 4-bit local build of GLM-5.3-Flash, and terminated it after 61 of 94 items. The transcript does not record the build; we attribute it from the matching totals below and the battery log. All 61 retained responses carry a reasoning channel. On the same 61 items the default-effort run produced 49,872 completion tokens and 160,766 characters of reasoning; the "none" arm produced 49,853

tokens and 160,702 characters. A 0.04 percent difference at temperature 0 means the switch had no effect in our serving configuration. We did not establish whether the template, server or build was responsible, and did not test other servers.

Acceptance of a request is not evidence that it was honoured. Agnes's preflight showed its server returning HTTP 200 for invented request fields, so for Agnes the effort effect is established by behaviour, rising completion tokens and reasoning share, and not by acceptance. Any ladder that includes "none" must assert on the response that the reasoning channel is empty.

7. What These Numbers Will Not Carry

Every run is 94 items and one k=1 pass per rung per build, except the FP8 build (run more than once) and the reliability runs. Per-dimension cells are 6 to 12 items. We tested two model families, one at two quantisations, plus DeepSeek-V4-Flash. No multiple-comparison correction was applied to Table 3, and single composites carry 2 to 3 points of spread.

What survives: the top rung loses to low by about 12 flat points on the FP8 and 4-bit builds, with p of 0.0018 or lower and 2.5 to 3.4 times the tokens; the same direction on Agnes among the items that finished, with a budget confound; damage mostly on tool selection, abstention, multi-step execution and grounding; a DeepSeek-V4-Flash plateau on an older bank version; and an effort setting that can be silently ignored.

What does not survive, or must be hedged: "low beats medium" as a general law, since it is significant on one build, weaker on another (the two significant FP8 pairs share one medium run) and absent on a third. "The gap sits in multi-step", which is the largest gap and not the only one. Any ranking of DeepSeek against the version 1.1 runs. The magnitude of the Agnes xhigh loss. And any claim that the inversion occurs at medium: it occurs at the top rung. The 9 September workspace-battery waves ran at low only, so no second-battery effort contrast exists for the 4-bit build or Agnes. One correction to how the 30 August figures should be read: the published "25.3 versus 50.5 minutes" are sums of per-probe seconds; clock time was 6 minutes 44 seconds and 12 minutes 54 seconds, a ratio of 1.9, the same conclusion.

8. What We Changed

The resident seat's default moved from medium to low on 28 August, and nothing here reverses that. Effort ladders now run at one raised budget for every rung, and any rung with unscorable items is reported with its floor composite beside the printed one. An effort arm must now show its behaviour (an empty reasoning channel for "none", reasoning tokens rising up the ladder) before its score is read. A budget-fair Agnes ladder and a pass^5 at medium on the 4-bit build have not been run.

9. Methods and Artefacts

Instrument: our internal frontier benchmark, bank version 1.1, 94 items, run through each server's OpenAI-compatible chat endpoint with tools passed natively. The sign tests and the floor composite were computed by us from the saved per-item results with a small read-only script and no new inference. Servers: vLLM (both Qwen builds), a nightly SGLang build of 8 September (Agnes), llama.cpp (local DeepSeek GGUF) and the hosted API (DeepSeek hosted lane). Measurements were taken on 31 July, 27 and 28 August, and 9 and 12 September 2026. Per-run summaries, raw transcripts with reasoning text and usage, and logs are retained, as is the partial effort-none transcript.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-051. The measurements were taken between 31 July and 12 September 2026; per-item results, raw transcripts and logs are retained.