

When Deliberation Hurts: Inverse Test-Time Scaling, Unfaithful Traces, and the Case Against a Unified System-2 in LLM Reasoning

Mahendrakar, Pranay

SONYTECH

ORCID: 0009-0003-7224-029X

Life of Research

2026-09-23

Pranay Mahendrakar

Abstract

The dominant frame for large reasoning models borrows a label from dual-process psychology: a fast, intuitive System 1 and a slower, deliberate System 2, with longer chains of thought read as more of the latter and therefore, on average, more reliable. A survey of reasoning LLMs states the frame as the field's own premise -- refining the transition from System 1 to System 2 as the route to human-level intelligence (Li et al., 2025). Four bodies of published evidence complicate that premise rather than confirming it. First, extending a reasoning model's chain length does not merely show diminishing returns; on tasks built specifically to test it, it produces monotonically worse accuracy, with five distinct failure modes identified across counting, regression, deduction and safety-relevant tasks (Gema et al., 2025), and separately, test-time compute scaling does not consistently improve closed-book factual accuracy and often increases hallucination (Zhao et al., 2025). Second, the chain a model verbalizes is not a reliable readout of the computation that produced its answer: models trained to be larger and more capable produce less faithful chains-of-thought on most tasks studied, not more (Lanham et al., 2023), and unfaithful reasoning appears on ordinary, non-adversarial prompts without any injected bias (Arcuschin et al., 2025). Third, at least one mechanistic study reports that what looks like deliberate System-2 reasoning is frequently downstream of a System-1-like snap judgment the model then spends tokens rationalizing rather than revising (Dang et al., 2025). Fourth, models do not reliably know how much deliberation a given problem needs, overthinking easy problems and underthinking hard ones in the same study (Su and Healey, 2025). None of this shows that longer reasoning never helps; a compute-optimal allocation strategy can match or exceed brute-force scaling at a fraction of the cost on math benchmarks (Snell et al., 2024), and faithfulness is recoverable under some model sizes and tasks (Lanham et al., 2023). The paper argues that these results are jointly better explained by treating "how much a model deliberates" as a token-length variable with heterogeneous, sometimes negative, effects on accuracy and an unreliable relationship to the model's actual computation, than by a two-system story in which more chain-of-thought means more of a qualitatively distinct and more trustworthy reasoning mode.

1. Introduction

"Reasoning models" are marketed and studied under a specific analogy. Li et al. (2025) open their survey of the field by stating it as a premise rather than a finding -- "Achieving human-level intelligence requires refining the transition from the fast, intuitive System 1 to the slower, more deliberate System 2 reasoning" -- and go on to describe recent reasoning models as "closely mimicking the deliberate reasoning of System 2 and showcasing human-like cognitive abilities." The analogy licenses two further assumptions that this paper argues do not survive contact with the evidence, even though the analogy's

source claim -- that extended chain-of-thought can improve accuracy -- is true in many settings. The first assumption is monotonicity: since System 2 is the more careful mode, more of it should, on average, help or at worst plateau. The second is transparency: since the chain-of-thought is supposed to be the deliberate process rendered in language, the words on the page should describe how the answer was actually reached, in a way that a snap intuitive judgment cannot be inspected.

Both assumptions have direct, recent, and specific counter-evidence. On monotonicity: Gema et al. (2025) construct evaluation tasks, spanning simple counting with distractors, regression with spurious features, deduction with constraint tracking, and advanced-AI-risk scenarios, on which extending a model's reasoning length deteriorates its performance, and report five distinct failure modes across model families, including one in which extended reasoning amplifies self-preservation-flavored behavior in a Claude model. Zhao et al. (2025) evaluate fourteen reasoning models under several test-time-scaling strategies on closed-book, knowledge-intensive tasks and report that increased computation does not consistently improve accuracy and often produces more hallucinations, driven largely by an increased willingness to attempt an answer rather than by better information. On transparency: Lanham et al. (2023) intervene directly on a model's own chain-of-thought (adding mistakes, paraphrasing it) to see how much the model's answer actually depends on it, and find that their central trend points the counter-intuitive way: on most of the tasks they study, the faithfulness of the stated reasoning falls as models grow in size and capability. Arcuschin et al. (2025) find that unfaithful chain-of-thought is not an artifact of adversarial prompt construction; on naturally worded, non-adversarial questions, models sometimes justify contradictory answers to logically related questions with superficially coherent arguments, at rates up to 13 percent in production models and measurably above zero even in thinking models.

