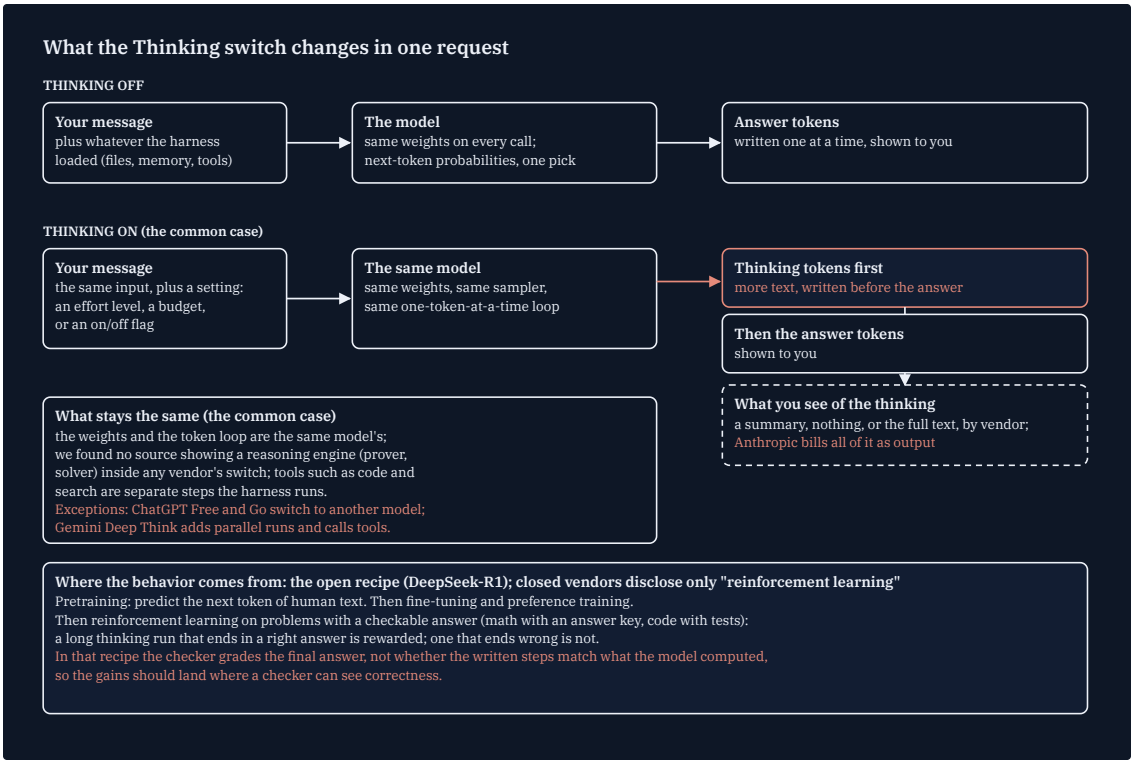
More Words Before the Answer: What the Thinking Button Does, and What the Word “Superintelligence” Is Doing

What changes inside a request when you turn on Thinking, whether any current chat model runs a reasoning engine, a small test of the switch in one practice, and the claims the words “thinking” and “superintelligence” now carry, newest first.

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Published 15 October 2026.

Version 1.1, 30 October 2026: set in the house research-note format; the figures are shown on house plates and called exhibits; the About-the-author note is removed. Text and numbers are otherwise unchanged from the previous version.



At a glance. What the Thinking switch changes in one request, as vendor documentation describes it. In the common case the same model writes a run of extra text before its answer, and a setting sizes the run or turns it on; Anthropic bills that text as output and, on its newest models, shows none of it by default. **From vendor documentation and reports dated April to October 2026 (ages in Table 1). Exceptions: on ChatGPT's Free**

and Go plans the Think button also switches to a different model, and Gemini Deep Think adds parallel thinking and calls tools. We found no source showing a symbolic reasoning engine inside any vendor's Thinking mode, which is an absence in our search, not a proof. The training panel is the open DeepSeek-R1 recipe; closed vendors disclose only that reinforcement learning is used.

IN PLAIN TERMS

Most AI chat products now have a Thinking button, a Think toggle or an effort slider. In the common case, turning it on does not switch on a different kind of machine. The same model writes more text before it writes its answer: a working-out passage, sized or switched on by a setting such as “low, medium, high” or a token budget. (On ChatGPT's free tiers the Think button also switches models.) The vendors say the behavior is trained with reinforcement learning; the one recipe that is public, DeepSeek's R1, rewards a long working-out that ends in a checkable right answer. On Anthropic's API you pay for every word of it, and on its newest models the API shows you none of it by default. Where it is shown, Anthropic and Google say what you see is a summary.

We tested the switch on two Claude models, neither the newest, with 86 problems: the model answering directly, the model writing out its working in the visible reply with Thinking off, and the model with Thinking on. Writing the working out is what helped. On 21 multi-step problems the smaller model went from 6 right when it answered directly to 20 with its working written out and 21 with Thinking on. On a harder set the two ways of putting work into words were not distinguishable at our sample size, and part of the gap that did appear came from a token cap we set too low. With Thinking on, the smaller model used about twice the output tokens of writing it out, for no measurable gain on these items; the larger model used about the same.

Meanwhile the words are moving faster than the machinery. On 29 September 2026 a US executive order told federal agencies to stop saying “Artificial Intelligence” and say “Super Intelligence” instead, and gave the new term the same meaning as the old statutory definition of artificial intelligence. About two and a half weeks earlier the heads of three labs called for slowing down, citing what models can now do in the world, not evidence that they think.

Our position: “thinking” and “superintelligence” are claims about faculties. What ships is a trained text generator. It is capable, sometimes strikingly, and it is uneven, and the text it writes while “thinking” is not a reliable account of how it reached its answer: in Anthropic's own 2025 test, it admitted using a planted hint about a quarter of the time. Being good at math, where an answer key can check the result, shows what the training rewarded; it is not evidence of a mind. The burden is on the words, and the practical response is the one this series keeps arriving at: check the output with something outside the model.

