Reasoning Is the Default Modality — and It Breaks Naive Agent Measurement

A Fathom measurement study across five frontier model vendors

Embedded Risk Analytics · A Fathom Case Study · August 2026

Fathom exists to make the measurement of an AI agent trustworthy: to tell a real coherence signal from an artifact of how the agent was run. Anyone measuring an agent today makes two quiet assumptions — that the model answers when it is asked, and that a non-reasoning baseline is something you can obtain for a fair comparison. On current frontier models both assumptions are false, and the ways they fail are invisible on a conventional scoreboard. We probed eight models across five vendors to show where, and why it is Fathom's ground.

One result stands out. On four of the five vendors tested, the flagship reasoning model cannot be run without reasoning — the API rejects an explicit request to disable it — and where a model's hidden thinking shares the response budget, it spends that budget thinking and returns nothing, which an ordinary harness records as a wrong answer rather than a broken measurement.

What we measured

We put each model to a trivial task (add two small numbers) and a harder one (a multi-step modular expression), each demanding a single line of output: an answer and nothing else. For every model we varied reasoning three ways — default (send no reasoning parameter), explicitly off, and explicitly high — at a tight 32-token output budget and an ample 2,000-token one, five times per cell. We recorded how many reasoning tokens the model spent, whether a valid answer survived, whether the answer was correct, and whether the reasoning was visible in the response or hidden. Every model was reached through OpenRouter on the same day. The probe carries its own planted self-tests and is shipped alongside this study, so every number below is reproducible.

Reasoning is the default, and on four of five vendors it is mandatory

Four flagship reasoning endpoints — OpenAI's GPT-5, Google's Gemini 2.5 Pro, DeepSeek R1, and xAI's Grok 4.6 — reason by default and reject an explicit disable outright, returning "Reasoning is mandatory for this endpoint." On these models a non-reasoning control is not merely unusual; it cannot be requested. Whether a model reasons by default is not even a fixed property: on the trivial task Claude Sonnet 5 spent nothing, but on the harder task it engaged reasoning on its own, spending thirty-four tokens per call with no parameter set. Reasoning is spent when the work warrants it, and the heavy reasoners scale steeply — Gemini and DeepSeek R1 each cross six hundred tokens to evaluate one arithmetic expression. In every case the reasoning is hidden: billed to the call, absent from the returned text.

Model (OpenRouter)	Reasons by default	Disable?	Reasoning tokens/call (trivial → hard)	Trace
openai/gpt-5	Yes	No — API rejects	85 → 141	hidden

Model (OpenRouter)	Reasons by default	Disable?	Reasoning tokens/call (trivial → hard)	Trace
google/gemini-2.5-pro	Yes	No — API rejects	270 → 679	hidden
deepseek/deepseek-r1	Yes	No — API rejects	148 → 709	hidden
x-ai/grok-4.6	Yes	No — API rejects	160 → 274	hidden
anthropic/claude-sonnet-5	Yes, on the hard task	Yes	0 → 34	hidden
anthropic/claude-opus-4.8	No (even on hard)	Yes	0 → 0	—
deepseek/deepseek-chat	No	Yes	0 → 0	—
qwen/qwen3-max	No	Yes	0 → 0	—

The trap: hidden thinking that truncates — on some endpoints

When a model's hidden thinking is billed against the same budget as its output, a tight budget is consumed by the thinking and the answer never appears. On the trivial task at a 32-token ceiling, GPT-5 and Gemini fell from a perfect answer rate to zero, and DeepSeek R1 to sixty percent — each cut off mid-generation. A harness that only reads the final answer scores these as wrong answers to "17 + 26," a broken cell mistaken for a legitimate low result. Grok is the exception that proves the hazard: it reasons as heavily as the others, spending one hundred and seventy-five tokens even at the tight budget, yet still answers, because its endpoint bills reasoning separately from output. Whether the trap fires is an opaque, per-endpoint accounting detail no measurer can predict without probing — which is an argument for instrumenting emission health, not against it.

Model	Valid-answer rate, ample → tiny budget	Outcome at the tight budget
openai/gpt-5	100% → 0%	thinking consumed the budget; no answer
google/gemini-2.5-pro	100% → 0%	same
deepseek/deepseek-r1	100% → 60%	partial truncation
x-ai/grok-4.6	100% → 100%	reasons 175 tokens but still answers (separate budget)

Why it exists: the non-reasoning path fails the work

Reasoning by default is not a quirk to route around; it is load-bearing. On the harder task, every model that reasoned answered correctly every time, while the models that did not reason failed it — DeepSeek Chat never got it right, Qwen managed it two or four times in ten, and even Claude Opus, which kept thinking off by default, dropped to eighty percent and recovered to perfect only once thinking was switched on. The non-reasoning baseline is not just a measurement artifact; on tasks that need reasoning it is genuinely less capable, which is exactly why vendors make reasoning the default and, increasingly, mandatory. The measurement hazard is the side effect of a capability decision — which is why it will not be fixed, and has to be instrumented around.

Model	Reasoned?	Correct on the hard task
gpt-5, gemini-2.5-pro, deepseek-r1, grok-4.6, sonnet-5	Yes	100%
anthropic/claude-opus-4.8 (default)	No	80% — and 100% with thinking on
qwen/qwen3-max	No	20–40%
deepseek/deepseek-chat	No	0%

Why this is Fathom's ground

When a model's reasoning is increasingly mandatory, hidden, and discarded between turns, the agent's model of its own work has moved off the observable stream and out of the operator's control. Three consequences follow, and each is where Fathom sits. Aggregate evaluations mis-score, because a truncated reasoning cell is indistinguishable from a capability failure unless an emission-health check and a parse-rate floor are watching — and no measurer can predict which endpoints truncate. The audit trail is gone by construction, because the reasoning is billed and thrown away, leaving a deterministic reconstruction of what the agent actually committed — from its own action and tool-return stream — as the only recoverable ground truth. And a fair non-reasoning comparison is simply unavailable on four of five vendors, so any attribution that leans on a matched non-reasoning arm must be built the way Fathom builds it, by difference of outcomes, not by a configuration flag the API refuses. Hidden, mandatory reasoning does not weaken the case for a committed-state record. It is the reason for it.

Scope

These are reproducible observations at one point in time, through one mainstream aggregator, on two tasks, five samples per cell. The mandatory behavior is a property of these endpoints as routed on the day of the run; a vendor's native API may differ, and endpoints change. This is not a verdict on any model's quality — it is a measurement hazard that anyone building agent evaluation on these endpoints inherits, and every claim here is tied to a figure the shipped probe reproduces.