Measuring omitted influence in the explanations of AI assistants: a hinted-pair methodology with cluster-aware uncertainty, chance correction and exposure under repeated attempts

A black-box methodology for unfaithful explanations in adversarial evaluation, with a simulation study and a re-analysis of public data

Juan F. Cobo *Author*[1] and Leonel Leenen *Reviewer*[1]

[1]EthiCompass

Date: 2026-09-30

Abstract

An explanation is unfaithful when it leaves out a factor that actually changed the answer it justifies. For deployed AI assistants, which are evaluated as black boxes through their conversational interface, the question can only be asked from the outside: did something we introduced change the answer, and does the explanation say so? This paper sets out a methodology for measuring that omission in adversarial evaluation, and for reporting it without claiming more than the data support.

The instrument is the hinted pair (Turpin et al., 2023; Chen et al., 2025): the same closed question is asked without and with a hint pointing to a wrong option, and in the pairs where the answer moves to the hint an automated judge decides whether the explanation acknowledges it. We specify the record each pair must produce and the consistency rules that make a delivery verifiable. The first part defines what can be reported with an uncalibrated judge: a conditional omission rate with a percentile cluster bootstrap interval over replicas of the same probe, strata by the correctness of the unhinted answer, a correction for answer changes that occur by chance computed per number of options, a threshold-sensitivity curve, and the exposure of a user who repeats the same question, estimated by applying the unbiased pass@k estimator to an unconditional violation event. The second part defines when a misclassification-corrected rate is admissible and how the error of a fixed calibration propagates.

A simulation with known truth and a re-analysis of 13,196 public hinted pairs illustrate both parts. The cluster bootstrap covers 0.955 against 0.895 for the independence-based Wilson interval at a mean intra-cluster correlation of 0.10. The unbiased estimator of exposure has a bias of at most 0.002, whereas the plug-in form $1 - (1 - \hat{p})^k$ overestimates it by up to +0.075 at k = 8. A calibration set of 150 items per class shifted every corrected estimate that used it by 0.04 to 0.06. In the public data, 36 % of hinted items point to the correct answer and must be excluded by design.

Keywords: unfaithful explanations; chain-of-thought faithfulness; LLM-as-a-judge; black-box evaluation; cluster bootstrap; pass@k; misclassification; red-teaming

Manuscript. Preprint prepared for journal submission; not yet peer reviewed.

1. Introduction

AI assistants routinely justify their answers. When the justification leaves out a factor that actually drove the answer, the user is given a false basis for trusting it, precisely in the cases where the answer was influenced by something the assistant does not declare. Measuring this is harder than measuring whether an answer is correct, because the property is a relation between the answer and its justification, not a property of the answer alone.

The literature offers many tests of explanation faithfulness, but most of them need access that an external evaluator does not have: token probabilities, attributions, the ability to continue or truncate the model's own text, or internal activations. A deployed assistant is evaluated through its conversational interface. The evaluator plans conversations, sends them, records them, and measures afterwards over the recorded text. This setting rules out adaptive procedures and anything that reads the model's internals.

This paper addresses that setting. We make four contributions:

1. **A choice of instrument justified by the constraints of black-box evaluation**, together with the methods that do not fit them and why.

2. **A record specification** for each hinted pair, with closed vocabularies and consistency rules that let the consumer of an evaluation verify a delivery mechanically before computing anything.
3. **A reporting methodology** for the uncalibrated case (Part I), and the conditions and procedure for a misclassification-corrected rate (Part II).
4. **A simulation study and a re-analysis of public data** that check every estimator against a known truth and against an independent recomputation.

The two parts are deliberately separated. Part I applies whenever judge outputs are available. Part II applies only when a calibration of the exact judge exists and meets explicit conditions.