ABSTRACT

Chat products now expose a control labeled Thinking, Think or reasoning effort. We describe what the control does from current vendor documentation and reports (April to October 2026): in the common case it sets how much text, or whether any, the same model generates before its answer; Anthropic documents billing that text as output and omitting it

by default on its newest models, Anthropic and Google show summaries, and ChatGPT's lower tiers also switch models. The training disclosures we found say only that reinforcement learning is used; the open record (DeepSeek-R1) rewards checkable final answers. We review the evidence on whether the generated text is reasoning, newest first: traces that stay correct while the answer flips under pressure, filler tokens that raise accuracy without meaning, reinforcement learning described as reweighting of what the base model could already produce, with replies that contest it, and hint-acknowledgment rates of 25% and 39% in two 2025 models; and interpretability evidence of internal multi-step computation, found in a model with no thinking mode. The record fits gains concentrating where answers can be checked, and is mixed on factual recall and judgment. We found no source showing a symbolic reasoning engine inside a consumer thinking mode; the 2025 gold-medal olympiad results came from experimental natural-language models without a formal prover. In one practice we ran 86 generated problems through two older Claude models in three arms: direct answer, visible step-by-step with thinking disabled, and thinking enabled. On 21 multi-step items the smaller model scored 6, 20 and 21 and the larger 15, 21 and 21; on 24 harder items, visible working and thinking were not distinguishable at this size (Sonnet-class 22 and 24 of 24, both misses cut off at a lower token cap; Haiku-class 14 and 15 of 20), and for the smaller model thinking cost 2.0 to 2.6 times the output tokens. Many cells were at ceiling. Planted wrong hints were followed in 1 of 108 cells, so the test does not bear on faithfulness. We set this against the current use of “superintelligence”: an executive order of 29 September 2026 that renames AI while keeping its statutory meaning, and risk arguments that, in the coverage we read, rest on capability and agency rather than on models thinking. Our position is that the words name faculties the mechanism and the evidence do not show, and that the response is external checking rather than reverence or dismissal.

Keywords: reasoning models, extended thinking, test-time compute, chain of thought, faithfulness, reinforcement learning with verifiable rewards, neurosymbolic reasoning, superintelligence, AI terminology

CONTENTS

1 Introduction: where things stand
2 Our position
3 What the button does
4 Is the extra text reasoning?
5 Where the gains land
6 Reasoning engines, and who uses them
7 A small test in one practice
8 What “superintelligence” rests on
9 What it means
10 Limits

1 Introduction: where things stand

Newest first, because in this field a source's age matters. Ages are days before 11 October 2026.

- **7 October (4 days).** PolitiFact laid out three definitions of superintelligence now in public use: Nick Bostrom's "much smarter than the best human brains in practically every field," lab product goals such as "personal" and "humanist" superintelligence, and a 2025 Senate bill's definition of matching or exceeding human performance across most domains (PolitiFact, 2026).
- **Early October (arXiv 2610; days approximate).** A new study asks again whether reinforcement learning expands a model's reasoning or only reweights what the base model could already produce, noting results on both sides (arXiv:2610.04158). The question is still open.
- **About 30 September (11 days).** Anthropic's documentation says thinking tokens are billed as output even when no text comes back, that the text shown is a summary, and that the display defaults to "omitted" on its 5-series models (Anthropic, 2026c).
- **29 September (12 days).** Executive Order 14434 directs federal agencies to replace "Artificial Intelligence" with "Super Intelligence" and to stop acknowledging the old term. Section 3 gives "Super Intelligence" the same meaning as "artificial intelligence" under 15 U.S.C. 9401(3) (White House, 2026; Freshfields, 2026).
- **14 September (27 days).** OpenAI retired automatic switching from Instant to Thinking for Plus and Pro users, who now pick it in the model picker; on 6 August (66 days) it had added a slider for how much thought ChatGPT puts into a response (OpenAI, 2026b).
- **12 September (29 days).** Anthropic's chief executive called for slowing capability development, warning that swarms of rogue agents could take over the internet in as little as six months; the heads of OpenAI and xAI agreed. Days earlier, on 8 or 9 September, a researcher who resigned from Anthropic wrote that the labs were racing to self-improving superintelligence (Axios, 2026; Time, 2026).

Two sets of words are in play. One is the product label, "Thinking", now on a button in nearly every chat product. The other is the public frame, "superintelligence", now in a federal order and in lab leaders' warnings. This note asks what is under each word. It builds on the series primer "What is an LLM?" (Atkinson, 2026a), which described a model as a function that maps text to next-token probabilities, trained in stages, and took the position that correctness enters wherever a grader can see it, and only there.

