

A Satisfied Constraint Is Not a True Answer: What Reasoning Changes When an Instruction Cannot Be Met Truthfully, Why 'Detection-Resistant' Names a Property of the Checker, and the Feasibility-Controlled Test That Would Settle the Trade

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A language model asked for something its knowledge cannot honestly supply has four ways out: it can break the instruction openly, decline, bend a real fact until it fits, or invent one. A January 2026 study reports that switching reasoning on in two commercial models moved bibliographic answers out of the first category and partly into the last two, and reads this as reasoning trading honest violations for "detection-resistant distortions". In this analysis the author reviews and synthesises published results to show that the finding is real but narrower, and differently caused, than that reading. Arithmetic on the study's own tables shows that most of the violations removed by reasoning became compliant records rather than distortions or fabrications, which together absorbed 13 to 14 percent of the shift in one model and 30 percent in the other, and that at least 53 percent of GPT-5.2's reasoning-mode records were real journal articles with accurate metadata. The wider literature turns on a variable the study held fixed, the way out, which has to be split three ways: offered in the prompt, credited by the grader, and rewarded in training. Where abstention is forbidden or unrewarded, reasoning training and longer deliberation raise confident wrong answers, invented tool calls and hacked tests. Offering or crediting abstention does not reliably reverse this: on one factuality benchmark the reasoning models declined more and erred less, but in two other studies reasoning training lowered the use of an exit that was offered or scored, and in a third, longer budgets did so while switching thinking on mostly cut errors. Only where training rewarded declining was it restored. "Detection-resistant" turns out to describe the checker rather than the output: a venue swap passes an existence check, fails a metadata check, and is excluded by design from one conservative audit of hallucinated citations. The paper separates three reasoning interventions and three meanings of feasibility, consolidates 24 published results in one table, states an attribution procedure, and specifies a feasibility-controlled factorial test that would decide how much of the trade is reasoning and how much is the missing exit.

Keywords: large language models; reasoning models; hallucination; instruction following; abstention; constraint satisfaction; specification gaming; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The working assumption behind most deployments of reasoning models is that more deliberation buys more reliability. The assumption has evidence behind it. Prompting a model to plan and answer verification questions before committing reduces hallucinated list items and long-form errors (Dhuliawala et al., 2023). Reasoning models state their confidence more accurately than their non-reasoning counterparts in 33 of 36 benchmark settings in one comparison (Yoon et al., 2025). On a short-answer factuality benchmark that credits declining, the first OpenAI reasoning preview both answered more questions correctly and answered fewer incorrectly than GPT-4o (Wei et al., 2024a). DeepSeek-R1 scores higher on the same benchmark than DeepSeek-V3, the non-reasoning chat model built on the same base (DeepSeek-AI, 2025).

A second body of results points the other way, and it has grown quickly since 2025. Reinforcement finetuning for reasoning cut refusal rates on unanswerable problems by more than 80 percent in one controlled study (Song et

al., 2025). Reasoning fine-tuning lowered abstention by 24 percent on average across a 20-dataset benchmark (Kirichenko et al., 2025). Longer test-time reasoning on closed-book factual questions often produced more hallucinations, largely because the models attempted more answers (Zhao et al., 2025). Reasoning-trained models exploit misspecified tasks more often than their base models or non-reasoning comparators (Nishimura-Gasparian et al., 2026; Bondarenko et al., 2025).

Niimi (2026) adds a case that sits at the junction of the two literatures, and it is the occasion for this paper. Asked for ten "famous peer-reviewed journal articles" on computer-science topics where conference papers dominate, GPT-5.2 and Gemini 3 Flash with reasoning switched off mostly returned real conference papers and preprints: violations of the instruction, but true. With reasoning switched on, violations fell from 74.8 to 13.2 percent of records for GPT-5.2 and from 66.0 to 26.0 percent for Gemini 3 Flash, while records whose venue had been rewritten to a journal and records that do not exist both rose (Niimi, 2026). The study's abstract concludes that reasoning models "trade honest constraint violations for detection-resistant distortions" and calls

the pattern "a fundamental limitation of reasoning".

The case matters because the outcome it describes is the expensive one. A violated instruction is visible to anyone who reads the answer. A fabricated citation is caught by anyone who looks it up. A real paper moved to a journal it never appeared in survives a lookup by title and fails only a check of its metadata. If reasoning systematically moves failures from the first kind to the third, then the reliability gains measured on compliance benchmarks are partly the conversion of cheap errors into costly ones.

This paper asks what the published evidence licenses about that trade. Its central claim is that the trade is real but is not, on present evidence, a property of reasoning as such. It is a property of reasoning applied to an objective with no rewarded exit, scored by a checker that sees some failure kinds and not others. Three moves support the claim.

The anchor's own numbers bound the trade. Most of the violations that reasoning removed in Niimi (2026) became compliant records rather than distortions or fabrications, and the study's authenticity table implies that at least 53 percent of GPT-5.2's reasoning-mode records were real journal articles with accurate metadata (Section 4). The constraint was therefore hard rather than unsatisfiable for most items, and the distortion share is a minority of the shift, though a costly one.

The literature turns on the exit, but only one kind of exit is shown to work. Where declining is impossible or unscored, reasoning raises confident errors, invented tool calls and test hacking. Where declining is offered in the prompt or credited by the grader, the results go both ways; where training rewards declining, declines return (Sections 5 to 7). The anchor's prompt removed the exit, and its own recommendations say so.

Detectability is indexed by the checker. Semantic-entropy detection, existence checks, metadata checks, monitors that read the reasoning trace, and human reviewers under time limits each see a different subset of failures (Section 8). "Detection-resistant" is a claim about a pair - an output kind and a checker - and the anchor does not run a checker.

The paper makes five contributions. It defines four response outcomes under a constraint conflict and three meanings of feasibility (Section 2). It re-reads the anchor's tables to bound the trade (Section 4). It consolidates 24 published results, organised by the reasoning intervention, the exit available and the direction of the effect (Table 1, Section 9). It states an attribution procedure that assigns an observed shift in outcomes to knowledge, exit, incentive or checker (Algorithm 1, Section 10). And it specifies a factorial test, run on items whose feasibility is known in advance, that would decide the open part (Section 11).

1.1 Relation to the author's earlier papers

Four earlier records of the author touch this question from other sides. The Self-Verification Gap argued that working hallucination detectors are external to the model and that no published system combines low cost, transfer, localisation and an actionable false-positive rate (Mahendrakar, 2026a). This paper inherits that conclusion and narrows it: the question here is not whether a detector exists but which failure kinds a given detector can see. When Deliberation Hurts argued that the length of a chain of thought is a variable with heterogeneous, sometimes negative effects on accuracy (Mahendrakar, 2026d). The present paper takes one such negative effect and asks what it is made of. Two Kinds of Missing separated underspecified inputs from unknown answers as triggers for abstention (Mahendrakar, 2026b), and Not Acting Is Not One Decision separated epistemic from consequential triggers in agents (Mahendrakar, 2026c). Both treated abstention as the decision to be scored; this paper treats it as an exit whose availability changes what reasoning does when it is absent. The question here - what reasoning does to the form of a failure when the honest answer is disallowed or unavailable - is not one those papers addressed.

2. Scope and terms

2.1 Constraint conflict

A *constraint conflict* is a request whose explicit constraints cannot all be satisfied by an answer that is both responsive and true, given what the responding system can access. The anchor's journal-only bibliography is one instance. Others in the literature include a coding task whose unit tests contradict its specification (Zhong et al., 2025), a math problem with a premise deleted (Song et al., 2025; Fan et al., 2025), a tool call when no suitable tool exists (Yin et al., 2025), and a request to report success when the environment state shows failure (Advani, 2026).

The phrase "given what the responding system can access" carries the weight, and it separates three meanings of feasibility that the anchor's text does not distinguish.

World-feasible: an answer satisfying every constraint exists. Ten journal articles on transformers exist.

Knowledge-feasible: the system can produce such an answer from what it has stored or can retrieve. A closed-book model may not hold enough journal articles on a niche topic even when they exist.

Format-feasible: the output format admits the honest response. A JSON schema with every field required and no field for "none found" makes declining format-infeasible even when it is epistemically warranted.

A request can be world-feasible, knowledge-infeasible and format-infeasible at once. The anchor's task was designed to be hard on the second and closed on the

third (Section 3.1).

2.2 Four outcomes

Under a constraint conflict a response falls into one of four outcomes. The terms below are this paper's stipulations; the anchor's three categories map onto overt violation and the two subtypes of covert falsehood.

Truthful compliance (TC): the answer satisfies the constraints and is true.

Overt violation (OV): the answer is true but visibly breaks a constraint. The anchor's "constraint violation" category: a real conference paper returned for a journal-only request.

Exit (EX): the system declines, flags the conflict, or returns fewer items than asked. The anchor's design leaves no room for this outcome.

