

---

# MEASURING FAITHFULNESS DEPENDS ON HOW YOU MEASURE: CLASSIFIER SENSITIVITY IN LLM CHAIN-OF-THOUGHT EVALUATION

---

Richard J. Young<sup>1,2</sup>

<sup>1</sup> University of Nevada, Las Vegas, Department of Management,  
Entrepreneurship and Technology, Lee Business School, Las Vegas, NV, USA

<sup>2</sup> DeepNeuro AI, Las Vegas, NV, USA  
ryoung@unlv.edu · richard@deepneuro.ai

## ABSTRACT

Recent work on chain-of-thought (CoT) faithfulness reports single aggregate numbers (e.g., DeepSeek-R1 acknowledges hints 39% of the time [1]), implying that faithfulness is an objective, measurable property of a model. This paper demonstrates that it is not. Three classifiers (a regex-only detector, a two-stage regex-plus-LLM pipeline, and an independent Claude Sonnet 4 judge) are applied to 10,276 influenced reasoning traces from 12 open-weight models spanning 9 families and 7B to 1T parameters. On identical data, these classifiers produce overall faithfulness rates of 74.4%, 82.6%, and 69.7%, respectively. Per-model gaps range from 2.6 to 30.6 percentage points. At the hint-type level, all pairwise McNemar tests are significant ( $p < 0.001$ ). The disagreements are systematic, not random: inter-classifier agreement measured by Cohen’s  $\kappa$  ranges from 0.06 (“slight”) for sycophancy hints to 0.42 (“moderate”) for grader hints, and the asymmetry is pronounced: for sycophancy, 883 cases are classified as faithful by the pipeline but unfaithful by the Sonnet judge, while only 2 go the other direction. Classifier choice can also reverse model rankings: Qwen3.5-27B ranks 1st under the pipeline but 7th under the Sonnet judge; OLMo-3.1-32B moves in the opposite direction, from 9th to 3rd. The root cause is that different classifiers operationalize related faithfulness constructs at different levels of stringency (lexical *mention* versus epistemic *dependence*), and these constructs yield divergent measurements on the same behavior. These results demonstrate that published faithfulness numbers cannot be meaningfully compared across studies that use different classifiers, and that future evaluations should report sensitivity ranges across multiple classification methodologies rather than single point estimates.

**Keywords** Chain-of-Thought · Faithfulness · Classifier Sensitivity · Reasoning Models · Evaluation Methodology

## 1 Introduction

Is DeepSeek-R1 39% faithful or 95%? Chen et al. [1] report that DeepSeek-R1 acknowledges sycophantic hints in its chain-of-thought only 39% of the time. A regex-and-LLM pipeline, applied to the same model on comparable questions, yields 94.8%. An independent Claude Sonnet 4 judge returns 74.8%. Although these three evaluations differ in prompt design, model version, and classification methodology, making direct numerical comparison imprecise, the magnitude of the spread illustrates how sensitive faithfulness estimates are to evaluation choices. Which number is right?

Chain-of-thought prompting [2] has become the dominant paradigm for eliciting reasoning from large language models, and a growing literature treats measured faithfulness rates as objective properties of models [3, 4, 1]. Papers report single numbers and compare them across models, drawing conclusions about which architectures or training methods produce more faithful reasoning [5, 6]. Yet these numbers are products of two things: the model’s actual behavior *and* the classifier used to judge that behavior. When classifiers disagree substantially, faithfulness comparisons become classifier comparisons in disguise.

This problem has well-established precedents in the measurement sciences. Jacovi and Goldberg [7] argued that faithfulness itself admits multiple valid definitions, and Parcalabescu and Frank [8] showed that different operationalizations yield divergent measurements. More broadly, Gordon et al. [9] demonstrated that correcting for annotator disagreement in toxicity classification reduced apparent performance from 0.95 to 0.73 ROC AUC, Plank [10] argued that such label variation is an inherent property of subjective tasks rather than noise, and Bean et al. [11] found systematic construct validity failures across 445 LLM benchmarks. The use of LLMs as automated judges [12] introduces additional classifier-specific biases that have been catalogued by Gu et al. [13] and Ye et al. [14]. Despite these known concerns, CoT faithfulness studies routinely report single-classifier results without sensitivity analysis.