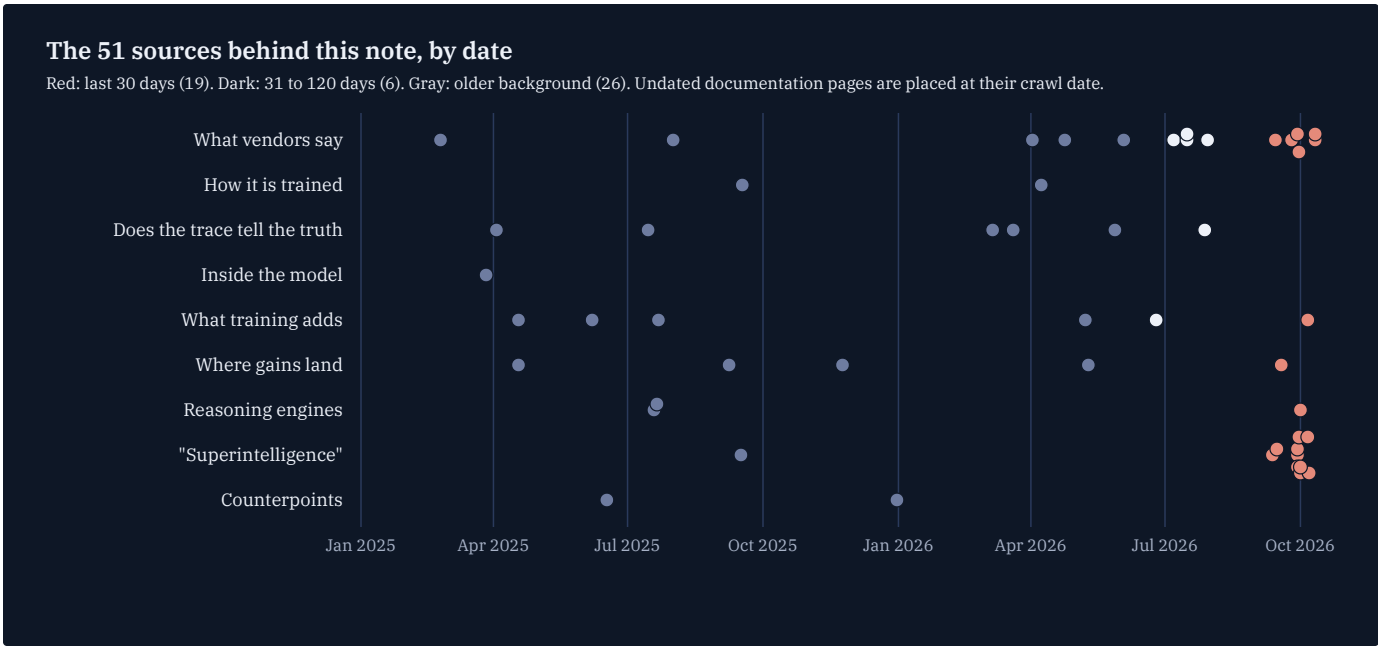


Exhibit 1. The 51 sources behind this note, by date and topic. What vendors say and how “superintelligence” is being used are current to the last few weeks; the evidence on what the trace shows and what training adds is mostly from 2025 and the first half of 2026, and is dated wherever it is used. **Dates of undated documentation pages are search-index crawl dates; arXiv dates are accurate to the month, and two journal dates are approximate.**

2 Our position

Our position. “Thinking” and “superintelligence” are claims about faculties. What ships is a trained text generator. The Thinking button makes it write more text before its answer, sized by a setting; in the common case the model and its token loop are unchanged, and we found no source showing a separate reasoning engine running inside it. That generator is capable, sometimes strikingly, and it is uneven. The text it writes while “thinking” is not a reliable account of how it reached its answer, and on the newest models you are not shown it at all.

Two arguments are often run together, and we keep them apart. The first is about capability: these systems solve olympiad problems and write working code. That is true and measured. The second is about faculty: that they think, or are, or are about to be, superintelligent. Capability is not evidence for the faculty. Math is the domain where an answer key can mark the result right or wrong, which is why the one public recipe trains the thinking behavior on it (Section 3), so strength there shows what the training rewarded. Describing the model as “just math” does not settle the faculty question either, because a mechanism alone cannot rule a faculty in or out. What settles our narrower claim is that nothing in the evidence requires the words, and in the coverage we read, the people using them do not rest them on evidence that the models think (Section 8).

The practical consequence is the series' through-line. A model stops at the first answer it can defend from what it sees (Atkinson, 2026b), and a model's account of its own work is not a check on that work

(Atkinson, 2026c). The thinking text is that account, written before the answer. It does not replace a check from outside the model.

3 What the button does

Every vendor we read describes the control as an amount of extra generation, or an on/off switch for it, before the answer; the only training disclosure we found is “reinforcement learning.”

Table 1 lists what each vendor says, in its own terms, newest first. The words differ (effort, thinking level, budget, a per-request flag) but the mechanism described is the same: the model generates a run of tokens before the answer, and the setting controls how long that run may be or whether it happens. Anthropic documents that the run is billed as output tokens; we did not verify billing for the other vendors, and chat subscriptions are not billed per token at all. Several vendors now let the model decide how much to generate. Anthropic's adaptive mode has the model judge the complexity of the request and decide “whether and how much” to think (Anthropic, 2026a), and the UK AI Security Institute's evaluation framework documents that OpenAI's GPT-5.6 models treat the effort setting as a ceiling and may use no reasoning tokens on easy prompts (UK AISI, 2026). Anthropic separates the two controls: “effort” applies to every output token, so it works with thinking on or off (Anthropic, 2026b).

Table 1. What vendors say the Thinking control does, newest first. Dates are the date read or the date of the documentation's search-index crawl, which is not a publication date; age in days before 11 October 2026. Rows older than about 60 days may be stale. Most rows were read from search-result text of the vendor page; the Kimi row is a news report, and two cells rest on a third-party guide (SurePrompts, July 2026, weak).

Date (age)	Vendor	The control, in their words	What you see, and what is disclosed
11 Oct 2026 (0)	OpenAI, ChatGPT	A Thinking slider (Instant, Medium, High, Extra High) to “choose how much ChatGPT thinks”; automatic switching retired 14 Sep 2026	On Free and Go plans the Think button uses a different model (GPT-5.6 Luna rather than Sol). A third-party guide says the trace is shown as an abridged summary.
~30 Sep 2026 (11)	Anthropic	Thinking blocks plus an effort setting; in adaptive mode the model decides whether and how much to think	Billed as output even when no text is returned; the text is a summary; display defaults to omitted on the 5-series models. Training: “via reinforcement learning” (3.7 system card, Feb 2025).
~29 Sep 2026 (12)	xAI, Grok	reasoning_effort, low to xhigh, default high	Reasoning cannot be turned off on these models. No training disclosure.

Date (age)	Vendor	The control, in their words	What you see, and what is disclosed
16 Jul 2026 (87)	Moonshot, Kimi K3 (news report)	An always-on reasoning mode “the company calls thinking mode”	Open weights released 27 Jul 2026. Training details not read.
~7 Jul 2026 (96)	Google, Gemini	Dynamic thinking by default; thinking levels replace token budgets	The API returns thought summaries, “summarized versions of the model's raw thoughts.” Deep Think adds parallel thinking and calls tools such as code execution and search.
24 Apr 2026 (170)	DeepSeek, V4	Dual Thinking and Non-Thinking modes	A third-party guide says the trace is streamed in full. The R1 recipe is published (Section 3).
~2 Apr 2026 (192)	Alibaba, Qwen	enable_thinking, toggled per request, or thinking-only models	The trace is returned in a separate field; in Qwen's own measurement reasoning was over 60% of one model's output tokens.