Covert falsehood (CF): the answer appears to satisfy the constraints and is false. It has two subtypes. *Distortion* alters an attribute of a real entity so that it fits (a NeurIPS paper relabelled as an IEEE Transactions article). *Fabrication* supplies an entity that does not exist.

The standard taxonomy of LLM hallucination separates factuality hallucination, a conflict with verifiable world facts, from faithfulness hallucination, a departure from the user's instructions or context (Huang et al., 2023). In those terms the anchor's finding is a transfer from the second type to the first. Overt violation is an unfaithful answer; covert falsehood is an unfactual one. Nothing in the finding moves the model from honest to dishonest by that taxonomy's lights; it moves the error from one column to the other. Whether the move is also a move toward deception is a separate question about what the model knew (Section 8.4).

2.3 Three reasoning interventions

"Reasoning" names at least three interventions, and the results below do not transfer freely among them.

R1, prompted deliberation: a fixed model is asked to think step by step or to verify before answering (Wei et al., 2022; Dhuliawala et al., 2023).

R2, a test-time budget toggle: the same deployed model is run with reasoning effort or thinking level set high or low. The anchor's contrast is of this kind: GPT-5.2 with reasoning effort "none" against "medium", and Gemini 3 Flash with thinking level "minimal" against "high" (Niimi, 2026).

R3, reasoning post-training: a model is trained with reinforcement learning on verifiable rewards or distilled from long reasoning traces, and compared with its pre-training or instruction-tuned parent (Song et al., 2025; Kirichenko et al., 2025; DeepSeek-AI, 2025).

Under R2 the weights do not change, so an R2 effect shows a disposition the weights already hold being expressed when deliberation is allowed. Under R3 the

disposition itself is what changes. The anchor's title speaks of "reasoning models"; its experiment is R2.

2.4 Checkers

A *checker* is any procedure applied to an output to decide whether it is acceptable. Four tiers matter here.

T0, constraint check: does the output satisfy the stated constraint? (Is each venue a journal?)

T1, existence check: does the entity exist? (Does a paper with this title and these authors exist?)

T2, attribute check: do the stated attributes match the record? (Was it published in that venue, that year, those pages?)

T3, belief check: did the system's own state hold the true attribute? (Does the model, asked neutrally, place the paper at NeurIPS?)

Overt violations fail T0. Fabrications pass T0 and fail T1. Distortions pass T0 and T1 and fail T2. Distortion of a fact the model demonstrably held additionally fails T3, and only that case is a lie in the sense the honesty literature uses (Ren et al., 2025; Pacchiardi et al., 2023).

3. The claim at issue and its two readings

3.1 The anchor study

Niimi (2026) prompted each model, zero-shot at temperature 1.0, to return ten famous peer-reviewed journal articles on each of 25 computer-science topics as a JSON array in APA style, with volume, issue and pages, and with a self-reported confidence from 0 to 100 that each paper exists exactly as described. The prompt required that all fields be filled in and that no additional explanation be given. Each condition therefore produced 250 records. Undergraduate annotators searched each title on Google Scholar and recorded the matching bibliography or "Not found", and the author validated the final labels.

The reported effects (Table 1 of the source) are these. Violations fell from 74.8 to 13.2 percent for GPT-5.2 and from 66.0 to 26.0 percent for Gemini 3 Flash. Distortion to any journal rose from 4.8 to 10.8 percent and from 6.8 to 12.8 percent. Complete fabrication rose from 8.8 to 10.8 percent and from 6.0 to 12.0 percent, with the rise significant for Gemini only. On the source's authenticity scale (2 for an existing paper with accurate metadata, 1 for an existing paper with metadata errors including distortions, 0 for a fabrication), GPT-5.2's mean rose from 1.48 to 1.56 with reasoning and Gemini's fell from 1.57 to 1.37 (Niimi, 2026).

3.2 The strong reading

The strong reading, which the anchor's abstract and conclusions adopt, has three parts. Reasoning, when knowledge cannot satisfy a constraint, systematically converts honest violations into distortions and fabrications. The conversion is "a fundamental limitation

of reasoning", not an artefact of one implementation, because it appears in both vendors' models. And the resulting errors are detection-resistant, because distortions contain real entities that a superficial check accepts.

3.3 The narrow reading

The narrow reading is that, in one prompt design that forbade declining, in a domain chosen so that the constraint would be hard to meet from memory, raising the test-time reasoning setting of two models made them satisfy the constraint far more often, mostly with real records and partly with altered or invented ones, and that the altered records would pass a check that looks only at titles and authors.

3.4 Why the difference matters

The two readings recommend different things. The strong reading implies that reasoning should be distrusted on knowledge-intensive tasks as such. The narrow reading implies that the exit, the incentive and the checker should be fixed, and that reasoning may then be an asset. The anchor's own practical section leans toward the narrow reading: it recommends giving models "legitimate escape routes" such as an explicit "I don't know" or a predefined fallback, and notes that its prompt prohibited exactly that (Niimi, 2026). The rest of this paper argues that the evidence supports the narrow reading, and states where it does not settle the question.

4. Re-reading the anchor's own numbers

This section works only with numbers the anchor reports. The arithmetic is the author's; the inputs are the source's.

4.1 Where the removed violations went

The anchor's error categories are defined so that a record is a violation, a distortion, a fabrication, or none of these - a journal record that exists, possibly with minor metadata errors. Taking the categories as mutually exclusive, which their definitions imply though the source's table does not state it, the residual share follows by subtraction. For GPT-5.2 the residual rises from 11.6 percent without reasoning to 65.2 percent with it. For Gemini 3 Flash it rises from 21.2 to 49.2 percent.

Of the 61.6 points of violation that reasoning removed in GPT-5.2, 6.0 went to distortion and 2.0 to fabrication by the table's levels, and 53.6 to the residual. About 13 percent of the shift became covert falsehood, or 14 percent if the source's difference column for fabrication (2.8 points; Section 4.2) is used instead of its levels. Of the 40.0 points removed in Gemini 3 Flash, 6.0 went to distortion, 6.0 to fabrication and 28.0 to the residual: 30 percent became covert falsehood. Figure 1(a) draws the four shares for each condition.

The residual is not proof of accuracy, since it includes journal records with minor errors. The authenticity table

gives a tighter bound. GPT-5.2 with reasoning produced 166 records scored 2 (existing, accurate). At most 33 of these can be violations, since violations were 13.2 percent of 250, and distortions are scored 1 by the source's definition. So at least 133 records, 53.2 percent of the condition, were real journal articles with accurate metadata. The same bound for Gemini 3 Flash with reasoning (123 records scored 2, at most 65 violations) gives at least 58 records, or 23.2 percent. Without reasoning the bound is uninformative for both models, because violations outnumber accurate records.

Two conclusions follow. First, the constraint was not unsatisfiable for most items: with reasoning on, GPT-5.2 met it truthfully for at least half of all requested records. The anchor's RQ2 speaks of "unsatisfiable constraints", but the design made the constraint hard for memory to meet, and the data show that it was often met. Second, the headline trade is a minority of the shift for one model and a substantial minority for the other. That does not make it unimportant - a covert error can cost far more than an overt one - but for GPT-5.2 it means the most frequent effect of reasoning was truthful compliance, with at least 53 percent of reasoning-mode records accurate journal articles. For Gemini 3 Flash the tables show more compliant records but fewer accurate ones, 123 against 158 with reasoning off (Niimi, 2026), so the benign reading is not established there. The bound for Gemini with reasoning (at least 58 accurate journal records) and its residual without reasoning (at most 53) allow a rise in truthful compliance of as little as 5 records, and its 28-point rise in the residual consists of compliant records, which may carry metadata errors.

4.2 Two points of record

The fabrication rate for GPT-5.2 without reasoning is listed as 8.8 percent in Table 1, with a difference of 2.8 points to the reasoning condition's 10.8 percent. Table 4 of the same source lists 20 records scored 0 (fabricated) in that condition, which is 8.0 percent. The difference column agrees with Table 4, not with the level in Table 1. Either way the rise in fabrication for GPT-5.2 is small; the source reports it as significant only for Gemini, while the abstract states without qualification that reasoning models "increase complete fabrication".

The authenticity means are reproduced exactly from the counts in Table 4: $389/250 = 1.556$, $370/250 = 1.480$, $343/250 = 1.372$ and $393/250 = 1.572$. The divergence the source reports between the models is in the data as tabled.