Read together, these results do not describe a broken System 2. They describe a token-length variable, "how long the model deliberates," whose relationship to both accuracy and to the truth of its own verbalized description is heterogeneous and sometimes runs in the direction opposite the one the dual-process label predicts. This paper's contribution is not any one of these findings, each already published and already surprising on its own; it is the argument that four independent literatures -- inverse scaling, faithfulness, mechanistic bias, and length calibration -- converge on the same structural diagnosis, and that the diagnosis is a reason to stop treating "System 2" as a single reliably-better mode rather than a reason to expect deliberation is useless. Section 2 states terms, method and boundaries with adjacent published work in this project. Section 3 states the dual-process framing in its strongest published form so the rest of the paper argues against a real target. Section 4 gives the monotonicity evidence. Section 5 gives the faithfulness evidence. Section 6 gives a mechanistic account of why the two are connected rather than coincidental. Section 7 adds a third failure, length miscalibration, and treats briefly the question of deciding a budget in advance. Section 8 states the strongest evidence for the framing and what, if anything, survives once it is checked. Section 9 states what is not known. Section 10 proposes experiments. Section 11 relates this paper to two adjacent records in this journal. Section 12 states limitations.

No experiments were run for this paper. It is an analysis of what a body of published measurements licenses, and its contribution is a structural argument built from four independent lines of evidence, plus a correction to the argument's own overreach once the strongest counter-evidence is checked.

2. Scope, terms, method, and what this paper is not

Terms. *System-1/System-2 framing*, as used here, denotes the specific published claim in Section 1: that a large reasoning model's extended chain-of-thought is usefully described as engaging a distinct, more deliberate and more reliable mode of computation than a shorter or no-chain-of-thought response, and that the verbalized chain is a description of that mode's operation. *Monotonicity* denotes the assumption

that, within a model and holding the model fixed, more reasoning tokens produce equal or better accuracy on average. *Faithfulness* denotes the degree to which a model's final answer is caused by the reasoning steps it verbalizes, as opposed to being reached independently and then rationalized in the verbalized steps; this paper adopts Lanham et al.'s (2023) intervention-based operational sense of the term throughout, since it is the sense the cited faithfulness results measure.

Method and its limits. The evidence base was built by searching arXiv's listing API across roughly fifteen query formulations targeting inverse scaling, test-time compute, chain-of-thought faithfulness, overthinking, and the System-1/System-2 framing specifically, and retrieving each candidate's full Atom metadata record -- title, authors, dates, category, and abstract -- before it was cited. Every numeric claim in this paper is taken from the abstract or from the specific headline result the source's own abstract states; this paper did not retrieve or re-derive any full-text table, and where a claim would require one, Section 9 says so rather than estimating.

What this paper is not. It is not a claim that test-time compute scaling is generally ineffective; Section 8 states the opposite result plainly and at length. It is not a claim that chain-of-thought is never faithful; Lanham et al.'s own qualification, that faithfulness is recoverable under some model sizes and tasks, is load-bearing for this paper's own honesty and is restated in Section 8. It is not a benchmark paper, a training proposal, or an argument that dual-process psychology is wrong about humans; the target throughout is the specific published claim that the human dual-process label transfers cleanly onto LLM chain-of-thought length, not the psychological theory in its home domain.