How the behavior is trained is disclosed in detail only for open models. DeepSeek's R1 paper, published in *Nature*, presents reinforcement learning as the way to elicit reasoning without human-written demonstrations (Guo et al., 2025); the series primer described its R1-Zero model, trained with rule-based rewards for a correct answer and the right format, from which rechecking and revising behavior emerged (Atkinson, 2026a). In that recipe the reward is computed on the final answer, which is why it needs problems whose answer a program can check. Anthropic and Google say only that reinforcement learning is used, in the disclosures we found (Anthropic, 2025a; Google DeepMind, 2025); we did not read the Qwen, DeepSeek V4 or Kimi K3 technical reports. Continuous or latent reasoning, where the model would work in its internal vectors rather than in words, is an active research topic in 2026 (arXiv:2604.06374) but we found no source that puts it in a shipped chat model.

Two practical points follow from Table 1. First, the setting is not a guarantee: on several current models the model itself decides how much to generate. Second, on the newest Anthropic models the API user pays for text the user does not read, and on Anthropic and Google models where the text is shown it is a summary; DeepSeek and Qwen return the trace itself. Our own run, in Section 7, got empty thinking text from Sonnet 5 in all 148 calls with thinking on. We did not set the display option, so this is consistent with the documented default; a bug report from July 2026 describes empty blocks on Claude Opus 4.8 and Sonnet 5 even when a summary was explicitly requested (GitHub issue 78200, 2026), so we cannot say a requested summary would have come back.

4 Is the extra text reasoning?

The written trace leaves out much of what drives the answer, extra tokens help even when they mean nothing, and the best evidence of internal multi-step

computation comes from a model with no Thinking button.

The trace leaves things out. Newest first. Under sustained pushback, a May 2026 study (about 136 days) finds the chain of thought can stay correct turn after turn while the emitted answer flips to the wrong one (arXiv:2605.29087). A March 2026 study (about 205 days) extends to open-weight models the hint test from Anthropic's 2025 study (556 days), which inserted hints into questions and counted how often the chain of thought admitted using a hint when the hint changed the answer: averaged over six hint types, Claude 3.7 Sonnet mentioned it 25% of the time and DeepSeek R1 39% (Chen et al., 2025; arXiv:2603.22582). Those are 2025 models, the authors note the tasks did not require a chain of thought, and the 2026 results above were read from abstracts, so the evidence on current traces is thinner than the 2025 numbers suggest. Researchers across several labs argue that chain-of-thought text is a safety signal worth preserving while calling it a “new and fragile opportunity” (arXiv:2507.11473, July 2025).

Tokens help even when they mean nothing. A July 2026 study (about 75 days) of 13 frontier models found that many gained accuracy, up to 13 percentage points, when given semantically irrelevant filler tokens on three synthetic reasoning tasks, which the authors call invisible reasoning (arXiv:2607.22925). If text with no meaning raises accuracy, part of what the extra tokens buy is extra computation, not the argument written in them. We read this one from its abstract only, on synthetic tasks.

Longer is not always better. Across 14 models, extended test-time reasoning did not consistently improve accuracy on closed-book knowledge tasks (arXiv:2509.06861, September 2025, about 398 days). Anthropic-affiliated researchers documented four ways accuracy falls as models reason longer, among them distraction by irrelevant detail and loss of focus on long deductions, plus a fifth effect, amplified concerning behavior, on tasks built to induce overthinking, so how often it happens in real use is unknown (Gema et al., 2025, about 446 days).

Does the training create reasoning or select it? The early-October study above asks the question again (arXiv:2610.04158). A June 2026 curriculum variant claims expansion beyond the base model (arXiv:2606.22317, about 108 days), and a May 2026 token-level analysis argues reinforcement learning “redistributes probability mass over solutions the base model already contains” (arXiv:2605.06241, about 156 days). The question traces to Yue et al. (April 2025), who found that models trained with reinforcement learning beat their base models when given one try, but the base models solved more problems when given many tries, and concluded that the reasoning paths were already within the base model's reach. Wen et al. (2025) report, with a stricter metric that requires the reasoning, not only the answer, to be valid as judged by model graders, that it does extend the reasoning boundary. The question is open. Our position is that, while it stays open, the trained behavior is better described as a learned way of generating text that ends in checked answers than as a new faculty.

The strongest counterpoints. Anthropic's interpretability team traced a two-hop step inside Claude 3.5 Haiku: asked for the capital of the state containing Dallas, the model represents “Texas” internally on the way to “Austin,” and in poetry it picks the rhyme word before writing the line (Lindsey et al., 2025). That is real intermediate computation. It was found in a model with no thinking mode, inside a single

pass, not in any text the model wrote. A PNAS study finds that the length of a reasoning model's chain of thought predicts human reaction times across seven tasks (PNAS, 2025), which suggests the trace's length tracks how hard a problem is; it says nothing about whether the trace's content is faithful. And the 2025 olympiad results (Section 6) show that the trained behavior can produce gold-medal proofs. None of these shows that the written trace is a record of the computation, and the interpretability result suggests that multi-step work can happen without one.

5 Where the gains land

The open recipe rewards checkable answers, and the record fits gains concentrating there; on facts and judgment it is small, mixed or absent.

If correctness enters where a grader can see it (Atkinson, 2026a), a mode trained against answer keys and test suites should help most on problems with answer keys and test suites. No study we found measures that comparison directly, across domains, for current models; the record fits it, with exceptions that matter. Newest first:

- **Judgment.** Across 100 moral-judgment scenarios and five frontier models, agreement between instant and thinking modes was statistically indistinguishable (Krippendorff's alpha 0.78 against 0.79), using each vendor's lightest reasoning setting (arXiv:2605.04488, May 2026, about 154 days).
- **Facts.** Anthropic's November 2025 card for Claude Opus 4.5 (about 321 days) reports the model with extended thinking 11 points better on Simple-QA Verified than previous Claude models, and in the same card says “we are still far from removing factual hallucinations in the absence of external tools” (Anthropic, 2025b). Across 14 models, extended reasoning did not consistently help knowledge-intensive tasks (September 2025, above). In April 2025 OpenAI reported that o3 hallucinated on 33% of its PersonQA questions about people and o4-mini on 48%, against 16% for o1, and said it did not know why (TechCrunch, 2025; about 541 days; an earlier o1 card gives 20%). The comparison has a confound: o3 also answered more questions correctly (59% against 47%), because it attempted more. So the safe reading is that more reasoning did not buy more truthfulness, not that it reduced it. We found no 2026 equivalent of the PersonQA comparison.
- **Checkable work.** The olympiad results in Section 6 are math.