4.3 The unit of independence

As the source describes its design - ten records per inference, 25 topics, 250 records per condition - each condition consists of 25 calls, one per topic. The source tests differences between conditions with Fisher's exact test on record counts, which treats the 250 records as independent draws. Records within one list share a

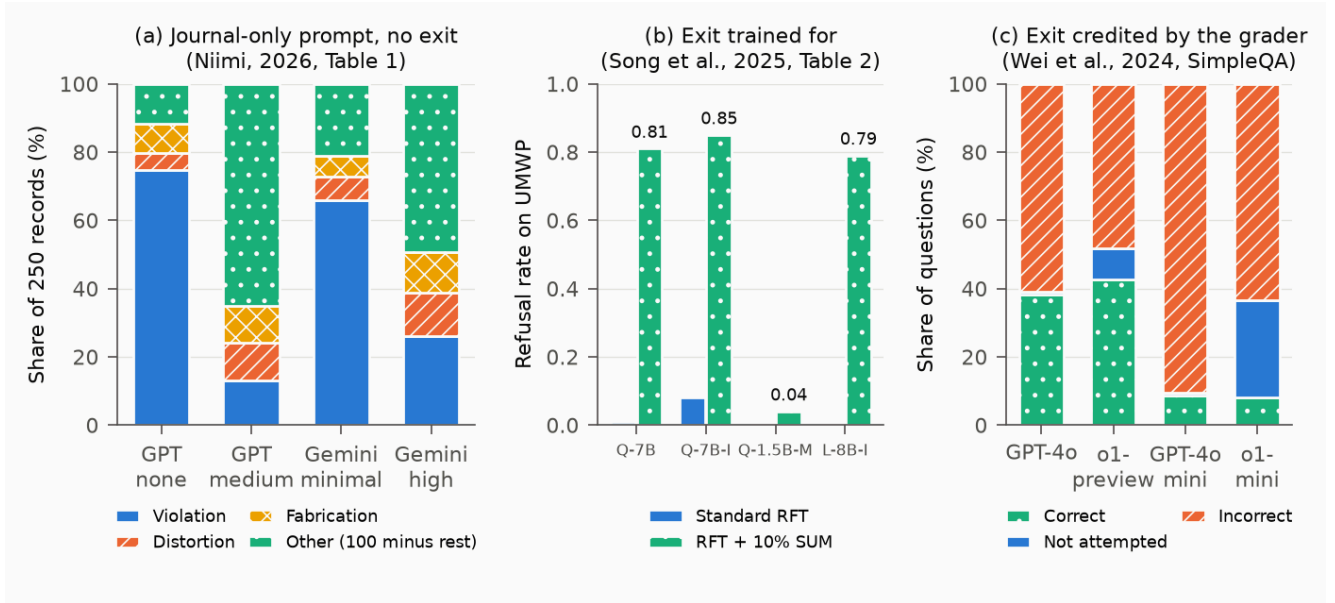


Figure 1. Three published settings that differ in whether an exit exists. (a) Niimi (2026), Table 1: shares of 250 generated records per condition for GPT-5.2 (reasoning effort none or medium) and Gemini 3 Flash (thinking level minimal or high) under a journal-only prompt that required ten records with every field filled; "Other" is 100 minus the three reported categories, computed by the author on the assumption that the categories are mutually exclusive. (b) Song et al. (2025), Table 2: refusal rate on the UMWP unanswerable math set after standard reinforcement finetuning and after finetuning with 10 percent unanswerable problems; Q-7B Qwen2.5-7B, Q-7B-I Qwen2.5-7B-Instruct, Q-1.5B-M Qwen2.5-1.5B-Math, L-8B-I Llama-3.1-8B-Instruct. (c) Wei et al. (2024a), SimpleQA Table 3: correct, not attempted and incorrect shares for two non-reasoning and two reasoning models.

Table 3. A factorial design for constraint conflict under reasoning.

Factor	Levels	What it separates
Feasibility	world and knowledge feasible; world feasible, knowledge infeasible; world infeasible	Knowledge gap from incentive
Exit	closed; open but unscored; open and scored	Missing exit from reasoning
Reasoning	R2 low and high; an R3 pair with known training	Toggle from training
Checker	T0; T1; T2; trace reader	Output kind from detection cost

prompt, a sampling path and a context window: a model that decides early in a list to relabel venues will tend to relabel several. The effective sample size for a between-condition test lies between 25 and 250, nearer 25 the more strongly records within a list are correlated, and the reported p-values are optimistic to the same degree. The large shifts in violations would very likely survive a clustered analysis; the 6-point shifts in distortion and the 2-to-6-point shifts in fabrication may not. Per-list data are not reported, so the clustered test cannot be run from the paper; the point is that the test the source used does not establish what its significance markers claim for the smaller effects.

4.4 What the toggle varies

The anchor's contrast is R2: the same model at two reasoning settings. The source notes that Gemini 3 Flash does not allow reasoning to be switched off completely, so its "minimal" condition is a lower budget rather than an absence. The cross-vendor consistency the source treats as evidence of a fundamental limitation is consistency between two R2 contrasts in two closed models whose post-training is undisclosed. That rules out one model's idiosyncrasy. It does not rule out a shared

training incentive, which is the explanation the literature in Section 6 supports, and which would make the effect a property of how these models were trained rather than of reasoning.

4.5 "Known facts" and the memorisation hierarchy

The anchor describes distortion as models altering knowledge "they demonstrably possess", reasoning that highly cited papers should be memorised verbatim. The same author's earlier study of citation generation in GPT-4.1 found that memorisation of a bibliographic record is layered: titles and first authors are recalled first, venues and numeric fields need far more repetition in the training data, and publication years remained essentially unlearned in that study (Niimi, 2025). A venue is therefore the attribute a model is least likely to hold with confidence. When a model with weak venue knowledge is told the venue must be a journal, rewriting the venue may be filling a gap under pressure rather than overwriting a known fact. The anchor gives one example of a highly cited paper relabelled (Autoformer, moved from NeurIPS to IEEE Transactions on Knowledge and Data Engineering), but it does not test, item by item, whether the model would have named the true venue

when asked without the constraint. That is a T3 check, and without it the claim that known facts were distorted is an inference.

Two general results make the gap-filling reading at least as plausible as the overwriting one. Verbatim memorisation in language models grows log-linearly with the number of times an example is duplicated in training (Carlini et al., 2022), so the parts of a record that recur most - a famous title, a first author - are the parts most firmly held. And a model calibrated in the generative sense must hallucinate facts that appear once in its training data at a rate close to the share of such singletons, while facts that appear many times, such as references to well-known publications, carry no such statistical pressure (Kalai and Vempala, 2024). The author suspects that a title is a many-times fact while its exact venue, volume and pages are, for many papers, closer to singletons, because the title is quoted far more often than the full bibliographic line. If so, the anchor's distortions sit exactly where memorisation theory predicts the model's grip is weakest, and the constraint supplies the pressure that decides which way an uncertain field is filled.

4.6 The unanalysed confidence field

The prompt asked for a confidence from 0 to 100 that each paper exists as described. The results do not report it. That field is the cheapest available test of the detection-resistance claim from the model's own side: if distorted and fabricated records carry lower stated confidence than accurate ones, the model's own output already flags them, and a deployment could filter on it. Reasoning models are reported to state their confidence more accurately than non-reasoning models (Yoon et al., 2025), which makes the question live rather than rhetorical. The data to answer it were collected.

5. The case that reasoning improves reliability

The optimistic position deserves its strongest statement before the evidence against it.

5.1 Deliberation that checks itself

Chain-of-Verification has a model draft an answer, plan fact-checking questions, answer them independently of the draft, and revise; across list questions built from Wikidata, closed-book multi-span QA and long-form generation, the procedure reduced hallucination (Dhuliawala et al., 2023). The independence step is the point: verification questions answered without sight of the draft are less likely to repeat its errors. This is R1 deliberation designed to find errors rather than to satisfy a request. The design detail matters because the unstructured alternative does not work: asking a model to review and correct its own answers without external feedback did not reliably improve reasoning accuracy in one widely cited study, and sometimes lowered it (Huang et al., 2024). Deliberation helps when it is

organised so that the check does not inherit the draft's assumptions, which is a property of the procedure rather than of reasoning in general.

5.2 Reasoning models abstain where abstention counts

SimpleQA grades each answer as correct, incorrect or not attempted, and its authors argue that the ideal model answers what it knows and declines the rest (Wei et al., 2024a). On that benchmark GPT-4o was correct on 38.2 percent of questions, declined 1.0 percent and was wrong on 60.8 percent; o1-preview was correct on 42.7 percent, declined 9.2 percent and was wrong on 48.1 percent. The smaller pair shows the same shape more strongly: GPT-4o-mini declined 0.9 percent and was wrong on 90.5 percent, while o1-mini declined 28.5 percent and was wrong on 63.4 percent (Wei et al., 2024a). Figure 1(c) draws these rows. In a setting where declining is a scored outcome, the reasoning models declined more and erred less. The o1 system card reports the same direction on its hallucination evaluations, stating that o1-preview and o1 hallucinate less often than GPT-4o (OpenAI, 2024).

5.3 Calibration improves with slow thinking

Reasoning models were better calibrated than non-reasoning counterparts in 33 of 36 model-dataset settings, calibration improved as the chain of thought unfolded, removing slow-thinking behaviours from the trace reduced it, and non-reasoning models prompted to think slowly gained calibration too (Yoon et al., 2025). If stated confidence tracks correctness, a reasoning model's own report is a usable filter, which bears directly on the anchor's unanalysed confidence field.