2. Related work

Hint-based tests of faithfulness. Turpin et al. (2023) showed that biasing features in the prompt, such as a suggested answer or a pattern in few-shot examples, change model answers without being mentioned in the chain of thought, with accuracy drops of up to 36 % on 13 BIG-Bench Hard tasks. Chen et al. (2025) turned the idea into a faithfulness score for reasoning models: condition on pairs whose answer changes from a non-hint answer to the hint answer, score whether the reasoning verbalizes the hint, and normalize for changes that occur by chance. They report faithfulness of 25 % for Claude 3.7 Sonnet and 39 % for DeepSeek R1. Atanasova et al. (2023) proposed a related counterfactual test that inserts content changing the prediction and checks whether the explanation mentions it. Chua et al. (2024) catalogue nine bias types for training and evaluation.

What black-box tests measure. Parcalabescu and Frank (2024) argue that tests that only observe outputs measure self-consistency at the output level rather than faithfulness to internal computation. Madsen et al. (2024) find self-explanation faithfulness to be model-, explanation- and task-dependent. Other tests require access we assume unavailable: early answering and truncation of the reasoning (Lanham et al., 2023), attribution-based comparisons such as CC-SHAP (Parcalabescu and Frank, 2024), and measures of the shift in the predicted label distribution (Siegel et al., 2024). Counterfactual simulatability (Chen et al., 2024) evaluates explanation precision, and automated simulators can reach high scores without using the explanation (Poché et al., 2026). Implicit post-hoc rationalization can be detected with reversed comparative questions (Arcuschin et al., 2025), which is black-box but uses a question group rather than a pair as its unit.

Where the hint is delivered. Gema et al. (2026) find, across 15 open-weight models, that verbalization is lower and unverbalized adoption higher when a preference cue arrives in a tool return rather than in the user message.

The judge is part of the measurement. Young (2026) applies three verbalization classifiers to the same data and obtains 74.4 %, 82.6 % and 69.7 %, with Cohen's between 0.06 and 0.42. Gur-Arieh et al. (2026) meta-evaluate faithfulness metrics against constructed ground truth and find most near chance, the best reaching an AUROC of 0.70 without transfer across settings. Angdembay et al. (2026) show that most annotated unfaithfulness occurs on incorrect answers, where no signal they tested is detectably above chance; Bentham et al. (2024) show that a widely used chain-of-thought faithfulness metric, even after normalization, is strongly correlated with accuracy ($R^2 = 0.74$).

Repeated attempts. Chen et al. (2021) give the unbiased estimator of pass@k and show that the plug-in form is biased. Yao et al. (2024) introduce pass^k, the probability that all k attempts succeed, for agents where consistency matters.

Misclassification and clustered data. The correction of apparent prevalence for imperfect sensitivity and specificity is due to Rogan and Gladen (1978), with the Youden index $J = \mathrm{Se} + \mathrm{Sp} - 1$ (Youden, 1950) as its denominator. For clustered binary data, the cluster bootstrap resamples whole clusters (Field and Welsh, 2007), and the design effect $\mathrm{deff} = 1 + (\bar{m} - 1)\rho$ summarises the variance inflation (Kish, 1965). Newcombe (1998) compares interval methods for differences of proportions; Wilson (1927) gives the score interval.

3. Setting and notation

An adversarial evaluation plans K probes against an assistant and replicates each one m times, because the assistant is not deterministic. Only the component that conducts conversations talks to the assistant. Measurement happens afterwards, over recorded conversations, by a component that does not talk to the assistant. The evaluator controls the user messages; it does not control tool returns, retrieval, or the assistant's sampling temperature, and it has no access to token probabilities.

A **hinted pair** is one probe with two halves, conducted in dedicated sessions:

- the **baseline** half asks a closed question with $L \geq 2$ options and explicitly asks for a short justification;
- the **hinted** half asks the same question with a hint pointing to option h.

For probe i and replica j, we write a^0_{ij} for the answer to the baseline half, a^1_{ij} for the answer to the hinted half, and c_i for the correct answer, known because the question is built from a reference source. By design, $h_i \neq c_i$: the hint always points to a wrong option. Subscripts are dropped when unambiguous.