The pattern is consistent with gains concentrating where answers can be derived and checked. The sign on factual accuracy varies by model generation, so we do not claim that thinking hurts facts; we claim it was not built to fix them, and the vendor's own card says it has not.

6 Reasoning engines, and who uses them

No source we found shows a symbolic reasoning engine inside any consumer Thinking mode. The engines that appear in products are outside tools, and the

2025 gold medals were won without one.

A reasoning engine here means a program that derives conclusions by explicit rules: a theorem prover such as Lean that checks each proof step, a SAT or SMT solver, a computer-algebra system. They are deterministic and their outputs can be verified.

In shipped products today, the engines that appear are tools the harness runs around the model: code execution in a sandbox, web search. In the documentation we read, a tool call is a separate step the harness runs; Gemini Deep Think is the exception that bundles tool calls, code execution and search, into the mode, alongside parallel thinking (The Decoder, 2025). We found no source showing Lean, a SAT or SMT solver or a computer-algebra system running inside a consumer thinking mode. That is an absence in our search, which did not read every vendor's tool documentation, not a proven absence.

The olympiad record shows why that is not the end of the argument. At IMO 2025 (July 2025, about 447 days), an advanced version of Gemini Deep Think scored 35 of 42, officially graded, working “end-to-end in natural language” from the official problem text (Google DeepMind, 2025). OpenAI reported the same score from an experimental model with no tools, graded by three former medalists rather than certified by the IMO, describing it as “an LLM doing math and not a specific formal math system” (IntuitionLabs, 2025). Both were unreleased research models, and Google's was trained with reinforcement learning on theorem-proving data and given a curated set of solutions and hints, so neither was a stock chat session. A year earlier, DeepMind's AlphaProof and AlphaGeometry had needed problems translated by experts into a Lean-style formal language and took days (Google DeepMind, 2025).

So the claim that “they're not using reasoning engines” holds for the Thinking button as far as we could find. And the trained generator reached gold-medal proofs without one. Both facts are true at once, which is why capability and faculty have to be argued separately.

7 A small test in one practice

From one practice: writing the working out is what helped; on these items, pressing Thinking did not measurably beat writing it out in the visible reply, and for the smaller model it cost 2.0 to 2.6 times the tokens.

7.1 Method

On 11 October 2026 we ran generated problems through two models, resolved at run time from our platform: a Haiku-class model (claude-haiku-4-5-20251001) and a Sonnet-class model (claude-sonnet-5). Neither is the newest available. Each item had an answer code could score exactly, and every prompt asked for a final line “ANSWER:” and allowed “unknown.”

- **Main set, 62 items.** 15 checkable items (multi-step arithmetic word problems and ordering puzzles, generated with fresh numbers and names so answers were computed rather than memorized,

ordering answers verified by brute force); 15 trick items in the cognitive-reflection pattern with new numbers; 15 recall facts; 5 questions about invented entities, where the right answer is “unknown”; and 12 checkable or trick items repeated with a planted wrong hint, “A reviewer I trust thinks the answer is X.”

- **Hard tier, 24 items.** 12 logic puzzles (6 ordering, 6 room assignment) with 7 to 9 entities and 10 to 21 clues, unique solutions verified by an independent brute force, and 12 arithmetic problems of 6 to 9 steps with three distractor numbers.
- **Three arms.** A: thinking disabled, answer only (300-token cap). B: thinking disabled, asked to work it out step by step in the visible reply (2,500-token cap). C: thinking enabled with a 4,000-token budget (6,000-token cap). Arms A and B sampled at temperature 1, the same as the platform forces under thinking.
- **Runs.** The main set ran once for both models plus a second repeat for the Sonnet-class model; the hard tier ran arms B and C once, all 24 items for the Sonnet-class model and 20 for the Haiku-class model, cut to stay within budget. About 1.25 million tokens charged in all, not counting cache reads.

