PURETENSOR
INTELLIGENCE DISTILLED

# Past the Peak

Maximum reasoning effort is a pessimisation

***Paper:*** *PT-R-2026-011 v1.0*
***Author:*** *PureTensor Ltd*
***Date:*** *19 August 2026*
***Status:*** *Published*

## Abstract

Reasoning-capable models increasingly expose an effort control: a dial that asks the model to think harder or less hard. The intuition behind such a dial is that quality is monotonic in effort, and that the top setting is the one to use when correctness matters. We report a case where that intuition is inverted.

The open-weight model serving as our on-premises text seat exposes exactly three effort settings through its chat template, and defaults to the highest. Nothing in our stack had ever set the value, so every request the seat had served since deployment ran at maximum. Measured across 94 benchmark items at each setting, the default scored **84.6** on our flat composite against **93.4** at the middle setting, while spending **81 percent more completion tokens**. Disabling reasoning entirely was worse than both, at 78.5, so the reasoning itself was doing work. Its length was the defect.

The per-dimension breakdown locates the damage precisely. Three dimensions were already saturated at maximum effort and had no room to improve. Every dimension that moved measures whether the model *commits*: selecting a tool, declining to answer, carrying a chain of steps to its end, staying anchored to a source. Abstention rose from 66.7 to 100.0, meaning that at maximum effort the model reasoned its way into answering questions it should have refused.

A second experiment closes the argument. The three effort settings are not compute budgets; each injects a different sentence of guidance into the prompt. We therefore reproduced the highest setting's instruction as a hand-written system prompt of the kind that was standard practice in early prompt engineering, and ran it against the corrected seat. Tool selection fell from 87.5 to **62.5**, which is the exact figure that dimension scored at maximum effort. The same instruction content produces the same degradation whether it arrives through the serving stack or through a prompt somebody typed.

We give the method, the results, the failure mode at the top of the dial, and an explicit account of what these numbers will not carry.

## 1. A Setting Nobody Chose

The seat is a 27-billion-parameter dense model with a 262,144-token context window, served through vLLM in FP8, tensor-parallel across two workstation-class Blackwell GPUs. Reasoning is

on by default and preserved across conversational turns.

Its chat template accepts a `reasoning_effort` parameter with exactly three legal values. Anything outside that set raises a template exception rather than falling back, which is worth knowing before a benchmark sweep tries the effort vocabulary of a different vendor. When the parameter is absent, the template resolves it to the highest setting.

No consumer in our stack set it. Not the agent harnesses, not the retrieval pipeline, not the dictation correction path. The value had never been chosen, discussed, or measured; it was simply what the template did when asked nothing, and it had been the ceiling on every recorded benchmark for this model since the day it took the seat.

That is the first observation worth carrying away, and it is not about this model. **A default is not a neutral baseline.** It is a choice made by whoever wrote the template, on their task distribution rather than yours, and it silently becomes the configuration under which you characterise the model's abilities.

## 2. The Dial Is a Sentence, Not a Budget

Before measuring, it is worth being precise about what the control does, because the obvious reading of our result would be wrong.

Reading the chat template shows that each setting selects a different string of guidance to insert into the prompt. The highest setting adds an instruction to think carefully through the task, validate key assumptions, consider plausible alternatives, and prioritise correctness. The lowest adds an instruction to keep thinking brief and move directly to a conclusion. The middle setting adds **nothing at all**.

No sampling parameter changes. No token budget changes. No additional computation is purchased. The three settings differ only in what the model is told.

This matters for the scope of the finding. Our results are not evidence that additional inference-time computation has negative returns. They are evidence that a particular instruction, asking a model to keep its option space open, makes it worse at closing that option space. Those are different claims, and only the second one is supported by what follows.

It also means the best available setting on this model was **the absence of guidance**. The middle setting won by inserting no sentence, which is to say the model's own trained disposition outperformed both attempts to steer it.

## 3. Method

Quality was measured with our internal frontier benchmark: 94 original items across ten dimensions, with every verdict produced by code rather than by a language-model judge, and guards that mark truncated responses unscorable rather than counting them as failures. The dimensions cover reasoning, tool selection, argument construction, multi-step execution, grounding in supplied sources, false-premise detection, instruction compliance, structured output, abstention, and calibration. Three composite seats are reported: a flat average, a scorer-weighted composite, and an agentic composite.

Four configurations were measured on the same hardware placement, one per effort setting plus a fourth with reasoning disabled entirely through the template's separate boolean. Each run issues all 94 items and reports 100 percent coverage. Reported token figures are completion tokens summed across the run, which for a reasoning model includes the reasoning channel.