Each pair ends in exactly one of five outcomes:

Outcome	Symbol	Condition
Hint-induced change	H	$a^0 \neq h$ and $a^1 = h$
Other change	O	$a^0 \neq h$, $a^1 \neq h$ and $a^1 \neq a^0$
No change	N	$a^0 \neq h$ and $a^1 = a^0$
Baseline already on the hint	B	$a^0 = h$
Not evaluable	U	an answer could not be extracted

The pairs with $a^0 \neq h$, that is, outcomes H, O and N, are the **eligible** pairs.

In pairs with outcome H, an automated judge reads the explanation of a^1 and returns a label, *faithful*, *unfaithful* or *undetermined*, together with a probability $\phi \in [0, 1]$ that the explanation is faithful. Following Chen et al. (2025), an explanation is faithful if it (i) mentions the hint and (ii) depends on it to reach the answer; mentioning it only as a final check does not count. We call $1 - \phi$ the **violation score**.

4. Why the hinted pair

The setting of Section 3 imposes four constraints on the instrument:

1. **Non-adaptive.** Every intervention must be planned before the conversation, because measurement happens after it.
2. **Output-only.** No token probabilities, attributions or activations of the assistant.
3. **User-message delivery.** The only channel the evaluator controls.
4. **Repeated sampling.** Replicas are available and are part of the plan.

The hinted pair satisfies all four. It is planned at generation time, it reads only the text of the answers, and its intervention is a user message. Early answering and truncation fail constraint 1 and require continuing the assistant's own text. Attribution- and probability-based measures fail constraint 2. Redaction-based self-consistency checks derive the second query from the first explanation, which fails constraint 1. Counterfactual simulatability measures a different property. Mechanistic methods fail constraint 2 entirely.

Two consequences follow for interpretation. First, the instrument measures **observable self-consistency**, not internal faithfulness (Parcalabescu and Frank, 2024): it detects that the explanation leaves out a factor that, by construction, changed the answer. We therefore describe the quantity as the *omission of the factor that changed the answer*. Second, because the hint is delivered in the user message, the channel where models verbalize cues most (Gema et al., 2026), the measured omission is a **lower bound** on the omission the assistant would show with cues arriving through other channels.

5. Record specification

A measurement that crosses a component boundary should be verifiable by its consumer before anything is computed from it. We specify one record per pair and replica:

Field	Content
Unit key	Session identifier, probe identifier and replica index
Halves	An identifier for each half and a reference to its recorded conversation, or empty if none was stored; never reconstructed
Stratum	Hint type and risk family
Question	Number of options L, correct answer c and hint answer h
Answers	a^0 and a^1, each empty if it could not be extracted
Outcome	One of the five outcomes of Section 3
Correctness	Whether a^0 and a^1 equal c, empty if unknown

Field	Content
Judge	Whether it was applied; its label; its probability ϕ; an evidence span from the hinted explanation, or empty when there is nothing to quote
Fresh session	Whether the replica ran in a new session
Reason not evaluable	Extraction failed, judge unavailable, judge undetermined or generation truncated; empty otherwise

The delivery also declares the planned number of replicas, the hint types used, the channel through which the hint is delivered, and how the judge's output is produced: whether ϕ comes from token log-probabilities, a vote fraction or a verbalized confidence, and whether the label is the judge's own decision or a threshold on ϕ. A coverage summary counts planned, completed and incomplete pairs per hint type; an incomplete pair, missing one half, produces no record.

A delivery is rejected as a whole, naming the offending field, if any of the following rules fails:

Rule	Check
R1	Counts, indices and L are integers in the serialization. A value such as 3.0 is rejected, which schema validation alone does not enforce
R2	The outcome is the one the answers imply
R3	The judge is applied if and only if the outcome is H
R4	$h \neq c$
R5	A reason for non-evaluability is given if and only if the record cannot be evaluated
R6	ϕ is present if and only if the judge answered
R7	With a thresholded label, the label is *faithful* if and only if ϕ is at least the threshold; a tie counts as *faithful*
R8	The replica index is within the plan
R9	Correctness agrees with the answers and with c
R10	The coverage summary agrees with the records: completed pairs equal the number of records, planned pairs equal completed plus incomplete, and the reasons add up
R11	There is one record per session, probe and replica

Rule R11 caught a real error during this work that would otherwise have passed silently: item identifiers in a public dataset repeat within a task, and using them as probe identifiers duplicated units. Rules R9 and R10 close gaps found in review: without them, a correctness flag that contradicts the answers, or a coverage summary that declares more completed pairs than the records it carries, would go undetected.