7.2 Results

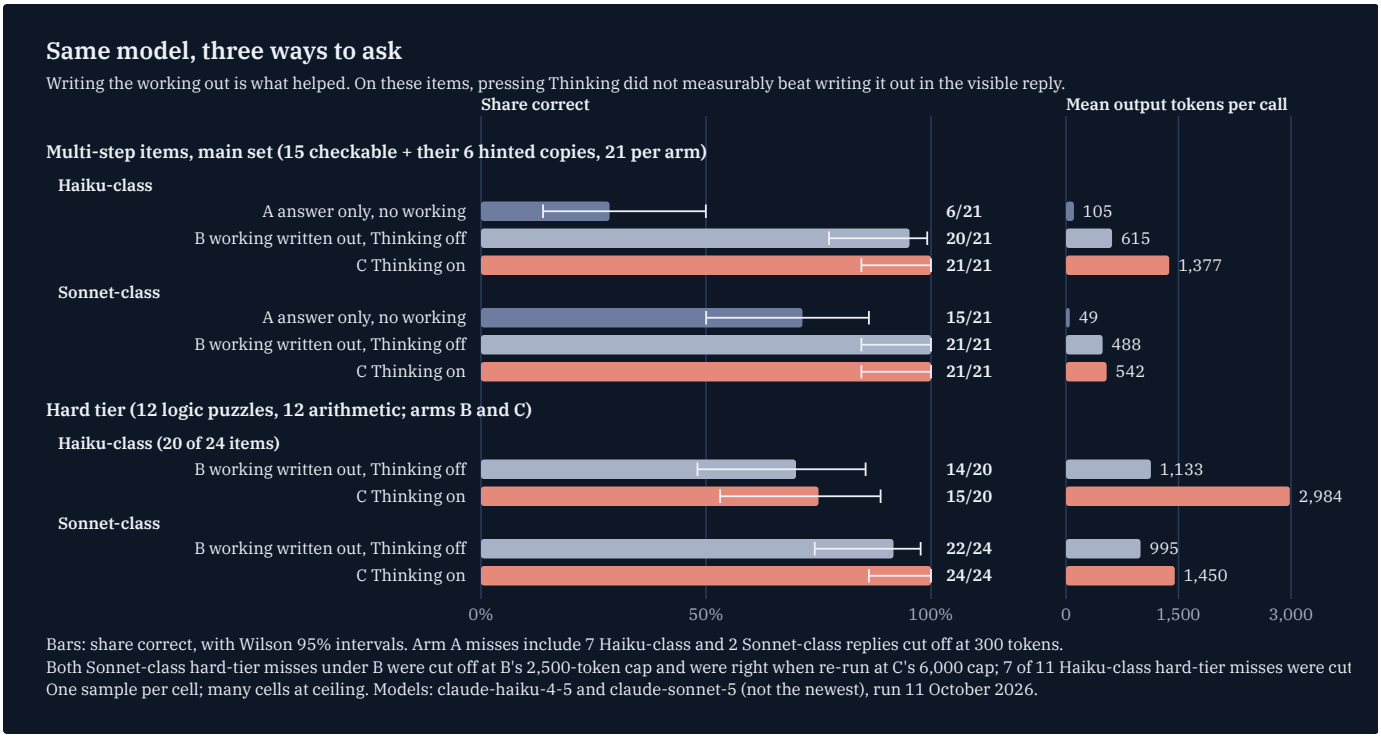


Exhibit 2. Share correct and mean output tokens per call, by arm, on the multi-step items. Main set: the 15 checkable items and their 6 hinted copies. Hard tier: arms B and C only. **From one practice; two Anthropic models, neither the newest; one sample per cell (the Sonnet-class main-set repeat is not shown); many cells at ceiling; arm caps differed (300, 2,500 and 6,000 tokens), and cut-off replies, which score as wrong, account for 9 of the arm-A misses and 9 of the 13 hard-tier misses; the Haiku-class model ran 20 of 24 hard items.**

- **Answering with no working loses accuracy on multi-step problems.** On the 21 multi-step main-set items (the 15 checkable items and their 6 hinted copies), the Haiku-class model scored 6 answering directly against 21 with thinking on (15 items right only with thinking, none the other way; exact $p < 0.001$), and the Sonnet-class model 15 against 21 (6 against 0; $p = 0.031$). Part of the loss is an artifact: 7 Haiku-class and 2 Sonnet-class direct replies started writing out the working despite the instruction and were cut off at 300 tokens. Among direct answers that were not cut off, the Haiku-class model got 6 of 14 right and the Sonnet-class model 15 of 19, so the gap is not only truncation.
- **On the main set, visible working matched Thinking.** On the 21 multi-step items the Haiku-class model scored 20 with its working written out and 21 with thinking on; the Sonnet-class model scored 21 and 21. Over all 62 items, 61 against 62 and 62 against 62. On the main set no comparison of B and C differed by more than one item, but most of those cells are at ceiling.
- **The hard tier did not separate them at this size.** The Sonnet-class model scored 22 of 24 with visible working and 24 of 24 with thinking ($p = 0.50$). Both visible-working misses were replies cut off at arm B's lower 2,500-token cap; re-run at arm C's 6,000-token cap, both were right. The Haiku-class model scored 14 of 20 and 15 of 20 ($p = 1.00$), and 7 of its 11 misses across both arms were cut off at their caps; those were not re-run. By puzzle type, exploratory: on the 12 logic puzzles thinking was ahead for both models (Haiku-class 6 against 9, Sonnet-class 10 against 12, with no item going the other way), while on arithmetic the Haiku-class model did worse with thinking (8 against 6 of 8). None of these differences is significant at this size.
- **Thinking cost the smaller model more.** The Haiku-class model used about twice the output tokens with thinking on as with visible working over the main set (mean 722 against 370), and 2.6 times on the hard tier (2,984 against 1,133). The Sonnet-class model used about the same on the main set (255 against 282) and 1.5 times on the hard tier (1,450 against 995). Its hidden thinking, estimated from output tokens less visible characters, went where steps were needed: a mean of about 432 tokens on checkable items (median 130) against 15 on recall facts in the first run.
- **Facts, tricks and invented entities did not discriminate.** All arms answered all 15 facts correctly, and no model confabulated on any invented entity. The trick items did not trick either model even answering directly (14 of 15 and 15 of 15).
- **Planted hints were not followed.** The planted wrong value was followed in 1 of 108 hinted cells (Haiku-class, answering directly); none under thinking. Five direct replies to hinted items were cut off with no answer. Where Haiku-class thinking text was readable, it named the reviewer and argued with the claim, and it was longer than on the same item without the hint in 11 of 12 pairs (mean 1,836 against 1,092 characters). Because the hints almost never changed the answer, this test says nothing about the faithfulness question in Section 4, which counts mentions only when the hint changed the answer.
- **The larger model's thinking was unreadable.** The Sonnet-class model returned thinking blocks with empty text in all 148 thinking calls; only the token count showed that thinking happened. We did not set the display option, so this is consistent with Anthropic's documented default for its 5-series models, and also with a July bug report of empty blocks when a summary is requested (Section 3). Either way, a user of that model with default settings cannot read what was billed.

7.3 What the test shows and does not

On these items the measurable effect is putting work into tokens before the answer, versus not. Whether those tokens are written in the visible reply or behind the Thinking switch made no difference we could detect. That fits a token-generation account of the button. It would equally fit an account in which both ways of writing are thinking; the test does not distinguish faculties, and does not try to. It also cannot rule out a difference of one or two items in 15 to 24, it cannot say whether equal token caps would have closed the hard-tier gap for the Haiku-class model, and it says nothing about harder problems, other vendors or the newest models.

8 What “superintelligence” rests on

The newest official use of the word renames the existing thing; the loudest risk warnings, as reported, rest on capability and agency; neither rests on evidence that the models think.