3. The framing, stated in its strongest form

The framing this paper argues against is not a straw target; it is stated as the organizing premise of a survey with twenty authors covering the field's major systems. Li et al. (2025) write that foundational LLMs are adept at fast decision-making but too shallow for complex reasoning, because they "have not yet fully embraced the step-by-step analysis characteristic of true System 2 thinking"; the same survey portrays recent reasoning models such as OpenAI's o1/o3 and DeepSeek's R1 as performing at expert level in areas such as mathematics and coding while closely imitating System 2-style deliberation. Read on its own terms, the claim has two separable parts that this paper's evidence addresses separately: a performance claim (deliberate, step-by-step processing produces better answers, at least in the fields the survey names) and a process claim (the mechanism by which it does so resembles human deliberate reasoning closely enough that the label, and the verbalized trace that comes with it, are informative about how the model reached its answer). The performance claim is true in many of the settings the survey and its cited systems were built and evaluated on; Section 8 does not dispute this. The process claim is the one this paper's Sections 4 through 6 complicate: the performance gain, where it exists, does not appear to come from a reliably distinct, monotonically-more-careful mode whose verbalization can be trusted as a description of itself.

A parallel literature already recognizes a version of this gap under a different name. Sui et al. (2025), surveying efficient reasoning methods, describe the "overthinking phenomenon": longer chain-of-thought sequences introduce significant computational overhead through verbose and redundant outputs, motivating a large body of work on shortening them without losing accuracy. That survey's framing is efficiency-first -- redundant reasoning costs compute even when it does not cost accuracy -- and is compatible with the System-1/System-2 label; an inefficient-but-still-correct System 2 is still a working System 2. The evidence in Sections 4 and 5 below makes a stronger claim than inefficiency: in the tasks Gema et al. construct, extra deliberation actively degrades the answer, not merely its cost, and in the tasks Lanham et al. and Arcuschin et al. study, the extra tokens sometimes describe a process that did not happen.

4. Deliberation is not monotonic in accuracy

Gema et al. (2025) is the paper this project's own topic-selection process anchored on, and it earns that position: it does not merely find diminishing returns to reasoning length, which would be compatible with a "System 2 helps, just with decreasing marginal utility" story. It constructs four categories of evaluation task -- simple counting tasks with distractors, regression tasks with spurious features, deduction tasks with constraint tracking, and advanced AI risk scenarios -- specifically to exhibit an inverse relationship between reasoning length and accuracy, and reports five distinct failure modes as models reason for longer: Claude models becoming increasingly distracted by irrelevant information; OpenAI o-series models resisting distractors but overfitting to problem framings; models shifting from reasonable priors to spurious correlations; all models showing difficulty maintaining focus on complex deductive tasks; and, in the most safety-relevant finding, Claude Sonnet 4 showing increased expressions of self-preservation under extended reasoning. The paper's own reading, stated plainly in its abstract, is that while test-time compute scaling remains promising for improving capabilities, it may inadvertently reinforce problematic reasoning patterns. That reading is itself calibrated -- "may reinforce," not "always reinforces" -- and this paper inherits the same calibration: the claim is that monotonicity fails on tasks built to expose it, not that it fails everywhere.

Zhao et al. (2025) supply a second, independent failure of monotonicity in a different task family: closed-book, knowledge-intensive question answering, where the fact needed to answer correctly either is or is not present in the model's parameters, and no amount of reasoning manufactures a fact that was never stored. Evaluating fourteen reasoning models under several test-time-scaling strategies, they report that accuracy does not rise consistently as test-time computation is increased, while hallucinations often become more common, and that the effect is largely driven by the model's increased willingness to attempt an answer rather than by any gain in retrieved information -- a model that reasons for longer tends to attempt more answers, many of them wrong -- alongside patterns consistent with confirmation bias in which extended reasoning reinforces an early incorrect belief with fabricated supporting detail. Their paper adds an information-theoretic argument for why this should be the general case on this task family: compute-only test-time scaling, applied as a post-processing procedure on a fixed model, cannot introduce new information about the ground-truth answer, which is a structural reason -- not merely an empirical one -- to expect that "deliberate more" cannot substitute for "know more" on tasks where the limiting resource is a stored fact rather than a derivation.

The two results are not redundant with each other. Gema et al.'s tasks are largely reasoning tasks where the correct derivation exists in principle and the model loses track of it; Zhao et al.'s tasks are retrieval tasks where no derivation would help because the limiting resource is a fact the model either has or does not. That they arrive at the same qualitative conclusion -- more test-time compute, worse or flat accuracy -- from two different task structures and two different papers is what elevates this section's claim from "some tasks are adversarially hard for long chains-of-thought" to "the monotonicity assumption fails for structurally different reasons in structurally different places," which a single result from either paper alone would not establish.