This paper quantifies the problem. Three classifiers (a regex-only detector, a two-stage regex-plus-LLM pipeline, and an independent Claude Sonnet 4 judge) are applied to 10,276 identical cases where 12 open-weight reasoning models [15] changed their answers in response to injected hints. Three results challenge the practice of reporting single faithfulness numbers:

1. **Large aggregate disagreement.** Overall faithfulness rates range from 74.4% (regex-only) through 82.6% (pipeline) to 69.7% (Sonnet judge). Per-model gaps range from 2.6 to 30.6 percentage points; at the hint-type level, all McNemar tests are significant ( $p < 0.001$ ).
2. **Hint-type-specific disagreement.** The gap is not uniform. Sycophancy hints [16] show a 43.4-percentage-point classifier gap ( $\kappa = 0.06$ , “slight” agreement), while grader hints show only 2.9 percentage points ( $\kappa = 0.42$ , “moderate” agreement). Classifier choice dominates the measurement for some hint types and is negligible for others.
3. **Asymmetric confusion structure.** The disagreements are directional, not random. For sycophancy, 883 cases are classified as faithful by the pipeline but unfaithful by the Sonnet judge, while only 2 go the other direction. The classifiers operationalize related faithfulness constructs at different levels of stringency: lexical *mention* versus epistemic *dependence*.
4. **Ranking reversals.** Qwen3.5-27B [17] ranks 1st by the pipeline (98.9% faithful) but 7th by Sonnet (68.3%); OLMo-3.1-32B [18] moves in the opposite direction, from 9th to 3rd ( $\rho = 0.67$ ,  $p = 0.017$ ). Conclusions about which models are “most faithful” depend on the measurement instrument.

The contribution is methodological rather than empirical: no single classifier is claimed to be correct. Rather, the field needs to treat classifier sensitivity as a first-class concern. Just as Gordon et al. [9] showed that ignoring annotator disagreement inflates reported performance, and Plank [10] argued that label variation should be modeled rather than suppressed, the present results show that faithfulness numbers are partly properties of the measurement instrument, not solely of the model under evaluation. Future work should report faithfulness ranges across multiple classifiers, disclose the exact classification methodology (including full prompts for LLM-based judges), and exercise caution when comparing numbers across papers that use different evaluation approaches.

## 2 Related Work

Faithfulness of chain-of-thought reasoning lacks a single agreed-upon definition. Jacovi and Goldberg [7] argued that faithfulness admits multiple valid formalizations depending on the intended guarantee, and Parcalabescu and Frank [8] showed that different operationalizations of the same concept yield divergent measurements in practice. Tanneru et al. [19] provided theoretical and empirical evidence that producing and verifying faithful CoT is inherently difficult, while Shen et al. [20] benchmarked instance-level faithfulness across tasks, confirming that aggregate metrics can obscure substantial variation. Bean et al. [11] conducted a systematic review of 445 LLM benchmarks and found widespread patterns that undermine construct validity, providing a broader framework for the measurement concerns raised here.

Empirical studies have probed CoT faithfulness through a variety of interventions. Turpin et al. [3] introduced biasing features into prompts and found that models rarely mention them in their reasoning, even when those features alter the final answer. Lanham et al. [4] applied causal interventions to measure the degree to which intermediate reasoning steps actually influence model outputs. More recently, Chen et al. [1] tested hint acknowledgment in reasoning models, reporting a 39% acknowledgment rate for DeepSeek-R1. Feng et al. [5] compared reasoning and non-reasoning models under similar protocols, Balasubramanian et al. [21] found that subtle biases are rarely articulated even in reasoning-specialized models, and Yee et al. [22] showed that distinct mechanisms drive faithful versus unfaithful error recovery, implying that classifier disagreement may reflect genuine ambiguity in the underlying phenomenon. Chua et al. [23] proposed an unlearning-based approach that sidesteps reliance on textual explanations altogether.

A parallel line of work has examined LLM-based evaluation methodology. Zheng et al. [12] established the LLM-as-judge paradigm, demonstrating strong agreement with human annotators for many tasks. However, subsequent work

has identified systematic limitations: Gu et al. [13] surveyed the growing LLM-as-judge literature and catalogued consistency and bias concerns, while Ye et al. [14] identified 12 specific bias types in LLM judges and showed that significant biases persist even in advanced models. Meek et al. [24] drew a useful distinction between monitorability and faithfulness, noting that the two properties can diverge when reasoning traces are evaluated by external models.

More broadly, the annotation disagreement literature has established that label variation is not merely noise but a fundamental property of subjective classification tasks. Plank [10] argued that human label variation should be modeled rather than suppressed, and Gordon et al. [9] demonstrated that annotator disagreement can cause standard metrics to dramatically overstate classifier performance. Despite this growing body of work, no prior study has systematically applied multiple classifiers to the same faithfulness dataset and quantified how measurement methodology shapes the resulting conclusions. The present paper addresses that gap.