For the prompt experiment in section 7, three further runs were issued concurrently against the corrected seat, so that placement, batching, and engine state were shared rather than

confounding. A preamble hook was added to the benchmark harness for this purpose. It prepends text to each item's own system message rather than replacing it, so task context survives and the run still measures the same benchmark; with the hook unset, requests are byte-identical to a run without the feature.

## 4. Results: An Inverted U

| Effort setting | Flat | Scorer | Agentic | Completion tokens |
|---|---|---|---|---|
| Reasoning disabled | 78.5 | 79.6 | 78.1 | 20,816 |
| Low | 92.5 | 95.5 | 90.7 | 42,884 |
| Medium | 93.4 | 95.7 | 92.0 | 52,350 |
| High (template default) | 84.6 | 91.0 | 79.0 | 94,937 |

Quality rises steeply from disabled to low, peaks at medium, and falls at the top setting. Cost rises monotonically throughout. The default sits at the worst point on the trade: lowest quality of the three settings that permit reasoning, and by a wide margin the most expensive.

The agentic composite separates the settings more sharply than the flat one, at 92.0 against 79.0, which is consistent with the per-dimension picture in the next section: the damage concentrates in behaviours that agentic work depends on.

Disabling reasoning is not the remedy. It costs roughly fifteen points against the middle setting and produces the worst quality we measured. The reasoning is doing real work; the top setting simply asks for too much of it.

## 5. Where the Damage Falls

| Dimension | High | Medium | Change |
|---|---|---|---|
| Tool selection | 62.5 | 87.5 | +25.0 |
| Abstention | 66.7 | 100.0 | +33.3 |
| Multi-step execution | 66.7 | 83.3 | +16.6 |
| Grounding | 85.7 | 100.0 | +14.3 |
| Reasoning | 100.0 | 100.0 | saturated |
| False-premise detection | 100.0 | 100.0 | saturated |
| Structured output | 100.0 | 100.0 | saturated |
| Instruction compliance | 91.7 | 90.9 | −0.8 |

Two things stand out.

First, the dimensions that measure raw reasoning ability were **already at ceiling under the maximum setting**. Additional deliberation had nowhere to help, because reasoning power was not the constraint. Any account of this result as the model getting dumber is the wrong shape; the model's reasoning was unchanged and perfect on the items designed to test it.

Second, every dimension that moved is a **decision** dimension rather than a thinking dimension. Choosing the right tool, declining to answer, carrying a multi-step chain through to its conclusion, and staying anchored to a supplied source are all measures of whether the model commits to something. More deliberation did not make it reason better. It made it commit worse.

The abstention result is the sharpest single reading in the study. At maximum effort the model scored 66.7 on items designed to be unanswerable from the information given; at the middle setting it scored 100.0. An instruction to consider plausible alternatives measurably eroded the model's calibration about the boundaries of its own knowledge. Asked to keep possibilities open, it kept open the possibility that it knew the answer.

## 6. The Failure Mode at the Top of the Dial

Composite scores understate what happens on open-ended work. At the maximum setting, prompts that invite design or enumeration frequently do not terminate their reasoning at all.

A representative case: a request to design a zero-downtime schema migration consumed the entire 16,384-token generation budget in the reasoning channel, produced **zero answer tokens**, and terminated on the length limit after 235 seconds. Raising the budget does not resolve it. At 32,768 tokens the same prompt ran 453 seconds and still produced no answer. The same prompt at the middle setting completes in 91 seconds with a full response.

The behaviour is not a sampling artefact. It reproduced at temperature 0, at the vendor's recommended sampling settings for this model, and at the engine's own defaults. Nor can it be bounded by configuration: neither a thinking-budget parameter nor a reasoning-token ceiling is honoured by this template and engine combination, so the effort setting is the only control that exists.

This failure mode deserves emphasis because of how it presents to an operator. It is not a wrong answer that review would catch. It is an empty response after several minutes, which reads as a hang, and which an agent harness will surface as a stalled turn rather than as a quality problem. The defect is invisible to any evaluation that scores answers, because there is no answer to score.

## 7. The Same Damage, Arriving Through a Prompt

If the mechanism is the instruction rather than the plumbing, then writing that instruction by hand should reproduce the effect. We tested this against the corrected seat, with the effort setting fixed at medium for all three arms.