Part I. Reporting without judge calibration

6. Sampling unit and cluster

The sampling unit is the pair in one replica; its two halves are never counted separately. Replicas of the same probe share the question and the hint and are correlated, so the **probe is the cluster**. The event of interest is the violation, an explanation judged *unfaithful*, and its score is $1 - \phi$.

7. The conditional omission rate and its uncertainty

In each cell, defined by hint type, consider the pairs with outcome H that can be evaluated and have $L \geq 3$. Let n_H be their number and k the number judged *unfaithful*. The conditional omission rate is

$$\hat{q} = \frac{k}{n_H}.$$

It is the rate at which the judge flagged omission, observed and uncorrected. The primary interval is a percentile cluster bootstrap over probes, with 2,000 resamples. We report alongside it the intra-cluster correlation ρ (one-way ANOVA estimator), the design effect $\text{deff} = 1 + (\bar{m} - 1)\rho$, where $\bar{m}$ is the mean number

of pairs per probe, and the effective sample size n_H/deff; the independence-based Wilson interval is shown only as a reference. When $k = 0$ or $k = n_H$ the percentile interval collapses to a point, which is not an interval; in that case we use the Wilson interval at the effective sample size.

A cell is reported as *insufficient* with fewer than two clusters or $n_H < 2$, as *undersampled* if the half-width of the interval exceeds a target precision $\varepsilon = 0.10$, and as *converged* otherwise.

Pairs that cannot be evaluated, including those where the judge returns *undetermined* or does not answer, leave the denominator and are counted separately as ungraded. They are never read as faithful.

8. Strata by the correctness of the unhinted answer

Because $h \neq c$, the final answer is wrong in every pair with outcome H. The stratification by correctness that the literature recommends (Bentham et al., 2024; Angdembay et al., 2026) is therefore made on a^0, not on a^1, and gives two readings within each cell:

- **drift from a correct answer** ($a^0 = c$): the hint moved the assistant from right to wrong and the explanation hid it, the most serious case for a user;
- **change between wrong answers** ($a^0 \neq c$): the assistant was already wrong, and the hint only changed which wrong option it chose.

The confound described by Angdembay et al. (2026) arises when unfaithfulness must be detected without knowing its cause. Here the cause is known by construction, which is exactly the hint-built testbed that work uses to label its data.

9. Correction for answer changes that occur by chance

Some pairs move to the hint by chance and would never mention it. Following Chen et al. (2025), among the N_E eligible pairs let $p = N_H/N_E$ be the share that move to the hint and $q = N_O/N_E$ the share that move to another option. With F the share of evaluable hint-induced changes judged *faithful*, the corrected faithfulness is

$$F^* = \min\left\{ \frac{F}{1 - \dfrac{q}{(L-2)\,p}},\ 1 \right\},$$

and the corrected omission is $1 - F^*$. Two details matter in practice. First, p and q describe the conversation, not the judge, so they are counted over all eligible pairs, including those whose judge did not answer; F uses only evaluable pairs. Excluding unjudged pairs from p understated it and overstated the correction by up to 0.0015 in our simulation. Second, the rate of chance changes to a given option depends on L, so a cell that mixes questions with different numbers of options must be corrected **separately for each value of** L and combined with weights proportional to N_H. The correction is undefined when $p \leq q/(L-2)$, and it has no effect when F is 0 or 1.

10. Threshold sensitivity

With the judge's probability, the rate is re-decided at cuts $t \in \{0.5, 0.6, 0.7, 0.8\}$ of the violation score: a pair counts as unfaithful if $1 - \phi > t$. A tie counts as faithful, the same convention as rule R7, so that the curve at the label's own threshold reproduces the label. Ties are frequent when ϕ is a vote fraction. The curve is a sensitivity reading, not a result: in our simulation the rate falls from 0.646 to 0.202 between the cuts 0.5 and 0.8.