### 3 Experimental Setup

**Data.** The dataset is drawn from a larger study of CoT faithfulness in open-weight reasoning models. It consists of 498 questions (300 from MMLU [25] and 198 from GPQA Diamond [26]), each paired with five hint types: sycophancy, consistency, metadata, grader, and unethical, the same categories used by Chen et al. [1]. Each hint steers the model toward a specific incorrect answer. Twelve open-weight reasoning models spanning 7B to 1T parameters across 9 model families are tested, including DeepSeek-R1 [15], OLMo [18], and Qwen-family models [17]. From the full set of hinted runs, the 10,276 *influenced* cases, instances where the model changed its answer from the baseline to match the hinted answer, are extracted, and all three classifiers are applied to this identical set.

**Classifiers.** Three approaches to determining whether a model’s chain-of-thought *acknowledges* the hint as a reason for its answer are compared:

1. **Regex-only.** The first stage of the pipeline in isolation: 38 regular expression patterns spanning all five hint types, designed for high precision. A match means the CoT explicitly references the hint (e.g., “the professor suggested,” “according to the metadata”). Free and instantaneous, but limited to surface-level lexical matching.
2. **Regex + Ollama judges** (full pipeline). Stage 1 regex as above, followed by Stage 2: a three-judge LLM panel (GLM-5, Kimi K2, Gemini 3 Flash) running locally via Ollama. Cases not caught by regex receive majority-vote classification. Effectively zero marginal cost.
3. **Claude Sonnet 4** (independent judge). Applied to all 10,276 cases via the Anthropic Batch API, following the LLM-as-judge paradigm [12]. Intended to assess whether the hint served as a *reason* for the model’s answer, not merely mentioned in passing. Cost: \$48.99 for the full dataset. Binary YES/NO output.

Additionally, the 39% acknowledgment rate reported by Chen et al. [1] for DeepSeek-R1 serves as a fourth data point, noting that their evaluation used a different prompt, a different model version, and a Sonnet-based classifier.

Figure 1 illustrates the experimental design: the same 10,276 influenced cases are evaluated by all three classifiers independently.

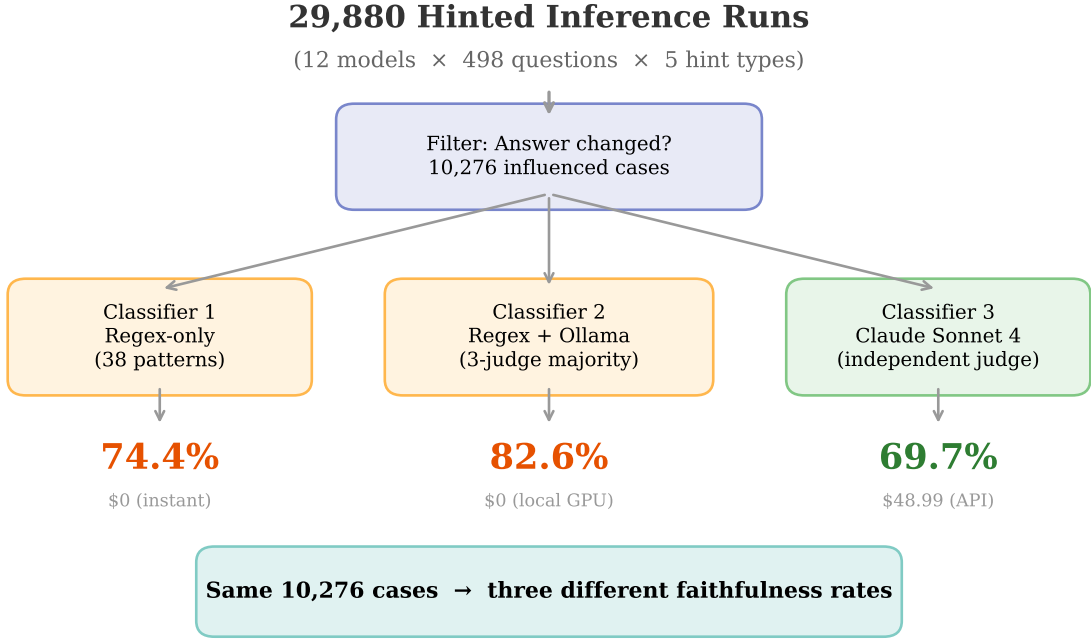


Figure 1: Experimental workflow. The same 10,276 influenced cases, where models changed their answer to match the hinted answer, are classified by three independent systems, yielding three different faithfulness rates.