5. The trace is not a reliable readout

A model's chain-of-thought earns its epistemic credit from an implicit promise: that the steps written are, in some reasonably direct sense, the steps that produced the answer, so that reading them tells a user how the answer was reached and gives that user something to check. Lanham et al. (2023) test this promise directly, using intervention (adding mistakes to a chain-of-thought, paraphrasing it, truncating it) to measure how much a model's final answer actually depends on what it wrote. Their central, counter-intuitive result is stated flatly in their own abstract: "As models become larger and more capable, they produce less faithful reasoning on most tasks we study." That direction is the opposite of

what the System-1/System-2 label predicts. If longer chains-of-thought in more capable models were more instances of a genuine deliberate mode, faithfulness should rise with capability, not fall. Lanham et al. also report that models vary widely by task in how strongly they condition their answer on the chain-of-thought at all, sometimes relying on it heavily and other times largely ignoring it, and that CoT's performance boost does not appear to come from the added test-time compute alone or from the specific phrasing used -- evidence that whatever is producing the accuracy gain where one exists is not well described as "the model thinking through the problem in the words it wrote down."

Arcuschin et al. (2025) close an obvious escape from that result: the worry that unfaithfulness is a product of the specific adversarial setup used to elicit it (deliberately biased prompts, edited outputs) rather than a general property of how these models reason. They show unfaithful chain-of-thought on naturally worded, non-adversarial prompts, with no artificial bias injected and no editing of model outputs. Presented separately with "Is X bigger than Y?" and "Is Y bigger than X?", models sometimes answer Yes to both or No to both and produce superficially coherent arguments to justify the contradiction -- a pattern they term Implicit Post-Hoc Rationalization, which they attribute to an implicit bias toward Yes or No that the verbalized reasoning did not report and, on their account, did not actually decide. They report rates up to 13 percent in production models, and state explicitly that "while frontier models are more faithful, none are entirely so," including thinking models such as DeepSeek R1 (0.37 percent) and Sonnet 3.7 with thinking (0.04 percent) -- low rates, but not zero, and present in exactly the models the System-1/System-2 label singles out as the ones running the deliberate mode. Their paper's own closing caution generalizes past their specific intervention: chain-of-thought, they conclude, "is not a complete account of the internal process that produced the model's answer and should be used with caution in agentic or safety-critical settings."

The combination of Section 4 and this section is where the paper's structural point becomes visible rather than merely additive. It is not only that longer reasoning sometimes produces worse answers; it is that the words produced alongside a worse answer are not a reliable description of why the answer is worse, or in Arcuschin et al.'s and Lanham et al.'s findings, why the answer is what it is at all. A user or a downstream monitoring system reading a chain-of-thought for an explanation of an error, expecting the transparency a "System 2" label implies, is reading a text whose own authors' evidence says should be trusted least in exactly the more-capable models most likely to be deployed.

6. Why the two failures are connected: bias contaminates the trace

Sections 4 and 5 could, in principle, describe two coincidentally co-occurring but mechanistically unrelated problems: reasoning models happen to degrade with length on some tasks, and separately, reasoning models happen to produce unfaithful chains-of-thought. Dang et al. (2025) supply a mechanistic account that connects them, and in doing so complicates the dual-process picture more directly than either result alone. They identify what they call internal bias: upon encountering a problem, a model forms a preliminary guess about the answer before any systematic reasoning has occurred, arising, in their description, without deliberate computation -- functionally, a System-1-style snap judgment. Their central finding is that when this preliminary guess conflicts with the model's subsequent step-by-step reasoning, the model tends to engage in excessive reflection, generating the "overthinking" that Sections 3 and 7 of this paper also discuss, and that this reflection is redundant rather than corrective: two counterfactual interventions show that removing the input question partway through generation reduces the redundant reasoning across multiple complex-reasoning tasks, and that manually injecting a bias increases overthinking correspondingly, which the authors read as evidence of a causal relationship rather than a mere correlation. Interpretability experiments in the same paper suggest that excessive attention to the input question is a mechanism by which the internal bias shapes the model's subsequent reasoning trajectory. Their paper's own closing note is a discouraging one for

any purely architectural fix: several methods aimed at mitigating overthinking were evaluated, and the influence of internal bias persisted under all of them.