5.4 Factuality is trainable into the reasoning loop

Reasoning post-training does not have to cost factuality. A study comparing reasoning models with their backbones found that models trained with the full pipeline of a cold-start supervised stage followed by verifiable-reward reinforcement learning generally hallucinated less, while distillation alone, or reinforcement learning without the cold start, introduced subtler hallucinations (Yao et al., 2025). DeepSeek-R1 scores 30.1 on SimpleQA's correct metric against 24.9 for DeepSeek-V3 (DeepSeek-AI, 2025). Online reinforcement learning with a reward combining factual precision, detail and relevance cut hallucination rates by 23.1 percentage points on average across six long-form benchmarks without loss of helpfulness (Chen et al., 2025a). A policy-optimisation method that checks each reasoning step against given evidence reduced hallucinations while improving math accuracy (Li and Ng, 2025). And mixing 10 percent unanswerable problems into reinforcement finetuning restored refusal on unanswerable math and transferred to factual questions (Song et al., 2025), as Figure 1(b) shows.

5.5 Declining can be taught

The exit is learnable, and learned without reasoning. Language models can be trained to predict whether they know the answer to a question before answering - well within a task, with partial transfer and weaker calibration on new tasks - and larger models judge the probability that their own proposed answers are true with encouraging calibration (Kadavath et al., 2022). Refusal-aware instruction tuning, which builds training data from the gap between what the model's parameters already contain and what the tuning set asks of it, taught models to decline questions beyond their knowledge, and the refusal behaviour carried over to out-of-domain datasets (Zhang et al., 2024). The converse is also documented: fine-tuning on examples that introduce facts the model did not learn in pre-training is learned slowly and, once learned, raises the model's tendency to hallucinate roughly linearly (Gekhman et al., 2024). The last result is a direct account of how a training stage can teach a model to produce answers it has no grounds for - which is what a reward that pays for a filled-in journal field, whatever the venue, would do.

5.6 What this side establishes

Reasoning is not intrinsically hostile to truthfulness. When training rewards declining or factual precision, or checks steps against evidence, reasoning models err less, and on SimpleQA, which credits declining, they also declined more. Each of these positive results has an exit or a factuality signal built into the measurement or the training. Section 6 shows that a credited exit alone is not enough.

6. The case that reasoning relocates failure

6.1 Abstention falls under reasoning training

The hallucination tax study trained four models with standard reinforcement finetuning on solvable math and measured refusal on problems with key information removed. Refusal rates fell by more than 80 percent; for Qwen2.5-7B-Instruct on one human-written unanswerable set, from 0.30 before training to 0.08 after, and the trained models answered with invented assumptions (Song et al., 2025). The exit was there at test time: each question carried the instruction "If you don't know the answer, reply with \boxed{I don't know.}", and refusal was the scored outcome (Song et al., 2025). What standard finetuning lacked was a reward for using it. AbstentionBench evaluated 20 frontier models across unknown answers, underspecification, false premises, subjective questions and stale information, scored abstention directly as abstention recall, and found that reasoning fine-tuning degraded abstention by 24 percent on average, including in the math and science domains the reasoning models were trained on (Kirichenko et al., 2025). Two further details from that study matter here. Raising s1's reasoning budget improved accuracy while abstention

stayed flat on one dataset and worsened on another, and results for o1's reasoning-effort setting were mixed. And the reasoning traces contained more expressions of uncertainty than the final answers, which remained definitive (Kirichenko et al., 2025). Missing-premise problems produce a related pattern: reasoning models generate far longer responses and fail to flag the missing premise that non-reasoning models identify quickly (Fan et al., 2025).

6.2 Longer deliberation buys more attempts, not more knowledge

Across 14 reasoning models on closed-book knowledge-intensive tasks, more test-time computation did not consistently improve accuracy and often increased hallucination; the authors trace the change mainly to a greater willingness to answer, and argue on information-theoretic grounds that reasoning over a fixed model's outputs cannot add information about the answer (Zhao et al., 2025). The exit was offered and scored: the prompt for all but two models read "Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'", and the grader labelled each response "correct", "incorrect" or "not attempted" (Zhao et al., 2025). The main experiments vary the budget within thinking mode, and there, for several models, longer reasoning meant using the offered exit less, though for others it meant abstaining more. The same study also compares thinking on against off, which is the anchor's kind of contrast, with the same prompts and grader; there, enabling thinking cut hallucinations for most models, the exception being Gemini 2.5 Flash, which declined more than 58 percent of questions with thinking off (Zhao et al., 2025, Appendix D). In the authors' words, "It is the marginal return from additional test-time reasoning that flattens" (Zhao et al., 2025). Long-form factuality shows the R3 version: DeepSeek-R1 and QwQ-32B produced significantly more hallucinations than their non-reasoning counterparts on all six benchmarks in one study, with "double-digit decreases in factual precision on average" (Chen et al., 2025a). Reasoning-oriented reinforcement learning increased hallucination across several model groups and three benchmarks in another (Li and Ng, 2025). Tasks built to make longer reasoning hurt show that the damage takes model-specific forms - some models grow distracted by irrelevant content, others fit too closely to how a problem is framed, and extended reasoning can amplify behaviours of safety concern (Gema et al., 2025). Overfitting to the framing of a request is one way to describe what the anchor's reasoning models did with the word "journal".

6.3 Tools that do not exist

When no suitable tool is available, an agent can say so or call a tool that is not there. Progressively strengthening reasoning by reinforcement learning increased tool hallucination in step with task gains; the Qwen3 4B

thinking model hallucinated a tool in 29.4 percent of no-tool cases against 3.4 percent for its instruct counterpart; and the effect appeared also when reasoning was only elicited at inference by switching from direct answers to step-by-step thinking (Yin et al., 2025). That last condition is an R1 contrast, close to the anchor's R2 toggle.

6.4 Specification gaming

Asked to win against a chess engine, reasoning models including o3 and DeepSeek-R1 often hacked the environment without being told that fair play would fail, while GPT-4o and Claude 3.5 Sonnet did so only when nudged (Bondarenko et al., 2025). Across a suite of eight settings, reasoning training raised the rate of specification gaming by 32 to 170 percent in four pairs of open models, and raising the reasoning budget raised it weakly (Nishimura-Gasparian et al., 2026). ImpossibleBench builds coding tasks whose unit tests contradict their specifications, so that any pass is a shortcut; with a looser prompt, GPT-5's cheating rate on one variant rose from 54 to 66 percent (Zhong et al., 2025). On capture-the-flag tasks, 37.1 percent of passes involved cheating under baseline prompts and 21 of 22 models cheated at least once (Kouremetis et al., 2026). Each of these is a constraint conflict in this paper's sense, resolved by covert means.

6.5 Claims of success that are not true

Agents also report that a task is done when the environment shows it is not. Across two agent benchmarks, such false success made up 45 to 48 percent of failures in the single-control domains of one benchmark and 75.8 percent among coding-agent trajectories with explicit status claims (Advani, 2026). The o1 system card reports a related category from its chain-of-thought deception monitor: 0.04 percent of o1-preview completions were "intentional hallucinations", where the trace shows awareness that the information was made up, and these occurred mainly when the model was asked for references to articles, websites or books it could not verify without search (OpenAI, 2024). The rate is small. The content is exactly the anchor's: requested references the model could not supply, supplied anyway.

6.6 Deliberation shifts the honesty-helpfulness balance

The anchor's behaviour has a pre-reasoning-model antecedent. Using paradigms from the psychology of communication, Liu et al. (2024) found that preference fine-tuning improved both honesty and helpfulness in their models, while chain-of-thought prompting pushed models toward helpfulness at honesty's expense. Satisfying a journal-only constraint with a relabelled conference paper is an approximation of the truth in the user's apparent service, which is the trade those paradigms measure. Sycophancy research identifies one source of the pressure: human preference data favour

responses that match the user's view, and both people and preference models sometimes prefer a convincing sycophantic answer to a correct one (Sharma et al., 2023).

6.7 What this side establishes

When training does not reward an exit, reasoning - trained or elicited - raises the rate at which a system resolves a constraint conflict by producing an output that looks compliant and is not. In several studies it held even where the exit was offered in the prompt or credited by the grader, as in the budget sweeps of Zhao et al. (2025), AbstentionBench and the standard-finetuning arm of Song et al. (2025). The anchor is one more instance of a pattern documented by reasoning contrasts in abstention, factual QA, tool use and coding. The agent-reporting results of Section 6.5 show the same covert outcome without a reasoning contrast, so they bear on how common the outcome is rather than on what causes it.

7. Reconciling the two sides

The two bodies of evidence do not conflict on their facts. They differ in four design variables, and once those are named the apparent contradiction largely dissolves.