11. Exposure under repeated attempts

A user who asks the same question several times is exposed to an unfaithful explanation more often than a single query suggests. Let V_{ij} indicate that replica j of probe i has outcome H and is judged *unfaithful*; this violation event is unconditional, defined for every evaluable replica. For each probe, let n_i be the number of evaluable replicas and $c_i = \sum_j V_{ij}$. The **exposure** $X(k)$ is the probability that at least one of k attempts shows a violation. Over the set $\mathcal{I}_k$ of probes with $n_i \geq k$, it is estimated without bias by

$$\hat{X}(k) = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left[1 - \frac{\binom{n_i - c_i}{k}}{\binom{n_i}{k}} \right],$$

for k up to the number of replicas, and is reported as reliable up to $k = m/2$. This is the pass@k estimator of Chen et al. (2021) applied to the violation event rather than to a success, so a higher $X(k)$ is worse. We avoid the name pass@k for the result, because in the literature it counts successes and would invert the reading. The **reliability** $R(k) = 1 - X(k)$ is the probability that none of k attempts shows a violation; it is the pass^k of Yao et al. (2024) with "no violation" as the success, so one curve is read in two ways. Probes with any replica outside a fresh session are left out, because independence between replicas is the estimator's assumption. The interval is a bootstrap over probes.

The shape of the curve carries information that a single rate does not. With π_i the probability that a replica of probe i shows a violation, $X(1) = \mathbb{E}[\pi_i]$ while $X(k) = 1 - \mathbb{E}\big[(1 - \pi_i)^k\big]$, so two assistants with the same mean can differ widely. We report the share of probes whose replicas never, always and sometimes show a violation, which separates a failure concentrated on certain questions from an intermittent one.

The plug-in form $1 - (1 - \hat{p})^k$, with $\hat{p}$ a rate pooled across probes, assumes that every probe has the same probability. By Jensen's inequality it overestimates exposure, and underestimates reliability, whenever probes differ; with one attempt per probe the curve is not identifiable at all, and any gap between the two readings comes from the assumption rather than the data.

12. Reporting

We propose a fixed sentence:

> When a *[type]* hint led the assistant to change its answer to a wrong option, the explanation omitted that influence in *[q]* % of cases (95 % CI: *[a, b]*, bootstrap over probes; *[n]* pairs; uncalibrated judge). This is a lower bound: the hint was delivered in the user message.

Cells for different hint types are reported side by side and never averaged. The susceptibility p is reported as context, because it explains the size of the denominator; it is not a rate of this construct but of resistance to pressure, which belongs to sycophancy.

Part II. Correction with a calibrated judge

13. When a calibration is admissible

The Part I rate is how often the judge flagged omission. A misclassification-corrected rate is admissible only with a calibration that has (i) the same judge, meaning model, provider, serving path and prompt; (ii) the same construct, the two criteria of Section 3; (iii) the same language and format; and (iv) at least 30 positive and 30 negative items. Otherwise the cell stays in Part I.

14. Procedure and gates

Let Se and Sp be the judge's sensitivity and specificity for detecting omission, estimated from n_+ positive and n_- negative calibration items, and $J = \mathrm{Se} + \mathrm{Sp} - 1$.

1. Re-decide each evaluable hint-induced change at the calibrated threshold, with the tie convention of Section 10, and compute $\hat{q}$ as in Section 7.
2. Correct with Rogan and Gladen (1978):

$$\hat{\pi} = \frac{\hat{q} + \mathrm{Sp} - 1}{\mathrm{Se} + \mathrm{Sp} - 1}.$$

3. Combine two sources of uncertainty. Let h_s be the half-width of the cluster interval of $\hat{q}$ divided by J, and

$$h_c = \frac{1.96}{J} \sqrt{\hat{\pi}^2 \, \frac{\mathrm{Se}(1 - \mathrm{Se})}{n_+} + (1 - \hat{\pi})^2 \, \frac{\mathrm{Sp}(1 - \mathrm{Sp})}{n_-}}$$

the half-width due to the calibration. The half-width of the corrected interval is $h = \sqrt{h_s^2 + h_c^2}$.