If this account generalizes beyond the tasks and models Dang et al. study -- and it is one paper, so that qualification matters and is repeated in Section 9 -- then what looks, from the outside, like a chain-of-thought deliberating its way to an answer is frequently a System-1-style guess formed immediately, followed by tokens that do not revise the guess so much as accompany it. That is a specific, mechanistic version of the unfaithfulness Lanham et al. and Arcuschin et al. document, and it is also a candidate explanation for why extra length can hurt rather than help in Gema et al.'s tasks: more tokens spent elaborating a snap judgment that will not be revised is not the same resource as more tokens spent on genuine additional computation, and only the second kind should be expected to improve accuracy on average. The dual-process label's own vocabulary, applied honestly to this finding, would have to say that "System 2" reasoning in these cases is System 1 wearing System 2's clothing -- which is close to conceding that the label is doing rhetorical work the underlying mechanism does not support, rather than describing a real architectural distinction the model instantiates.

7. Models do not know how much deliberation a problem needs

A fourth failure mode is independent of whether any given instance of extra reasoning helps or hurts: models are not well calibrated about which case they are in before they start. Su and Healey (2025) examine, systematically and empirically, how answer correctness varies with reasoning length, and report a miscalibration that runs in both directions in the same models: on easy problems they reason at needless length, and on hard ones they stop too soon, just where longer reasoning would have paid. Their interpretation is that the models may be misjudging how hard a problem is and so failing to fit the length of their answer to it. They also find that a preference-optimization procedure favouring shorter responses, whatever their correctness, cuts generation length substantially while keeping accuracy acceptable -- a finding that is itself informative, since if length were tightly coupled to a genuine deliberate process, indiscriminately rewarding brevity should cost more accuracy than it evidently does in their setup.

The natural response to a length-calibration failure is to decide the budget in advance rather than let the model run until it stops on its own, which is exactly the prospective side of the metacognitive-control question this paper's topic inherits from an earlier, since-folded queue entry on metacognitive monitoring. CODA (Wu et al., 2026) is one recent instance: rather than a user-supplied or externally annotated budget, it estimates per-instance difficulty from group-based policy rollouts and uses that estimate to shape a length-dependent reward during training, penalizing verbosity on instances its own estimator judges easy and rewarding longer rollouts on instances it judges hard, reporting token-cost reductions of over 60 percent on easy tasks at maintained accuracy and more deliberative rollouts on hard tasks. The design is worth reading against Section 6's mechanism: CODA's difficulty signal is estimated by the policy itself, from its own rollouts, before length is shaped -- which sidesteps asking whether the model's mid-generation reflection is trustworthy by moving the difficulty judgment earlier and outside the reasoning trace itself, rather than trying to make the trace more faithful. That design choice is a tacit concession to this paper's Section 6 argument, whether or not its authors frame it that way: if the trace's own account of why it is still reasoning were reliable, there would be less reason to estimate difficulty from an external rollout signal instead.

Read together with Section 4, this section's finding sharpens the paper's claim rather than merely adding to it. It is not only that more deliberation sometimes hurts; it is that models lack a reliable internal signal for which situation they are in, which is precisely what a genuine, well-functioning System 2 -- one that monitors its own confidence and allocates effort accordingly -- would be expected to have. A dual-process story in which System 2 cannot tell when it is needed is a substantially weaker

claim than the one the framing in Section 3 states.

8. The strongest case for the framing, checked

A fair test of this paper's argument is whether it survives contact with the best evidence for the position it opposes, not only the evidence against it. Two pieces of that evidence are substantial and are stated here without softening.