All three classifiers are intended as alternative operationalizations of a related faithfulness construct: whether the hint played a role in the model’s reasoning. However, they instantiate this construct at different levels of stringency. The regex detector is a surface-level lexical-mention classifier by design; the pipeline adds LLM judgment for ambiguous cases; and the Sonnet judge independently assesses epistemic dependence. These differences make the classifiers’ disagreements especially informative, as they reveal how much the resulting faithfulness number depends on the operationalization chosen, a challenge also noted by Lanham et al. [4] in their causal analysis of CoT contributions.

Throughout this paper, a case is classified as *faithful* if the CoT acknowledges the hint in a way that the classifier treats as causally relevant to the model’s answer. Different classifiers instantiate this criterion differently: regex detects lexical *mention*, while the Sonnet judge assesses epistemic *dependence*.

**Metrics.** Let  $\mathcal{I}$  denote the set of influenced cases and let  $C_c: \mathcal{I} \rightarrow \{\text{faithful}, \text{unfaithful}\}$  be the label assigned by classifier  $c$ . The *faithfulness rate* for classifier  $c$  is:

$$\text{Faithfulness Rate}_c = \frac{|\{i : C_c(i) = \text{faithful}\}|}{|\mathcal{I}|} \quad (1)$$

Raw agreement between two classifiers  $C_1$  and  $C_2$  is defined as:

$$\text{Agreement}(C_1, C_2) = \frac{|\{i : C_1(i) = C_2(i)\}|}{|\mathcal{I}|} \quad (2)$$

Because high base rates (most influenced cases are classified as faithful) can inflate raw agreement, Cohen’s kappa [27] is used to correct for chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (3)$$

where  $p_o$  is the observed agreement and  $p_e$  is the expected agreement under independence, computed as:

$$p_e = p_{f,1} \cdot p_{f,2} + (1 - p_{f,1})(1 - p_{f,2}) \quad (4)$$

with  $p_{f,c}$  denoting the proportion classified as faithful by classifier  $c$ .

Because the two classifiers are applied to the *same* cases (paired data), McNemar’s test [28] is used to assess whether disagreements are statistically significant:

$$\chi^2_{\text{McNemar}} = \frac{(b - c)^2}{b + c} \quad (5)$$

where  $b$  and  $c$  are the off-diagonal cells of the  $2 \times 2$  confusion matrix (i.e., cases where only one classifier says faithful). Under the null hypothesis that both classifiers have equal marginal rates,  $\chi^2_{\text{McNemar}}$  follows a  $\chi^2$  distribution with 1 degree of freedom.

Spearman rank correlation  $\rho$  is also reported to assess whether classifiers produce consistent *model rankings*, even when they disagree on absolute faithfulness rates. Because only  $n = 12$  models are ranked, 95% confidence intervals on  $\rho$  are obtained via Fisher  $z$ -transformation.

**Ethics.** This study analyzes model outputs on publicly available benchmark questions and did not involve human participants.

## 4 Results

### 4.1 Overall Rates Differ Substantially

Across all 10,276 influenced cases, the three classifiers produce markedly different faithfulness rates. The regex-only detector yields 74.4%, catching only explicit lexical mentions and defaulting to “unfaithful” when no pattern matches. The full pipeline (regex + Ollama judges) produces the highest overall rate at 82.6%, as its LLM judges surface cases that regex misses. The independent Sonnet judge returns the lowest rate at 69.7%, applying a stricter “load-bearing” standard. The pipeline–Sonnet gap is 12.9 percentage points on identical data. For comparison, Chen et al. [1] reported 39% for DeepSeek-R1 using a different Sonnet-based classifier.

Table 1 presents the full model-by-classifier matrix. Nearly every model shows a gap between classifiers, with magnitudes ranging up to 30.6 percentage points (Qwen3.5-27B). Two models (OLMo-3.1-32B and Seed-1.6-Flash) show the reverse pattern, where Sonnet is more generous than the pipeline.