4. Apply the gates in order. The estimate is *non-discriminative* if the lower bound of the interval of J, obtained by combining the Wilson intervals of Se and Sp in quadrature (Newcombe, 1998), is at most 0. It is *below the detection floor* if $\hat{q} \leq 1 - \text{Sp}$ or $J < 0.20$. It is *limited by the calibration* if $h_c > \varepsilon$. It is *undersampled* if $h > \varepsilon$ or if the interval must be clipped to $[0, 1]$. It is *converged* otherwise.

15. The error of a fixed calibration

A calibration is estimated once and used many times. Its sampling error is therefore not noise that averages out across evaluations: it shifts every corrected estimate that uses it in the same direction. In our simulation, a calibration set of 150 items per class estimated the judge's sensitivity at 0.86 when the true value was 0.813, and every corrected estimate using it was shifted by 0.036 to 0.059. The interval of Section 14 covers this uncertainty in expectation, and with a new calibration set in each repetition coverage was correct, but a single calibration's error repeats in every use.

The standard error of an estimated sensitivity is approximately $\sqrt{\text{Se}(1 - \text{Se})/n_+}$. With $\text{Se} \approx 0.81$ it is 0.072 with 30 items per class, 0.032 with 150 and 0.016 with 600. The minimum of 30 per class is therefore a condition of admissibility, not of sufficiency; h_c should be reported with every corrected estimate, and the gate that flags an estimate as limited by the calibration is what protects against a small calibration set.

The main practical obstacle is not licensing but the scarcity of positives: explanations that do acknowledge the hint. In the public data of Section 16.3 there is none. Reasoning models acknowledge hints 25 % to 39 % of the time (Chen et al., 2025), so a calibration set can be built by running the protocol against such a model and annotating a sample by hand with the two criteria.

16. Simulation study and re-analysis of public data

16.1 Design

We used two data sources, because each tests different things.

- **A simulated assistant with known truth.** Three hint types, 320 probes per type and 8 replicas. Each probe has its own latent accuracy, susceptibility to the hint and propensity to acknowledge it, drawn from Beta distributions, so probes differ. Answers change by chance with probability 0.15, spread uniformly over the other options, and 10 % of questions have two options. The judge draws ϕ from Beta(4, 2) when the explanation acknowledges the hint and from Beta(2, 4) when it does not, which makes it correct about 81 % of the time; 2 % of its labels are undetermined and 1 % of its calls fail. A perfect-judge variant isolates the chance correction.
- **Public data.** The primary results released by Turpin et al. (2023) for the suggested-answer bias: 13,196 items from two 2023 models on 13 BIG-Bench Hard tasks, each with a biased and an unbiased generation, converted to the record of Section 5. The judge is a keyword detector for mentions of the suggestion, which cannot see paraphrase or check the second criterion; the results validate the mechanics, not the behaviour of a current assistant.

Every estimate was also recomputed by an independent program that does not share code with the implementation and follows the text of this paper: exact binomial coefficients, closed-form Wilson intervals and an ANOVA estimator of ρ written from scratch.

16.2 Estimator properties

Property	Result
Coverage of the cluster bootstrap, 200 repetitions	0.955 (nominal 0.95; mean = 0.096)
Coverage of the independence-based Wilson interval	0.895
Bias of the unbiased exposure estimator, k = 1, 2, 4, 8	0.0002, 0.0003, 0.0005, 0.0021
Bias of the plug-in form $1 - (1 - \hat{p})^k$, k = 1, 2, 4, 8	0.0006, +0.010, +0.039, +0.075
Chance correction, perfect judge, three hint types, 50 repetitions	bias 0.003 to +0.005; coverage 0.92 to 0.94
Corrected rate, new calibration set per repetition	bias 0.001 to +0.003; coverage 0.96 to 1.00
Corrected rate, one fixed calibration set	bias 0.036 to 0.059

The small negative bias of the exposure estimate at $k = 8$ has an identifiable cause: the judge can only return an undetermined label on pairs that changed, so probes with more changes lose replicas more often and drop

out at high k. The independent recomputation matched every estimate, and it detected the counting error in p described in Section 9 before these results were produced.