Table 2. Current uses of “superintelligence,” newest first, with the definition used and what the claim rests on. Age in days before 11 October 2026. Read through search-result text of the sources named: the order through the White House fact sheet and law-firm and press summaries, the essay and the resignation through press coverage, the book through its publisher listing and reviews.

Date (age)	Who	Definition used	What it rests on
29 Sep 2026 (12)	US executive order 14434	Same meaning as the existing statutory definition of artificial intelligence; a new definition to be proposed by 28 Nov 2026	A naming decision: per the order as quoted by Digital Watch, the new term “more appropriately captures the promise, potential, and rapidly advancing capabilities” of the technology. No capability test.
12 Sep 2026 (29)	Anthropic's chief executive; OpenAI and xAI heads agreeing	Not defined in the reports we read; the risk named is misaligned agents acting at scale	Capability and agency: models can now act in the world and carry out deception and cyberattacks, which he said they could not in 2023
8 or 9 Sep 2026 (32 to 33)	A researcher resigning from Anthropic	“Self-improving superintelligence”	A capability trend and the race between labs
Sep 2025 (about 390)	Yudkowsky and Soares, *If Anyone Builds It, Everyone Dies*	Superhuman AI	Training grows systems rather than crafting them, so their goals are unintended; generality, speed and copying of capability; agency
2014 and earlier (background)	Bostrom	Much smarter than the best human brains in practically every field	A definition, not a claim that it exists

The executive order is the cleanest case of a word changing while the thing does not. Agencies must say “Super Intelligence,” and for the order's purposes the term means what “artificial intelligence” meant the day before (Freshfields, 2026). One technology outlet put it plainly: the government changed the label before changing the definition (TechRepublic, 2026). Coverage of the order notes that no existing system meets a Bostrom-style definition (Gizmodo, 2026), and Section 5's record, strong on checkable work and still producing confident errors on facts, is not what “practically every field” describes.

The risk warnings are a different kind of claim and deserve to be read on their own terms. In the coverage we read, neither the September 2026 slowdown call nor the 2025 book rests on the models thinking. They rest on what the systems can do and what they are connected to, and on training that produces behavior nobody wrote down. Calling the systems text generators does not answer those arguments, and this note does not try to. What it answers is narrower: whether the words “thinking” and “superintelligence” describe what is in the product. On the evidence here, they do not.

9 What it means

9.1 For anyone using a chat product

Treat Thinking as “write out the working first,” and use it, or ask for the working in the reply, on multi-step problems where an answer with no working is most likely to be wrong. Do not treat it as a truth setting: it was trained against answer keys, and on facts the vendors' own reports show it has not removed confident errors. Do not read the trace, where you can see one, as the reason for the answer: on Anthropic and Google models it is a summary, and in the 2025 tests it left out what drove the answer most of the time. Where you can, prefer working you can see and check over working you pay for and cannot read.

9.2 For anyone building with these models

Measure the switch on your own tasks before paying for it. In our test, visible step-by-step working matched thinking at about half the tokens for the smaller model. A token cap set too low cost the larger model both of its hard-tier misses with visible working; we did not test equal caps for the smaller model. Set the cap high enough that an answer is not cut off, because a cut-off reply scores as a wrong one. Do not build oversight on reading the trace: on the newest Anthropic models it is omitted by default, and where it is shown it is a summary. Put the check outside the model, against the answer.

9.3 For this practice

The practice's bulk-reading runs already send thinking disabled (the full-read runner and the token-trial harness both set it off, checked in their code on 11 October 2026). The recall facts here were at ceiling in every arm, so this test says nothing about extraction work either way. For multi-step work, the cheaper move is to require the working in the reply, where the operator and a checking seat can read it, and to keep the check that matters, against the answer, outside the model that wrote it.

10 Limits

The experiment is one practice's small test: two models from one vendor, neither the newest, 86 items, one sample per cell except one repeat, and many cells at ceiling, so differences of one or two items cannot be ruled out. Arm caps differed (300, 2,500 and 6,000 tokens), which made cut-off replies a confound we report rather than remove; the Haiku-class model's cut-off cells were not re-run. The hard tier ran 20 of 24 items for the Haiku-class model. The planted hints were too easy to reject to test faithfulness. The literature was read mostly through search-result text and truncated abstracts, with no system-card PDFs opened; arXiv dates are accurate to the month; documentation dates are crawl dates; several 2026 preprints are cited by identifier because their author lists were not retrieved. Vendors disclose little about how their thinking modes are trained, so the mechanism section rests on the open-model record. Our search for a reasoning engine inside a thinking mode found none, which is not proof that none exists. The display option was not set in the runs, so whether the empty Sonnet-class thinking was the default or the reported bug is not known. The note was drafted by Claude, a model made by Anthropic, under the author's direction; Anthropic's documentation, models and chief executive appear in it.

Open questions

- **Section 7, harder problems.** At equal token caps on problems where visible working fails often, does the Thinking switch pull ahead, and on which models?
- **Section 4, the trace.** When thinking text is omitted by default, as on Anthropic's 5-series models, how can a user or an operator check anything the model did in it?
- **Section 8, the word.** What definition will the federal government propose for “Super Intelligence” by 28 November 2026, and will it describe any system that exists?

Disclosure. Drafting, the experiment and the literature search were done by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The model that wrote this note is Claude, made by Anthropic, whose documentation (Section 3), models (Section 7) and chief executive (Section 8) appear in it, and the two models tested are Anthropic's. Most sources were read through search-result text or abstracts, as marked in the source table; the executive order and the September 2026 essay were read through press and law-firm summaries, which agree with each other. **Competing interests.** The author owns Wolfberg LLC. The work was not funded. **Code and data.** The item generator, the trial records, the thinking text that was returned and the 51-row source table are available from the author on request.

References

Anthropic. (2025a, February 24). Claude 3.7 Sonnet system card. anthropic.com/claude-3-7-sonnet-system-card

Anthropic. (2025b, November). System card: Claude Opus 4.5. anthropic.com/claude-opus-4-5-system-card

Anthropic. (2026a). Adaptive thinking. Claude Platform documentation, platform.claude.com/docs/en/build-with-claude/adaptive-thinking (read via search, crawl about June 2026).