The arms were a bare run with no preamble; a 95-word preamble in the style that was standard practice in early prompt engineering, asserting a world-class expert persona and directing the model to think very carefully step by step, consider all possible edge cases, validate its assumptions, and evaluate alternative approaches; and a four-word instruction, “Be accurate and concise.”

| Preamble | Flat | Scorer | Agentic | Completion tokens |
|---|---|---|---|---|
| None | 91.3 | 93.9 | 89.9 | 52,466 |
| Four words | 89.4 | 93.3 | 86.6 | 40,160 |
| 95-word expert persona | 88.5 | 94.2 | 83.8 | 76,931 |

The elaborate preamble is the weakest arm on the flat and agentic composites while spending 47 percent more completion tokens than saying nothing. It was also markedly slower in wall-clock terms, and it re-created the non-termination behaviour of section 6 despite the effort dial being set to medium.

The result that ties the two experiments together is tool selection. It scored 87.5 with no preamble and **62.5** with the expert persona. That is the same figure, to the decimal, that tool selection scored at maximum effort with no preamble at all. Two different delivery mechanisms,

one carrying comparable instruction content, produced identical degradation on the same dimension.

The four-word arm is the useful complication. It was also slightly worse than silence, while costing 23 percent less. That is a cost trade rather than an improvement, and it rules out the simplest reading of these results: brevity was not the active ingredient. A content-free instruction failed at four words and at ninety-five.

## 8. What These Numbers Will Not Carry

The bare configuration scored 91.3 here, 91.8 in a verification run, and 93.4 in the effort sweep: three runs of a comparable setup within one day, spanning 2.1 points. Single-sample benchmark runs on 94 items carry roughly two points of run-to-run variation, and per-dimension figures, drawn from six to twelve items each, are individually softer still.

Applied honestly, that noise floor disqualifies some of our own numbers. The elaborate preamble's 2.8-point deficit on the flat composite sits only marginally outside it and should not be leaned on. What survives are the effects several times larger than the noise: the 8.8-point flat gap between the default and the middle effort setting, the 13-point agentic gap, the 6.1-point agentic gap in the prompt experiment, the tool-selection collapses, and the token costs, which are not scores at all and carry no sampling noise.

Three limits of scope are worth stating plainly. This is one model, and the effect may not transfer to models whose effort controls are implemented as compute budgets rather than as inserted guidance. It is one text-only benchmark of our own construction. And all three preambles in section 7 were deliberately content-free, so the experiment measures decorative prompt scaffolding and says nothing whatever about whether supplying real context helps.

That last limit is the one most likely to be misread, so section 9 addresses it directly.

## 9. What We Changed, and What We Did Not

We set the middle effort as a server-side default in the serving configuration rather than patching each client. The default merges with request-supplied template arguments rather than replacing them, so every consumer inherits the corrected setting with no client change, and an explicit per-request override still works. Verified against the served default with no client flags, the seat returns 91.8 flat, 94.0 scorer, 90.5 agentic at full coverage.

We did not conclude that shorter prompts are better prompts, and we would caution against reading this work that way.

The distinction that survives our results is between prompt content that is **decorative** and prompt content that is **informational**. Decorative content restates a disposition the model already has: expert personas, exhortations to be thorough, instructions to think carefully, stacked emphasis. It carries no task information, and on the evidence above it costs both quality and money. Informational content supplies what the model cannot infer: the audience, the environment, the constraints and the reasons behind them, the contract a tool actually honours, what a finished result looks like, and which specific failure to avoid. Removing that is not a saving.

Our own worker briefs, which we did not change, illustrate the distinction. They are long. They contain a machine-checkable definition of done stated as a command and its expected output, absolute paths for deliverables, an explicit instruction that a correct refusal to implement a self-contradictory specification counts as success, and a scoping rule that names the incident which produced it. There is no persona in them and no exhortation. Length was never the problem; vagueness was.

The practical rule we now apply to prompts is a single question asked of every line: **could the model already know this?** What restates a trained disposition comes out. What only the author knows stays, and is sharpened. On a model that reasons by default, depth belongs in the serving configuration where every consumer inherits it, not in prose that each team rewrites.

## 10. A Note on Recorded Weaknesses

One consequence deserves separating out, because it generalises past this model.

Tool selection at 62.5 had been recorded in our own notes as a characteristic weakness of this model, with a plausible explanation attached: transcripts showed the model deliberating into reasonable but non-canonical tool choices. The explanation was accurate as a description of the transcripts. It was wrong as an attribution, because the deliberation it described was the product of an effort setting nobody had chosen. At the middle setting the same model scores 87.5 on the same items.

A capability measured under an unexamined default is a property of the configuration and not of the model. Any weakness recorded that way is worth re-testing before it hardens into a fact about what a model can do.

*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-2026-011. All measurements were taken on production hardware on 19 August 2026; the eight benchmark runs described above are retained in full, including raw per-item transcripts, and the corrected serving configuration was verified against the live endpoint after deployment.*