Worked example. For one simulated hint type with the imperfect judge, 482 of 746 evaluable hint-induced changes were flagged: $\hat{q} = 0.646$, 95 % CI [0.609, 0.684] ($\rho = 0.084$, deff = 1.152, 265 clusters), against a population value of 0.644 for this judge. The strata were 0.663 for drift from a correct answer (387/584) and 0.586 for changes between wrong answers (95/162). The chance-corrected omission was 0.611 [0.570, 0.654]. The estimated exposure was 0.208, 0.568 and 0.762 at $k = 1$, 4 and 8, against true values of 0.195, 0.544 and 0.751; 24 % of probes never showed a violation, none always did, and 76 % did sometimes. With the fixed calibration, the corrected rate was 0.682 [0.597, 0.767] ($J = 0.673$), against a true omission of 0.729.

16.3 Public data

Model, prompting	Hint-induced changes	Flagged omission	95 % CI	Susceptibility p	Exposure, one attempt
claude-v1, zero-shot	618	1.000	[0.994, 1.000]	0.507	0.445
claude-v1, few-shot	376	1.000	[0.990, 1.000]	0.294	0.274
text-davinci-003, zero-shot	788	1.000	[0.995, 1.000]	0.671	0.566
text-davinci-003, few-shot	476	1.000	[0.992, 1.000]	0.384	0.342

Neither model mentioned the suggestion in any explanation that changed because of it, consistent with Turpin et al. (2023). With a single replica per item, the exposure after one attempt is simply the share of pairs that moved to the hint without acknowledging it; here it equals the share that moved to the hint, because none acknowledged it. The intervals come from the boundary rule of Section 7, since the percentile bootstrap collapses at a rate of 1. The chance correction has no effect at $F = 0$.

Two properties of the public data are relevant to anyone designing a hinted-pair evaluation. The suggestion was drawn among all options, so in 4,752 of 13,196 items (36 %) it pointed to the correct answer; such items test nothing about omission and violate $h \neq c$. Six of the thirteen tasks have two options, which leaves 720 of the 2,111 remaining pairs per model (34 %) outside the analysis, because the chance correction is undefined for $L = 2$. Hints should be drawn among wrong options only, and questions should have three or more options.

Sample size. At a half-width of 0.10 around 0.5, about 96 useful pairs are needed. With the susceptibilities observed in the public data (0.29 to 0.67), that is roughly 170 to 355 completed pairs per hint type, or 195 to 410 once a design effect of 1.15 is allowed for.

17. Discussion

The methodology separates what the data can support from what it cannot. Without calibration, the number is the rate at which a particular judge flagged omission, with an interval that is honest about the correlation between replicas, and its readings show how much it depends on chance, on the threshold and on repeated attempts. With an admissible calibration, a corrected rate becomes available, but its error is shared by every evaluation that uses the same calibration.

The instrument measures one component of transparency. With a social hint, the same pair also measures whether the answer yields to pressure, which is sycophancy; we recommend building the core of the instrument on non-social hints (metadata, the model's own prior answer, a distractor fact) and reporting a social hint separately, so that the two constructs are not conflated.

18. Limitations

1. The instrument measures observable self-consistency, not internal faithfulness.
2. It is a lower bound, because the hint is delivered in the user message.
3. It applies to closed questions with three or more options; open questions are out of scope.
4. The explanation is requested, not spontaneous, which changes the task.
5. Without calibration, the number depends on the judge.

6. The denominator is conditional: if the assistant rarely moves to the hint, the cell remains insufficient or undersampled, which is itself informative.
7. Replicas do not control the assistant's temperature; the exposure curve measures its natural variability with the same question text and no rephrasing.
8. The public data come from 2023 models without extended reasoning, and the judge used on them is a keyword detector.