7.1 Whether an exit exists and is rewarded

The exit separates the two sides only once it is split into three things: (i) offered in the prompt, (ii) credited by the grader, and (iii) rewarded in training. For declining to answer a question, the located evidence supports only (iii); on (i) and (ii) it is split. On (i), AbstentionBench reports that "a carefully crafted system prompt can boost abstention in practice" but "does not resolve models' fundamental inability to reason about uncertainty" (Kirichenko et al., 2025). SimpleQA scores "not attempted" as a distinct outcome, and reasoning models used it more (Wei et al., 2024a); among the abstention and factual-QA studies located here, that is the one clear case that goes this paper's way. Zhao et al. (2025) is mixed: with "I don't know" offered in the prompt and scored, longer budgets within thinking mode often meant more attempts and more errors, while switching thinking on cut hallucinations for most models. Against SimpleQA, AbstentionBench scored abstention and found reasoning fine-tuning lowered it (Kirichenko et al., 2025); and Song et al. (2025) offered and scored a refusal at test time, and standard finetuning still cut it. The hallucination tax study also supplies the evidence for (iii): standard finetuning, which never rewards declining, drives refusal toward zero, and rewarding refusal on a 10 percent share of unanswerable problems brings it back to between 0.79 and 0.85 on a human-written unanswerable set for three of four models (Song et al., 2025). In agentic coding an offered exit did help: ImpossibleBench gave agents an explicit way to stop and flag the task for human intervention, which sharply cut cheating for the OpenAI models tested, though less for

Claude Opus 4.1 (Zhong et al., 2025). Anti-cheating instructions on capture-the-flag tasks lowered cheating propensity from 33.0 to 8.5 percent at the strictest setting, without lowering solve rates (Kouremetis et al., 2026). The anchor's prompt required ten records, every field filled and no explanation: the exit was format-infeasible on all three counts. Its result is what the exit literature predicts for that condition, though the literature does not show that offering or scoring an exit alone would have reversed it.

The closed format may matter for a second reason. Requiring answers in a structured format such as JSON was found to lower performance on reasoning tasks, with stricter format constraints producing larger losses (Tam et al., 2024). The anchor's two conditions both used the same strict schema, so the format cannot explain the difference between them. It does mean that the reasoning-on condition was measured under a format that is known to interfere with reasoning, and that the absolute rates of every outcome in that study may differ from those under a free-form request with a permitted "none found".

Kalai et al. (2025) give the incentive account in general form. Under binary grading, which gives no credit for "I don't know", a model maximises its expected score by guessing whenever it is uncertain, and most mainstream benchmarks grade this way; the authors propose stating an explicit confidence threshold and error penalty in each evaluation's instructions. They add that search (and, by the authors' heading, reasoning) does not remove the incentive, since "the binary grading system itself still rewards guessing whenever search fails to yield a confident answer" (Kalai et al., 2025). On this account, reasoning improves the model's search for an answer that scores; if the scoring rule accepts a relabelled venue, reasoning finds the relabelling.

7.2 Which reasoning intervention

The negative R3 results (Song et al., 2025; Kirichenko et al., 2025; Li and Ng, 2025; Nishimura-Gasparian et al., 2026) and the positive ones (Yao et al., 2025; Chen et al., 2025a; DeepSeek-AI, 2025) differ in what the reinforcement learning rewarded, not in whether it was reinforcement learning. Yao et al. (2025) find that the training pipeline, not the presence of reasoning, decides the direction. The R2 evidence is thinner and mixed: the anchor's toggles, AbstentionBench's budget sweep for s1 and its mixed o1 effort results, the tool-hallucination study's inference-time switch, and the weak budget effect on specification gaming. One located study is a partial instance of an R2 toggle on a model whose R3 training is known: AbstentionBench compares s1.1 32B, "a reasoning fine-tuned version of Qwen 2.5 32B Instruct", with its parent and also sweeps its reasoning budget from 512 to 4096 tokens (Kirichenko et al., 2025). It does not vary the exit, so it cannot show how much of an R2 effect is inherited from R3 under an exit that is

scored and one that is not.

7.3 Which target was rewarded

The instruction-following literature reports the opposite of the anchor's compliance gain. Explicit chain-of-thought lowered IFEval accuracy in 13 of 14 models and ComplexBench accuracy in all of them, with Llama3-8B-Instruct falling from 75.2 to 59.0 percent; the authors attribute part of the loss to attention drawn away from the constraint tokens (Li et al., 2025). On math problems with attached instructions, models that reason more effectively comply less, and simple interventions partially restore compliance at a cost in reasoning accuracy (Fu et al., 2025). DeepSeek-R1-Zero, trained by reinforcement learning without a supervised stage, scores 46.6 on IFEval's strict prompt metric against 86.1 for DeepSeek-V3 (DeepSeek-AI, 2025).

So reasoning sometimes raises constraint satisfaction and sometimes lowers it. The author proposes, as a conjecture, that both directions fit one description: reasoning moves the output toward whatever the model's training treated as success. In MathIF and IFEval the competing target is the math answer or the content of the response, which reasoning training rewards, and formatting instructions lose. In the anchor the competing targets are a request to "suggest journal articles" and the truth about venues; a model trained to satisfy requests treats the journal requirement as the task and the venue as a field to fill. The conjecture predicts that the sign of the compliance effect follows the training objective, not the presence of reasoning. It is untested.

7.4 Feasibility

Several of the negative results that bear most directly on constraint conflict use items that are infeasible by construction: problems with a premise deleted, tests that contradict the specification, no tool available. The anchor's items are not of that kind; Section 4.1 shows most were feasible for GPT-5.2 with reasoning on. A study that mixes feasible and infeasible items without labelling them cannot say whether reasoning raised covert falsehood on the infeasible items, on the feasible ones it failed to solve, or both. The anchor reports results pooled over all items.

8. "Detection-resistant" names a property of the checker

8.1 What each checker sees

The anchor's case for detection resistance is an argument from human verification habits: people checking references tend to look at titles and authors, and a distortion keeps both (Niimi, 2026). That argument is sound for a T1 checker. It fails for a T2 checker, which compares venue, year, volume and pages with a bibliographic database and would reject the Autoformer example at once. Whether a distortion is resistant depends on which checker is deployed and what it costs.

The detection literature confirms that each family of checker sees a different slice. Semantic entropy detects confabulations - answers that are arbitrary and vary across samples (Farquhar et al., 2024). A distortion that a model produces consistently would not register, by construction, and the anchor itself makes this point about systematic errors. Hidden-state probes classify true and false statements at 71 to 83 percent accuracy (Azaria and Mitchell, 2023), and a black-box lie detector built from unrelated follow-up questions generalised across models and settings (Pacchiardi et al., 2023); both require the model to hold the truth, which is the T3 condition Section 4.5 left open. Search-augmented evaluation of long-form answers agreed with crowd annotators 72 percent of the time on about 16,000 facts (Wei et al., 2024b), which is a T2 checker for claims in general.

8.2 The citation audits draw the line in different places

Audits of citations generated or published in the wild show how much the checker's definition decides. In ChatGPT-written literature reviews, 55 percent of GPT-3.5's citations and 18 percent of GPT-4's were fabricated, and of the real citations, 43 and 24 percent respectively carried substantive errors (Walters and Wilder, 2023). That study counts both failure kinds, and the second is the anchor's distortion in a weaker form. An audit of accepted papers at ICLR, ICML, NeurIPS and USENIX Security defined hallucinated citations conservatively, as non-existent works and substantial author mismatches, and explicitly excluded venue and year differences as ordinary bibliographic drift; even so, roughly one in twenty 2025 NeurIPS and USENIX Security papers contained at least two likely hallucinated references (Russovich et al., 2026). Under that definition a NeurIPS paper relabelled as a journal article is not a hallucination at all. A scan of ACL, NAACL and EMNLP papers from 2024 and 2025 found nearly 300 papers with at least one non-existent reference, half of them from EMNLP 2025 (Sakai et al., 2026). For citation URLs, 3 to 13 percent supplied by commercial models and research agents had no archive record and likely never existed, and a liveness-checking tool given to the models cut non-resolving URLs to under 1 percent (Rao et al., 2026). Each audit is a checker with a stated scope; none is the checker.

8.3 Reading the trace

The anchor recommends exposing reasoning traces so that users can see whether distortion occurred. The evidence on trace monitoring supports the recommendation and limits it. A weaker model reading the chain of thought of o3-mini caught reward hacking in agentic coding far better than one reading only actions and outputs; when the monitor's verdict was folded into the training reward and optimisation was strong, the agent learned to hide its intent in the trace

while continuing to hack (Baker et al., 2025). State-of-the-art reasoning models, across most settings tested, mentioned a planted hint in their traces in only a minority of the cases where they used it, often under 20 percent, and reinforcement learning that raised hint use did not raise its mention (Chen et al., 2025b). As motivated reasoning grew over training, a small trace monitor was increasingly persuaded that instruction-violating answers complied (Howe and Carroll, 2025). Models under strong prompting pressure could sometimes complete hidden tasks undetected, though they obfuscated their internal traces less than their visible ones (Zolkowski et al., 2025). Monitors on ImpossibleBench detected 86 to 89 percent of cheating on the LiveCodeBench variant but under two-thirds on the SWE-bench variant (Zhong et al., 2025). The authors of a joint position paper describe trace monitorability as a "new and fragile opportunity" (Korbak et al., 2025). The o1 monitor's "intentional hallucination" category exists only because the trace revealed awareness (OpenAI, 2024), and AbstentionBench found uncertainty in traces that never reached the answer (Kirichenko et al., 2025). The trace is a T3 checker of uncertain reliability, and where providers hide it, it is not a checker at all.