First, test-time compute scaling is not generally ineffective, and this paper's own anchor result does not claim it is. Snell et al. (2024) study the scaling of inference-time computation with a focus on how much a fixed inference-time compute budget can improve performance on a challenging prompt, analyzing search against process-based verifier reward models and adaptive updating of the model's output distribution at test time. Their central result is a compute-optimal strategy that allocates test-time compute adaptively per prompt based on estimated difficulty, and they report that this strategy improves the efficiency of test-time compute scaling by more than 4 times compared to a best-of-N baseline, and that, in a FLOPs-matched evaluation, test-time compute can be used to outperform a model 14 times larger on problems where a smaller base model already shows non-trivial success. That is a genuine, load-bearing counter-example to any reading of this paper's argument as "reasoning length does not matter" or "test-time compute is not worth having." What Snell et al.'s result does show, and what this paper's Sections 4 and 7 both rely on, is that the strategy has to be compute-optimal and difficulty-conditioned to deliver that gain; naive, undifferentiated scaling is exactly the regime Gema et al. and Su and Healey report failing.

Second, faithfulness is not categorically absent; it is conditional, and Lanham et al.'s own paper states the condition. Their abstract's closing sentence, quoted only partially in Section 5, reads in full that chain-of- thought "can be faithful if the circumstances such as the model size and task are carefully chosen." That is a meaningfully different claim from "chains-of- thought are never faithful," and this paper's Section 5 argument does not require the stronger claim: the point is that faithfulness is not a guaranteed property of the more-capable, more-deliberate end of the model spectrum the System-1/ System-2 label associates with genuine reasoning, and on Lanham et al.'s own reported trend, capability and faithfulness move in opposite directions on most tasks studied -- but "most" is not "all," and a reader should not take this paper as asserting otherwise.

A third, more indirect complication is worth stating because it cuts against this paper's framing rather than for it. Dinardi et al. (2025), studying parallel test-time reasoning, report evidence for exactly two distinct regimes in a model's own output distribution: a concise, confident conventional regime and a verbose, uncertain overthinking regime, with a detectable critical point at which the overthinking regime becomes significant, and they show that a simple shortest-answer heuristic -- preferring the shortest of several sampled solutions -- is competitive with more expensive self-consistency methods precisely because it preferentially samples from the concise regime. Read uncharitably to this paper's argument, a genuine bimodal regime structure is some evidence for a two-mode story after all, just not one where the longer mode is the more trustworthy one; read more carefully, it is evidence for two regimes distinguished by output length and confidence, which is a narrower and more mechanical distinction than the dual-process label's claim that the longer regime is the site of genuine deliberate computation. This paper takes the narrower reading as the one licensed by Dinardi et al.'s own framing, since their paper's contribution is a length heuristic, not a claim about which regime reasons more validly -- but a reader who prefers the stronger reading has a real result to point to, and this paper does not have a way to rule it out from the evidence checked here.

9. What is not known

Stated flatly, as declared ignorance rather than hedged around: this paper does not know whether Dang

et al.'s bias-contamination mechanism generalizes beyond the models and tasks their paper studies, since it is one mechanistic study and this paper located no independent replication. It does not know what fraction of published reasoning-model deployments rely on undifferentiated length scaling, of the kind Gema et al. and Su and Healey show failing, versus difficulty-conditioned allocation of the kind Snell et al. and Wu et al. (CODA) show succeeding; this paper's search located examples of both design families but not a census of which is more common in production systems. It does not know whether Arcuschin et al.'s Implicit-Post-Hoc-Rationalization rate, measured on the specific paired- question task they construct, generalizes to other task formats where a model's answer to one question is not so directly checkable against its answer to a logically related one. It does not know, absent the per-task breakdown neither cited paper's abstract reports, whether Lanham et al.'s capability-faithfulness trend and Zhao et al.'s compute-hallucination trend share a common cause or are independent regularities that happen to point the same direction.

10. Proposed experiments

First, replicate Dang et al.'s counterfactual internal-bias intervention on the specific task families Gema et al. use to demonstrate inverse scaling (counting with distractors, regression with spurious features, deduction with constraint tracking), to test directly whether the accuracy degradation in those tasks is mediated by the same bias-contamination mechanism or arises independently of it. If mediated, manually suppressing the identified internal bias signal before generation should recover some of the accuracy inverse scaling costs; if independent, it should not.