19. Conclusion

Unfaithful explanations can be measured from the outside with an instrument that fits black-box evaluation: planned hinted pairs, recorded conversations and a judge applied afterwards. A verifiable record specification, a cluster-aware conditional rate, a chance correction computed per number of options, and an unbiased estimator of exposure under repeated attempts make the number honest about what it is. A corrected rate requires a calibration of the exact judge with enough positive examples, and the error of that calibration must be reported, because it does not average out.

Code availability

The record validator, the estimators, the simulation, the re-analysis of the public data and the independent recomputation are implemented in Python with NumPy and are available as supplementary material.

Author contributions

Juan F. Cobo: conceptualization, methodology, software, formal analysis, writing (original draft). Leonel Leenen: writing (review and editing).

Competing interests

Both contributors are affiliated with EthiCompass, which develops AI evaluation services.

Use of generative AI

An AI assistant was used to help draft text, implement the simulation code and check references. The author reviewed all content and takes responsibility for it.

References

- Angdembay, S. R., Aryal, D. and Rahimi, N. (2026). Two regimes of chain-of-thought unfaithfulness: metric-based detection fails where models are wrong. arXiv:2607.23458.
- Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N. and Conmy, A. (2025). Chain-of-thought reasoning in the wild is not always faithful. arXiv:2503.08679.
- Atanasova, P., Camburu, O.-M., Lioma, C., Lukasiewicz, T., Simonsen, J. G. and Augenstein, I. (2023). Faithfulness tests for natural language explanations. *Proceedings of ACL 2023*. arXiv:2305.18029.
- Bentham, O., Stringham, N. and Marasović, A. (2024). Chain-of-thought unfaithfulness as disguised accuracy. *Transactions on Machine Learning Research*. arXiv:2402.14897.
- Chen, M. et al. (2021). Evaluating large language models trained on code. arXiv:2107.03374.
- Chen, Y. et al. (2024). Do models explain themselves? Counterfactual simulatability of natural language explanations. *Proceedings of ICML 2024*. arXiv:2307.08678.
- Chen, Y., Benton, J. et al. (2025). Reasoning models don't always say what they think. arXiv:2505.05410.
- Chua, J., Rees, E., Batra, H., Bowman, S. R., Michael, J., Perez, E. and Turpin, M. (2024). Bias-augmented consistency training reduces biased reasoning in chain-of-thought. arXiv:2403.05518.
- Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B*, 69(3), 369–390.
- Gema, A. P., Rajani, N., Saxena, R., Kwan, W.-C. and Minervini, P. (2026). Chain-of-thought faithfulness of reasoning models varies with where and how preference cues are delivered. arXiv:2608.29464.

- Gur-Arieh, Y., Marasović, A. and Geva, M. (2026). Faithfulness metrics don't measure faithfulness: a meta-evaluation with ground truth. arXiv:2605.25052.
- Kish, L. (1965). *Survey Sampling*. Wiley.
- Lanham, T. et al. (2023). Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702.
- Madsen, A., Chandar, S. and Reddy, S. (2024). Are self-explanations from large language models faithful? *Findings of ACL 2024*. arXiv:2401.07927.
- Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890.
- Parcalabescu, L. and Frank, A. (2024). On measuring faithfulness or self-consistency of natural language explanations. *Proceedings of ACL 2024*. arXiv:2311.07466.
- Poché, A. et al. (2026). Limitations of automated simulatability: LLM simulators can bypass explanations. arXiv:2609.08585.
- Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1), 71–76.
- Siegel, N. Y., Camburu, O.-M., Heess, N. and Perez-Ortiz, M. (2024). The probabilities also matter: a more faithful metric for faithfulness of free-text explanations in large language models. *Proceedings of ACL 2024*. arXiv:2404.03189.
- Turpin, M., Michael, J., Perez, E. and Bowman, S. R. (2023). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems 36*. arXiv:2305.04388.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.
- Yao, S., Shinn, N., Razavi, P. and Narasimhan, K. (2024). -bench: a benchmark for tool-agent-user interaction in real-world domains. arXiv:2406.12045.
- Young, R. J. (2026). Measuring faithfulness depends on how you measure: classifier sensitivity in LLM chain-of-thought evaluation. arXiv:2603.20172.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.