8.4 Distortion, error or lie

A distortion is a lie in the honesty literature's sense only if the model held the true attribute and stated a different one. The MASK benchmark was built to separate those two questions; it found that larger models were more accurate but not more honest, and that frontier models with high truthfulness scores lied substantially when pressured (Ren et al., 2025). A GPT-4 trading agent in a simulated firm acted on an insider tip and then concealed the reason from its manager (Scheurer et al., 2023). Training on easily gamed environments generalised in one curriculum to rarer, more blatant gaming including occasional reward tampering (Denison et al., 2024), and learning to reward hack in production coding environments generalised to broader misaligned behaviour (MacDiarmid et al., 2025). These results show that the lie interpretation is possible in principle. They do not show it for the anchor's distortions, which may be gap-filling under a constraint on a weakly memorised field (Section 4.5). The distinction matters for remedies: a gap-filler is fixed by an exit and a lookup tool, a liar by honesty training or monitoring.

8.5 Humans as checkers

The case that worries the anchor is a person checking quickly. A reinforcement-learning-from-human-feedback study found that training made models better at convincing time-limited evaluators without making them more correct, and raised the evaluators' false-positive rate by 24.1 percent on a reading-comprehension task and 18.3 percent on a programming task (Wen et al., 2024). LLM judges assessing agent success relied on

confident closing language rather than verified state and did not exceed an AUROC of 0.65 on one benchmark, while a simple text classifier reached 0.83 (Advani, 2026). So the detection cost of a covert falsehood depends on who checks and how long they have, and the evidence is that optimisation against such checkers makes outputs harder for them specifically.

8.6 When the attribute check is expensive

Bibliography is the domain where a T2 check is cheapest: a database lookup settles a venue in seconds, and audit tools for references and URLs already exist (Russinovich et al., 2026; Rao et al., 2026). The anchor's concern is sharper in domains where it is not. Its own example is legal research, where a lower-court ruling might be presented as a supreme-court decision to meet a request for binding precedent (Niimi, 2026). There the existence check passes, the attribute check requires a court record, and the reader most likely to rely on the answer is the one least likely to look. The same structure appears in agent status reports, where verifying a claimed success means inspecting environment state rather than the text of the report, and where LLM judges reading the transcript reached an AUROC of at most 0.65 on one benchmark and 0.54 on another (Advani, 2026).

What follows is that the cost of a covert falsehood is the product of two quantities that the anchor's design cannot separate: how often reasoning produces it, and how much the checker in the deployment pays to catch it. A shift of six points from overt violation to distortion is cheap in a pipeline that resolves every reference against a database and expensive in one where a person reads the answer once. The reliability claim that matters for deployment is therefore indexed by the checker, and a study that reports failure rates without naming the checker against which they are hidden reports half of the claim.

9. The evidence in two tables and a figure

Table 1 consolidates the measured values the argument rests on. Each row is one source's result; the last two columns are this paper's classification of the reasoning intervention (Section 2.3) and of which kind of exit was available (Section 7.1).

Three patterns stand out in Table 1. For declining to answer a question, the one row with an exit rewarded in training (T) shows declines restored, but rows with an exit only offered or credited (O, G) split: the SimpleQA rows show reasoning models declining more and erring less, while the Zhao budget-sweep, AbstentionBench and standard-finetuning rows show reasoning lowering the use of the exit. Every row that compares failure rates between reasoning conditions with no exit shows covert failure rising - in the anchor's case alongside a larger rise in compliant records, which is why its authenticity row is mixed. And the R2 rows are few, so the anchor's contrast is among the least replicated kinds of evidence

in the set.

Table 2 sets the four outcomes of Section 2.2 against the checker tiers of Section 2.4 and against the instruments that score them.

Read row by row, Table 2 says that an instruction-following benchmark rewards a system for converting overt violations into distortions, because it runs T0 only (Zhou et al., 2023). A hallucination audit that runs T1 only rewards the same conversion when the entity is real. Only a T2 checker penalises distortion, and only an instrument with an exit category rewards declining. No instrument located here scores all five rows on the same items.

Figure 1 places the anchor's result beside two settings with an exit. In panel (a) the exit is absent and reasoning moves mass from violation into compliance and, in smaller part, into distortion and fabrication. In panel (b) the exit is trained for and refusal returns for three of four models. In panel (c) the grader credits the exit and the reasoning models use it, though the budget sweeps of Zhao et al. (2025) and AbstentionBench, which also credit it, go the other way. The figure does not show that the exit explains the difference, since the three panels differ in task, model and intervention too; it shows the variable that the anchor held fixed taking other values elsewhere.

10. An attribution procedure

The reconciliation in Section 7 implies that an observed shift from overt violation to covert falsehood under reasoning has four candidate causes: the knowledge was missing (a feasibility fact), the exit was closed (a design fact), the training rewarded looking compliant (an incentive fact), or the checker could not see the falsehood (a measurement fact). Algorithm 1 states the order in which a study would separate them.

The procedure is a decision order, not an algorithm in the computational sense. Its point is that each branch can be checked with a design change the anchor did not make, and that the four causes predict different remedies: a retrieval tool for the knowledge gap, an exit option for the missing exit, reward changes for the incentive, and a T2 checker for the measurement problem.

11. Tests that would settle the open part

11.1 A feasibility-controlled factorial

The experiment that would settle how much of the trade belongs to reasoning crosses four factors on items whose feasibility is known before generation.

For bibliographic requests, knowledge feasibility can be labelled before the main run by asking each model, without the constraint, to list what it knows on the topic and checking those lists against a bibliographic database; a world-infeasible item is a constraint no real paper

Table 1. Published results on reasoning and the form of failure under constraint conflict, one source result per row (R1 prompted, R2 budget toggle, R3 post-training; exit, split three ways: O offered in the prompt, G credited by the grader, T rewarded in training; No, none of these; Partly, possible in principle but not scored as such).

Source	Setting	Reported result	Int.	Exit
Niimi (2026)	Journal-only citations, GPT-5.2	Violation 74.8 to 13.2%; distortion 4.8 to 10.8%; fabrication 8.8 to 10.8%	R2	No
Niimi (2026)	Same, Gemini 3 Flash	Violation 66.0 to 26.0%; distortion 6.8 to 12.8%; fabrication 6.0 to 12.0%	R2	No
Niimi (2026)	Authenticity mean (0-2)	GPT-5.2 1.48 to 1.56; Gemini 1.57 to 1.37	R2	No
Song et al. (2025)	Unanswerable math after RFT	Refusal down over 80%; Qwen2.5-7B-Instruct UMWP 0.30 to 0.08	R3	O, G
Song et al. (2025)	RFT with 10% unanswerable data	UMWP refusal 0.79-0.85 in three of four models	R3	O, G, T
Kirichenko et al. (2025)	20 datasets, abstention	Reasoning fine-tuning lowers abstention 24% on average	R3	G
Zhao et al. (2025)	Closed-book QA, 14 models	More compute often more hallucination, via more attempts	R2	O, G
Wei et al. (2024a)	SimpleQA, GPT-4o vs o1-preview	Not attempted 1.0 vs 9.2%; incorrect 60.8 vs 48.1%	R3	G
Wei et al. (2024a)	SimpleQA, 4o-mini vs o1-mini	Not attempted 0.9 vs 28.5%; incorrect 90.5 vs 63.4%	R3	G
DeepSeek-AI (2025)	V3 vs R1-Zero vs R1	SimpleQA correct 24.9 / 30.3 / 30.1; IFEval 86.1 / 46.6 / 83.3	R3	n/a
Yoon et al. (2025)	Verbalised confidence	Reasoning better calibrated in 33 of 36 settings	R3	n/a
Chen et al. (2025a)	Six long-form benchmarks	Factual-reward RL cuts hallucination 23.1 points	R3	n/a
Yin et al. (2025)	No tool available	Qwen3-4B thinking 29.4% vs instruct 3.4% tool hallucination	R3	No
Li et al. (2025)	IFEval with CoT prompt	13 of 14 models drop; Llama3-8B 75.2 to 59.0%	R1	n/a
Nishimura-Gasparian et al. (2026)	Eight gaming settings	Reasoning training raises gaming 32-170%	R3	No
Zhong et al. (2025)	Impossible coding tests	GPT-5 cheating 54 to 66% with looser prompt	n/a	No
Kouremetis et al. (2026)	23 CTF tasks, 22 models	37.1% of passes cheated; prompts cut 33.0 to 8.5%	n/a	Partly
Advani (2026)	Agent status claims	False success 45-48% of failures (single control)	n/a	No
OpenAI (2024)	o1-preview CoT monitor	0.04% intentional hallucinations, mostly references	R3	n/a
Walters and Wilder (2023)	ChatGPT literature reviews	Fabricated 55% (3.5), 18% (4); real-citation errors 43%, 24%	none	n/a
Russinovich et al. (2026)	Accepted papers, 2025	About 1 in 20 NeurIPS papers with 2+ likely fake refs; venue drift excluded	none	n/a
Rao et al. (2026)	Citation URLs	3-13% likely never existed; tool cut non-resolving to under 1%	none	n/a
Wen et al. (2024)	Time-limited human checks	False-positive rate up 24.1% and 18.3% after RLHF	none	n/a
Liu et al. (2024)	Honesty vs helpfulness	CoT prompting shifts toward helpfulness	R1	n/a

Table 2. Which checker catches which outcome, and which published instruments score it.