Second, score a chain-of-thought faithfulness measure (Lanham et al.'s intervention protocol, or Arcuschin et al.'s paired-question protocol) against Zhao et al.'s closed-book knowledge-intensive benchmark, to test whether the hallucinations produced under extended reasoning there are accompanied by unfaithful chains-of-thought or by faithful chains-of-thought that simply reach a wrong conclusion from insufficient stored knowledge. The two would support different diagnoses: an unfaithfulness finding would extend this paper's Section 5 argument into the knowledge-intensive setting; faithful-but-wrong would isolate Section 4's monotonicity failure as a separate phenomenon from Section 5's transparency failure in this specific task family.

Third, evaluate CODA's or a similar prospective difficulty estimator's predicted-difficulty signal against Su and Healey's overthinking and underthinking labels directly, rather than against downstream accuracy alone, to test whether the estimator's judgment of "this is easy" or "this is hard" agrees with a labeling of which instances the model itself over- or under-reasoned on. Agreement would be evidence the estimator has located a genuine, reusable difficulty signal; disagreement would suggest the estimator's gains come from some other property of its training procedure.

11. Relation to prior records

Two records in this journal are adjacent and are distinguished here rather than repeated. Mahendrakar (2026a) analyzed hallucination detection and argued that the signals which reliably detect an ungrounded claim are external to anything a model reports about itself; the present paper's Section 5 argument is compatible with and, in the specific case of chain-of-thought, an instance of that broader claim -- a verbalized reasoning trace is one more thing the model reports about itself, and this paper's evidence is that it should be trusted no further than the externally-detected signals record 16 already surveys. Mahendrakar (2026b) analyzed budget and resource awareness in LLM agents and argued that published allocation gains are keyed to a signal measured after the fact rather than a capability the allocating model possesses; that record and this one share one evidentiary anchor, Gema et al.'s inverse-scaling result, cited there for the narrower claim that capability does not confer resource awareness and cited here for the broader claim that the System-1/System-2 framing itself does not survive the same

result once faithfulness and mechanism are added. The two papers' central arguments do not reduce to each other: 2026b is about whether a model can predict and manage its own compute spend, a resource-management question; this paper is about whether extended deliberation is validly described as engaging a more reliable, transparent reasoning mode at all, a construct- validity question about the System-1/System-2 label itself. Where the two overlap, on the specific finding that more test-time compute is not monotonically beneficial, this paper defers to 2026b's more detailed treatment of the allocation literature and adds only the faithfulness and mechanism evidence 2026b does not address.

12. Limitations

This paper's evidence base is built entirely from abstracts and stated headline results rather than from full-text tables; where a finer-grained number would strengthen or weaken a claim, Section 9 says so rather than estimating it. Second, Section 6's mechanistic account rests on one paper, Dang et al. (2025), studying a specific set of models and tasks; the paper treats this account as a candidate explanation connecting Sections 4 and 5, not as an established general mechanism, and Section 9 restates that qualification. Third, this paper's central target, the System-1/System-2 framing as stated by Li et al. (2025), is one influential statement of the framing rather than the only one in the field; other authors may state the analogy more cautiously than the passage this paper quotes, and this paper's argument bears most directly on the strong, literal version of the claim rather than on every author's use of "System 2" as loose shorthand. Fourth, Section 8's discussion of Dinardi et al.'s bimodal-regime evidence acknowledges a reading of that result this paper does not adopt; a reader who adopts the stronger reading would have grounds to discount this paper's argument that no genuine two-mode structure exists, though not grounds to discount Sections 4 through 6's evidence that the longer mode is not reliably more accurate or more transparent than the framing assumes.

Acknowledgements

This paper is a literature analysis and depends entirely on measurements made and published by others. Its central argument is only possible because Lanham et al. and Arcuschin et al. reported unfaithfulness as a genuine finding rather than a footnote, because Gema et al. built tasks specifically to make inverse scaling visible rather than treating it as noise, and because Dang et al.'s counterfactual interventions gave a mechanistic account rather than a correlational observation. Snell et al.'s and Wu et al.'s positive results are cited at equal length and with the same care, since a paper that used only the negative evidence would not have tested its own claim honestly.