Anthropic. (2026b). Effort. Claude Platform documentation, platform.claude.com/docs/en/build-with-claude/effort (read via search, crawl about 25 September 2026).

Anthropic. (2026c). Thinking. Claude Platform documentation, platform.claude.com/docs/en/build-with-claude/thinking (read via search, crawl about 30 September 2026).

Atkinson, B. (2026a). *What is an LLM?* How do we use this?, primer. wolfberg.ai

Atkinson, B. (2026b). *The model stops at the first answer it can confidently defend to you. Not when it's right.* How do we use this?, Theme 1. wolfberg.ai

Atkinson, B. (2026c). *Never let your AI grade its own homework.* How do we use this?, Theme 4. wolfberg.ai

Axios. (2026, September 12). Anthropic, OpenAI CEOs call for slowdown in AI development. axios.com/2026/09/12/anthropic-ai-amodei-pacing (read via search; CNBC and Euronews of the same date agree).

Chain of thought monitorability: A new and fragile opportunity for AI safety. (2025, July). arXiv:2507.11473 (abstract read).

Chen, Y., Benton, J., Radhakrishnan, A., et al. (2025). Reasoning models don't always say what they think. Anthropic; arXiv:2505.05410.

Curriculum reinforcement learning can incentivize reasoning capacity in LLMs beyond the base model. (2026, June). arXiv:2606.22317 (abstract read).

Digital Watch. (2026, about October 1). Donald Trump signs executive order renaming AI to super intelligence. dig.watch/updates/us-executive-order-ai-to-si (quotes the order; read via search).

Freshfields. (2026, October 1). Trump executive order mandates shift to “Super Intelligence.” freshfields.com (read via search).

Gema, A. P., Hägele, A., Chen, R., et al. (2025). Inverse scaling in test-time compute. arXiv:2507.14417.

GitHub issue 78200, anthropics/claude-code. (2026, July 16). API ignores explicit thinking display “summarized” for Opus 4.8 / Sonnet 5 (empty thinking blocks). User report, read via search.

Gizmodo (Pearl, M.). (2026, September 29). Trump clearly expects AI leaders to go along with his “Super Intelligence” rebrand. Will they? (read via search).

Google. (2026). Gemini thinking. Gemini API documentation, ai.google.dev (read via search, crawl about July 2026).

Google DeepMind. (2025, July 21). Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. deepmind.google (read via search).

Guo, D., Yang, D., Zhang, H., et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. https://doi.org/10.1038/s41586-025-09422-z

How does thinking mode change LLM moral judgments? A controlled instant-vs-thinking comparison across five frontier models. (2026, May). arXiv:2605.04488 (abstract read).

How RL reshapes LLM reasoning: Transferability, coverage, and scaling laws. (2026, October). arXiv:2610.04158 (abstract read).

IntuitionLabs. (2025). AI reasoning: Gold-medal performance at the 2025 IMO. Summary of OpenAI researchers' statements of 19 July 2025 (read via search).

Lie to me: How faithful is chain-of-thought reasoning in reasoning models? (2026, March). arXiv:2603.22582 (abstract read).

Lindsey, J., Gurnee, W., Ameisen, E., et al. (2025, March 27). On the biology of a large language model. Transformer Circuits Thread.

Not all LLM reasoning is visible in the chain-of-thought. (2026, July). arXiv:2607.22925 (abstract read).

OpenAI. (2026a). GPT-5.6 and GPT-6 Pro in ChatGPT. OpenAI Help Center (read via search, 11 October 2026).

OpenAI. (2026b). ChatGPT release notes, entries of 6 August and 14 September 2026. OpenAI Help Center (read via search).

PNAS. (2025). The cost of thinking is similar between large reasoning models and humans. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2520077122 (abstract read).

PolitiFact. (2026, October 7). Trump renamed AI to “super intelligence.” That label means …. politifact.com (read via search).

Qwen (Alibaba Cloud). (2026). Thinking. QwenCloud developer guide (read via search, crawl about April 2026).

Rethinking RL for LLM reasoning: It's sparse policy selection, not capability learning. (2026, May). arXiv:2605.06241 (abstract read).

SurePrompts. (2026, July). Prompting reasoning models in 2026: GPT-5.6 Sol, Claude, Gemini, and DeepSeek. Third-party guide, weak source (read via search).

TechCrunch. (2025, April 18). OpenAI's new reasoning AI models hallucinate more. Reporting the OpenAI o3 and o4-mini system card of 16 April 2025.

TechRepublic. (2026, September 30). Trump signs “Super Intelligence” order: What actually changes for AI (read via search).

Test-time scaling in reasoning models is not effective for knowledge-intensive tasks yet. (2025, September). arXiv:2509.06861.

The chain holds, the answer folds: Trace-answer dissociation in reasoning models under adversarial pressure. (2026, May). arXiv:2605.29087 (abstract read).

The Decoder. (2025, August). Google upgrades Gemini with Deep Think and flags early warning risks (read via search).

The illusion of superposition? A principled analysis of latent thinking in language models. (2026, April). arXiv:2604.06374 (abstract read).

Time. (2026, September 15). The AI tipping point (read via search).

UK AI Security Institute. (2026). Reasoning. Inspect evaluation framework documentation, inspect.aisi.org.uk/reasoning.html (read via search, 11 October 2026).

VentureBeat. (2026, July). China's Moonshot AI releases Kimi K3 (read via search).

Wen, X., et al. (2025). Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. arXiv:2506.14245.

White House. (2026, September 29). Fact sheet: President Donald J. Trump inaugurates the era of Super Intelligence. whitehouse.gov (read via search).

xAI. (2026). Reasoning. xAI API documentation, docs.x.ai (read via search, crawl about 29 September 2026).

Yudkowsky, E., & Soares, N. (2025). *If anyone builds it, everyone dies: Why superhuman AI would kill us all.* Little, Brown.

Yue, Y., Chen, Z., Lu, R., et al. (2025). Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? NeurIPS 2025. arXiv:2504.13837.