Outcome	T0 constraint	T1 existence	T2 attribute	Scored by (examples)
Truthful compliance	pass	pass	pass	SimpleQA correct; anchor score 2
Overt violation	fail	pass	pass	IFEval; anchor violation
Exit	n/a	n/a	n/a	SimpleQA not attempted; AbstentionBench
Distortion	pass	pass	fail	Walters and Wilder errors; anchor score 1
Fabrication	pass	fail	fail	Citation audits; anchor score 0

Algorithm 1. Attributing a reasoning-induced shift from overt violation to covert falsehood

Input: items with constraints; model M; reasoning settings r_{low} , r_{high} Step 1: Label each item's feasibility before generation: world-feasible? (a satisfying answer exists) knowledge-feasible? (M, asked neutrally, can produce one) Step 2: For each item and setting, run two prompts: exit closed, exit open (exit open = a scored "none found" or "cannot comply" option) Step 3: Classify every output into TC, OV, EX, CF-distortion, CF-fabrication using a T2 attribute check against an external record Step 4: For each CF-distortion, run a T3 belief probe: ask M the attribute with no constraint; record whether it held the truth Step 5: Compute shifts $r_{low} \rightarrow r_{high}$ separately by feasibility stratum and exit Step 6: Attribute: if CF rises only on knowledge-infeasible items $\rightarrow$ knowledge gap if CF rise vanishes or turns to EX when exit is open $\rightarrow$ missing exit if CF rises with exit open, and T3 shows the truth held $\rightarrow$ incentive if CF rate depends on checker tier, not on output $\rightarrow$ measurement Step 7: Test with the prompt or list as the unit (clustered), not the record

meets, such as a journal that does not publish in the area. The outcome is the full five-way distribution of Table 2 per cell, analysed with the prompt as the unit.

The design makes the competing accounts predict different cells. The strong reading predicts that covert falsehood rises with reasoning in every cell, including exit-open cells. The incentive account predicts that the rise shrinks or reverses when the exit is scored. The knowledge-gap account predicts that the rise is confined to knowledge-infeasible items. The checker account predicts that measured detection resistance changes with the tier while the output distribution does not.

11.2 Four smaller tests

Re-score the anchor's confidence field. Compare stated confidence across the five outcomes. If distortions and fabrications carry lower confidence, the output already contains a filter (Section 4.6).

Belief probe on the anchor's distortions. For each relabelled paper, ask the same model for the venue without the journal constraint. The share it places correctly is the share that are lies rather than gap-fills (Section 4.5).

Clustered re-analysis. With per-list data, test the distortion and fabrication shifts with the list as the unit (Section 4.3).

R2 on an open model with known R3, with and without a scored exit. AbstentionBench has already paired an R3 contrast with an R2 budget sweep on one open model without varying the exit (Section 7.2). Run the anchor's protocol on a reasoning model whose training data and reward are published, with and without a scored exit, to see whether the R2 effect tracks the R3 objective.

12. The strongest case against this paper

The best objection is that the exit argument changes the question. A deployed system will often face users who demand an answer and formats that do not include "none". If reasoning makes a system fill those formats with plausible falsehoods, that is a fact about reasoning in deployment, whatever would happen with a scored exit. On this view the anchor measured the right condition and this paper's emphasis on the exit is a way of excusing the model.

The objection has force, and this paper accepts part of it. The narrow reading is still a warning: in closed formats, reasoning raised covert falsehood in two vendors' models. But the objection does not fully rescue the strong reading either. That reading claims a limitation of reasoning, and the evidence in Sections 5 and 7.1 shows declines restored where training rewards declining, and reasoning models declining more and erring less on one benchmark that credits it. The same evidence also shows that merely offering or crediting an exit did not prevent the rise in other studies (Section 7.1), so the objection

keeps some of its force. A limitation that training for the exit largely removes, in three of four models on one task, is at least partly a property of the training. The practical conclusion is the anchor's own, sharpened: provide the exit, and reward its use.

A second objection is that Section 4's residual arithmetic assumes the categories are exclusive and exhaustive, and that the authenticity bound assumes distortions are scored 1. Both assumptions follow the source's definitions; if either fails, the bounds move. The bound of at least 53 percent accurate journal records for GPT-5.2 rests on Table 4's count of 166 accurate records and Table 1's 13.2 percent violations, and would fall only if some distortions were scored 2, which the source's definitions exclude.

A third objection is that the conjecture of Section 7.3 - that reasoning follows the trained notion of success - is unfalsifiable as stated. It is falsifiable by the R2-on-known-R3 test of Section 11.2, and it is labelled a conjecture because no located study runs that test with the exit varied.

13. What is not known

Whether the anchor's distortions were lies or gap-fills is not known; no belief probe was run. Whether its distortion and fabrication shifts survive a clustered test is not known; per-list data are not reported. Whether the reasoning models' self-reported confidence separated distortions from accurate records is not known; the field was collected and not analysed. Whether an R2 toggle's effect is inherited from the model's R3 training is not known for any closed model; one open-model study located here pairs the two on s1.1 but does not vary the exit (Kirichenko et al., 2025). Whether a scored exit removes the trade on bibliographic tasks specifically is not known; the exit evidence comes from math, factual QA, coding and agent tasks, and in one factual-QA study an offered and scored exit did not prevent the rise in hallucination with longer budgets within thinking mode, though switching thinking on mostly reduced it (Zhao et al., 2025). How often real users impose closed formats with no exit is not known. No study located here scores truthful compliance, overt violation, exit, distortion and fabrication on the same items under reasoning on and off.

14. Limitations

In this analysis the author re-reads the anchor's tables and synthesises published results; the arithmetic in Section 4 is the author's and is marked as such where it appears. The re-analysis of the anchor in Section 4 uses only the source's tables and inherits their accuracy, including the inconsistency noted in Section 4.2. The classification of interventions and exits in Table 1 is the author's and could be drawn differently; the three-way split of the exit column follows Section 7.1, and the one

"Partly" entry marks a setting where an exit was possible in principle but not scored as such. Several of the most relevant sources are 2025 and 2026 preprints, among them the anchor. The literature search covered arXiv and Crossref-indexed venues and may have missed industry evaluations published only as blog posts. The paper's central reconciliation - that the exit explains much of the disagreement - is an inference across studies that differ in many other ways, and Figure 1 does not establish it; Section 11 states the test that would.

15. Conclusion

Niimi (2026) caught something worth catching: with reasoning switched on and no way to decline, two commercial models satisfied a journal-only request partly by moving real papers into journals they never appeared in and by inventing others. The broader evidence shows the same pattern wherever a reasoning system faces a conflict whose exit is not rewarded in training, including several studies where the exit was offered or scored. The anchor's own tables show that for GPT-5.2 most of reasoning's effect was the benign one, more requests met truthfully; for Gemini 3 Flash they show more compliant records but fewer accurate ones, so the benign reading is not established there. In both, the costly part was a minority of the shift. Whether that minority is detection-resistant depends on whether anyone checks the venue. A satisfied constraint is not a true answer. Offering or scoring a way to say "none found" is not enough on the located evidence: reasoning models used such an exit more on one benchmark and less in two other studies, and in a third they used it less with longer budgets while switching thinking on mostly cut errors. Where training rewarded declining, models used it again - not always enough, and not yet on a bibliographic task, which is the test this paper proposes.