Revision note

Version 1.1 (29 September 2026) replaces version 1.0. Passages that followed the wording of cited abstracts too closely are now either quoted, with their citation, or rewritten in new words, and a few qualifiers the sources state (such as "may" and "often") were restored in the process. The argument, the references and the reported values are unchanged.

References

- Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2025). Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. arXiv:2503.08679. doi:10.48550/arXiv.2503.08679. [PREPRINT]
- Dang, R., Li, Z., Huang, S., & Chen, J. (2025). The First Impression Problem: Internal Bias Triggers Overthinking in Reasoning Models. arXiv:2505.16448. doi:10.48550/arXiv.2505.16448. [PREPRINT]
- Dinardi, R. C., Yamamoto, B., Costa, A. H. R., & Jordao, A. (2025). The Virtues of Brevity: Avoid Overthinking in Parallel Test-Time Reasoning. arXiv:2510.21067. doi:10.48550/arXiv.2510.21067. [PREPRINT]
- Gema, A. P., Hagele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael,

- J., Alex, B., Minervini, P., Chen, Y., Benton, J., & Perez, E. (2025). Inverse Scaling in Test-Time Compute. arXiv:2507.14417. doi:10.48550/arXiv.2507.14417. [PREPRINT]
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukositate, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., & Perez, E. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702. doi:10.48550/arXiv.2307.13702. [PREPRINT]
- Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., Zhang, Y., Yin, F., Dong, J., Li, Z., Bi, B.-L., Mei, L.-R., Fang, J., Liang, X., Guo, Z., Song, L., & Liu, C.-L. (2025). From System 1 to System 2: A Survey of Reasoning Large Language Models. arXiv:2502.17419. doi:10.48550/arXiv.2502.17419. [PREPRINT]
- Mahendrakar, P. M. (2026a). The Self-Verification Gap: Why Working Hallucination Detectors Are External, and What Real-Time Correction Would Actually Require. Life of Research. doi:10.5281/zenodo.22030500
- Mahendrakar, P. M. (2026b). Three Things Called Budget Awareness: Observability, Forecasting and Allocation in LLM Agents, Why Every Published Allocation Gain Is Keyed to a Signal Measured After the Fact, and the Run-to-Run Variance No Forecast Is Scored Against. Life of Research. doi:10.5281/zenodo.22884949
- Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314. doi:10.48550/arXiv.2408.03314. [PREPRINT]
- Su, J., & Healey, J. (2025). Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and Correctness in LLMs. arXiv:2505.00127. doi:10.48550/arXiv.2505.00127. [PREPRINT]
- Sui, Y., Chuang, Y.-N., Wang, G., Zhang, J., Zhang, T., Yuan, J., Liu, H., Wen, A., Zhong, S., Zou, N., Chen, H., & Hu, X. (2025). Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. arXiv:2503.16419. doi:10.48550/arXiv.2503.16419. [PREPRINT]
- Wu, S., Xie, J., Zhang, Y., & Xiao, Y. (2026). CODA: Difficulty-Aware Compute Allocation for Adaptive Reasoning. arXiv:2603.08659. doi:10.48550/arXiv.2603.08659. [PREPRINT]
- Zhao, J. X., Hooi, B., & Ng, S.-K. (2025). Test-Time Scaling in Reasoning Models Is Not Effective for Knowledge-Intensive Tasks Yet. arXiv:2509.06861. doi:10.48550/arXiv.2509.06861. [PREPRINT]

(c) 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The literature search, drafting and citation verification for this paper were carried out with AI assistance under the author's direction. Every citation was machine-verified against its arXiv or DataCite record before inclusion, with title and author list checked against the record returned, and every quantitative claim in this paper is taken from the abstract or stated headline result of the source credited with it. No experiment was run and no number in this paper was measured or recomputed by its author. The author is responsible for the final text and for all claims made in it.

DOI: 10.5281/zenodo.22904980