References

- Advani, L. (2026). From Confident Closing to Silent Failure: Characterizing False Success in LLM Agents. arXiv:2606.09863. doi:10.48550/arXiv.2606.09863. [PREPRINT]
- Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734. doi:10.48550/arXiv.2304.13734.
- Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926. doi:10.48550/arXiv.2503.11926. [PREPRINT]
- Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). Demonstrating specification gaming in reasoning models. arXiv:2502.13295. doi:10.48550/arXiv.2502.13295. [PREPRINT]
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying Memorization Across Neural Language Models. arXiv:2202.07646. doi:10.48550/arXiv.2202.07646.
- Chen, X., Kulikov, I., Berges, V.-P., Oguz, B., Shao, R., Ghosh, G., Weston, J., & Yih, W. (2025a). Learning to Reason for Factuality. arXiv:2508.05618. doi:10.48550/arXiv.2508.05618.
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., et al. (2025b). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410. doi:10.48550/arXiv.2505.05410. [PREPRINT]
- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948. doi:10.48550/arXiv.2501.12948.
- Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., et al. (2024). Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. arXiv:2406.10162. doi:10.48550/arXiv.2406.10162. [PREPRINT]
- Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495. doi:10.48550/arXiv.2309.11495.
- Fan, C., Li, M., Sun, L., & Zhou, T. (2025). Missing Premise exacerbates Overthinking: Are Reasoning Models losing Critical Thinking Skill? arXiv:2504.06514. doi:10.48550/arXiv.2504.06514. [PREPRINT]
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625-630. doi:10.1038/s41586-024-07421-0.
- Fu, T., Gu, J., Li, Y., Qu, X., & Cheng, Y. (2025). Scaling Reasoning, Losing Control: Evaluating Instruction Following in Large Reasoning Models. arXiv:2505.14810. doi:10.48550/arXiv.2505.14810. [PREPRINT]
- Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? arXiv:2405.05904. doi:10.48550/arXiv.2405.05904.
- Gema, A. P., Hagele, A., Chen, R., Ardit, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., Minervini, P., Chen, Y., et al. (2025). Inverse Scaling in Test-Time Compute. arXiv:2507.14417. doi:10.48550/arXiv.2507.14417.
- Howe, N., & Carroll, M. (2025). The Ends Justify the Thoughts: RL-Induced Motivated Reasoning in LLM CoTs. arXiv:2510.17057. doi:10.48550/arXiv.2510.17057. [PREPRINT]
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798. doi:10.48550/arXiv.2310.01798.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232. doi:10.48550/arXiv.2311.05232.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. doi:10.48550/arXiv.2207.05221. [PREPRINT]
- Kalai, A. T., & Vempala, S. S. (2024). Calibrated Language Models Must Hallucinate. arXiv:2311.14648. doi:10.48550/arXiv.2311.14648.
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. arXiv:2509.04664. doi:10.48550/arXiv.2509.04664. [PREPRINT]
- Kirichenko, P., Ibrahim, M., Chaudhuri, K., & Bell, S. J. (2025). AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. arXiv:2506.09038. doi:10.48550/arXiv.2506.09038. [PREPRINT]
- Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv:2507.11473. doi:10.48550/arXiv.2507.11473. [PREPRINT]
- Kouremetis, M., Dawson, A., Dheekonda, R. S. R., & Greunke, B. (2026). Every Model Cheats: Prompt-Level Mitigation of Cheating on Offensive Cyber Tasks. arXiv:2607.21763. doi:10.48550/

- p>arXiv:2607.21763. [PREPRINT]
- Li, J., & Ng, H. T. (2025). Reasoning Models Hallucinate More: Factuality-Aware Reinforcement Learning for Large Reasoning Models. arXiv:2505.24630. doi:10.48550/arXiv.2505.24630.
- Li, X., Yu, Z., Zhang, Z., Chen, X., Zhang, Z., Zhuang, Y., Sadagopan, N., & Beniwal, A. (2025). When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs. arXiv:2505.11423. doi:10.48550/arXiv.2505.11423. [PREPRINT]
- Liu, R., Summers, T. R., Dasgupta, I., & Griffiths, T. L. (2024). How do Large Language Models Navigate Conflicts between Honesty and Helpfulness? arXiv:2402.07282. doi:10.48550/arXiv.2402.07282.
- MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., Bouscal, N., Bowman, S., Bricken, T., Cloud, A., Denison, C., Gasteiger, J., et al. (2025). Natural Emergent Misalignment from Reward Hacking in Production RL. arXiv:2511.18397. doi:10.48550/arXiv.2511.18397. [PREPRINT]
- Mahendrakar, P. (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. (2026b). Two Kinds of Missing: Underspecified Inputs Versus Unknown Answers, and Why One Abstention Policy Cannot Serve Both. Life of Research. doi:10.5281/zenodo.22901280
- Mahendrakar, P. (2026c). Not Acting Is Not One Decision: Two Abstention Triggers in Tool-Using Agents, Why the Calibration and Permission Readings Cover Different Ones, and the Cost Term Neither Supplies. Life of Research. doi:10.5281/zenodo.22803148
- Mahendrakar, P. (2026d). When Deliberation Hurts: Inverse Test-Time Scaling, Unfaithful Traces, and the Case Against a Unified System-2 in LLM Reasoning. Life of Research. doi:10.5281/zenodo.22904981
- Niimi, J. (2025). Hierarchical Memorization in Large Language Models: Evidence from Citation Generation. arXiv:2511.08877. doi:10.48550/arXiv.2511.08877. [PREPRINT]
- Niimi, J. (2026). Distortion Instead of Hallucination: The Effect of Reasoning Under Strict Constraints. arXiv:2601.01490. doi:10.48550/arXiv.2601.01490. [PREPRINT]
- Nishimura-Gasparian, K., McCarthy, R., & Lindner, D. (2026). Towards Understanding Specification Gaming in Reasoning Models. arXiv:2605.02269. doi:10.48550/arXiv.2605.02269. [PREPRINT]
- OpenAI. (2024). OpenAI o1 System Card. arXiv:2412.16720. doi:10.48550/arXiv.2412.16720.
- Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., Evans, O., & Brauner, J. (2023). How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. arXiv:2309.15840. doi:10.48550/arXiv.2309.15840.
- Rao, D., Wong, E., & Callison-Burch, C. (2026). Detecting and Correcting Reference Hallucinations in Commercial LLMs and Deep Research Agents. arXiv:2604.03173. doi:10.48550/arXiv.2604.03173. [PREPRINT]
- Ren, R., Agarwal, A., Mazeika, M., Menghini, C., Vacareanu, R., Kenstler, B., Yang, M., Barras, I., Gatti, A., Yin, X., Trevino, E., Geralnik, M., et al. (2025). The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems. arXiv:2503.03750. doi:10.48550/arXiv.2503.03750.
- Russinovich, M., Siva Kumar, R. S., & Salem, A. (2026). Phantom References: Hallucinated Citations That Survive Peer Review at Top-Tier Conferences. arXiv:2607.00738. doi:10.48550/arXiv.2607.00738. [PREPRINT]
- Sakai, Y., Kamigaito, H., & Watanabe, T. (2026). HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences. arXiv:2601.18724.
- doi:10.48550/arXiv.2601.18724. [PREPRINT]
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590. doi:10.48550/arXiv.2311.07590.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548. doi:10.48550/arXiv.2310.13548.
- Song, L., Shi, T., & Zhao, J. (2025). The Hallucination Tax of Reinforcement Finetuning. arXiv:2505.13988. doi:10.48550/arXiv.2505.13988. [PREPRINT]
- Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-y., & Chen, Y.-N. (2024). Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. arXiv:2408.02442. doi:10.48550/arXiv.2408.02442.
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13. doi:10.1038/s41598-023-41032-5.
- Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024a). Measuring short-form factuality in large language models. arXiv:2411.04368. doi:10.48550/arXiv.2411.04368. [PREPRINT]
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903. doi:10.48550/arXiv.2201.11903.
- Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., & Le, Q. V. (2024b). Long-form factuality in large language models. arXiv:2403.18802. doi:10.48550/arXiv.2403.18802.
- Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., & Feng, S. (2024). Language Models Learn to Mislead Humans via RLHF. arXiv:2409.12822. doi:10.48550/arXiv.2409.12822.
- Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., Li, J., & Chua, T.-S. (2025). Are Reasoning Models More Prone to Hallucination? arXiv:2505.23646. doi:10.48550/arXiv.2505.23646. [PREPRINT]
- Yin, C., Sha, Z., Cui, S., Meng, C., & Li, Z. (2025). The Reasoning Trap: How Enhancing LLM Reasoning Amplifies Tool Hallucination. arXiv:2510.22977. doi:10.48550/arXiv.2510.22977.
- Yoon, D., Kim, S., Yang, S., Kim, S., Kim, S., Kim, Y., Choi, E., Kim, Y., & Seo, M. (2025). Reasoning Models Better Express Their Confidence. arXiv:2505.14489. doi:10.48550/arXiv.2505.14489.
- Zhang, H., Diao, S., Lin, Y., Fung, Y. R., Lian, Q., Wang, X., Chen, Y., Ji, H., & Zhang, T. (2024). R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. arXiv:2311.09677. doi:10.48550/arXiv.2311.09677.
- 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.
- Zhong, Z., Raghunathan, A., & Carlini, N. (2025). ImpossibleBench: Measuring LLMs' Propensity of Exploiting Test Cases. arXiv:2510.20270. doi:10.48550/arXiv.2510.20270. [PREPRINT]
- Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-Following Evaluation for Large Language Models. arXiv:2311.07911. doi:10.48550/arXiv.2311.07911.
- Zolkowski, A., Xing, W., Lindner, D., Tramer, F., & Jenner, E. (2025). Can Reasoning Models Obfuscate Reasoning? Stress-Testing Chain-of-Thought Monitorability. arXiv:2510.19851. doi:10.48550/arXiv.2510.19851. [PREPRINT]