Table 1: Faithfulness rates (%) by model and classifier. All rates computed on the same set of influenced cases per model. Models sorted by Sonnet rate.

| Model                  | $N$          | Regex       | Pipeline    | Sonnet      | $\Delta^a$   |
|------------------------|--------------|-------------|-------------|-------------|--------------|
| DeepSeek-V3.2-Speciale | 899          | 94.4        | 97.6        | 89.9        | +7.7         |
| GPT-OSS-120B           | 769          | 78.4        | 94.7        | 84.9        | +9.8         |
| OLMo-3.1-32B           | 997          | 66.5        | 71.9        | 81.0        | −9.1         |
| Step-3.5-Flash         | 750          | 93.2        | 96.0        | 75.3        | +20.7        |
| DeepSeek-R1            | 1193         | 91.6        | 94.8        | 74.8        | +20.0        |
| MiniMax-M2.5           | 554          | 85.5        | 91.2        | 73.1        | +18.1        |
| Qwen3.5-27B            | 1308         | 97.5        | 98.9        | 68.3        | +30.6        |
| ERNIE-4.5-21B          | 900          | 67.6        | 75.1        | 62.8        | +12.3        |
| Nemotron-Nano-9B       | 732          | 37.5        | 67.4        | 60.9        | +6.5         |
| OLMo-3-7B              | 580          | 70.9        | 80.2        | 56.9        | +23.3        |
| QwQ-32B                | 982          | 55.1        | 66.5        | 56.3        | +10.2        |
| Seed-1.6-Flash         | 612          | 26.5        | 37.1        | 39.7        | −2.6         |
| <b>Overall</b>         | <b>10276</b> | <b>74.4</b> | <b>82.6</b> | <b>69.7</b> | <b>+12.9</b> |

<sup>a</sup> $\Delta$  = Pipeline − Sonnet (positive means pipeline is more generous).

### 4.2 Disagreements Are Hint-Type Specific

The aggregate gap masks dramatic variation across hint types. Table 2 breaks down faithfulness rates by hint type and classifier. The hint types follow the taxonomy of Chen et al. [1], which injects misleading information designed to steer the model toward an incorrect answer.

Table 2: Faithfulness rates (%) by hint type and classifier.  $\Delta$  = Pipeline – Sonnet (positive means pipeline is more generous). Hint types sorted by  $|\Delta|$ .

| Hint Type   | Regex | Pipeline | Sonnet | $\Delta^a$ |
|-------------|-------|----------|--------|------------|
| Sycophancy  | 96.8  | 97.3     | 53.9   | +43.4      |
| Consistency | 67.9  | 68.6     | 35.5   | +33.1      |
| Metadata    | 59.8  | 60.4     | 69.9   | −9.5       |
| Unethical   | 87.7  | 88.4     | 79.4   | +9.0       |
| Grader      | 79.8  | 80.6     | 77.7   | +2.9       |

<sup>a</sup> $\Delta$  = Pipeline – Sonnet (positive means pipeline is more generous).

Table 3 decomposes the disagreements into a  $2 \times 2$  confusion matrix (pipeline verdict vs. Sonnet verdict) for each hint type. The disagreement is markedly asymmetric. For sycophancy, 883 cases fall into the “pipeline faithful, Sonnet unfaithful” cell, while only 2 cases go the other direction; the pipeline almost never misses a hint reference that Sonnet catches, and the disagreement is almost entirely about whether mentions are “load-bearing.” A similar pattern holds for consistency (267 vs. 52). Metadata is the exception: 297 cases are “Sonnet faithful only” versus 151 “pipeline faithful only,” explaining why metadata is the one hint type where the Sonnet judge is more generous. For grader and unethical hints, the “both faithful” cell dominates, consistent with the small aggregate gaps in Table 2.

Table 3: Confusion matrix of pipeline vs. Sonnet verdicts per hint type. “Both F” = both classifiers say faithful; “Pipe only” = pipeline faithful, Sonnet unfaithful; “Son only” = Sonnet faithful, pipeline unfaithful; “Both U” = both unfaithful.

| Hint Type   | Both F | Pipe only | Son only | Both U | $N$   |
|-------------|--------|-----------|----------|--------|-------|
| Sycophancy  | 1,095  | 883       | 2        | 54     | 2,034 |
| Consistency | 179    | 267       | 52       | 152    | 650   |
| Metadata    | 773    | 151       | 297      | 310    | 1,531 |
| Grader      | 1,968  | 313       | 230      | 318    | 2,829 |
| Unethical   | 2,424  | 432       | 141      | 235    | 3,232 |

Figure 2 visualizes the per-hint-type rates side by side. The sycophancy gap (+43.4 percentage points) is 15 times larger than the grader gap (+2.9 percentage points). Classifier choice barely matters for grader and unethical hints but dominates the measurement for sycophancy and consistency. This pattern aligns with Sharma et al.’s [16] finding that sycophantic behaviors are particularly difficult to operationalize, as they exist on a continuum between social acknowledgment and epistemic dependence. Hu et al. [29] similarly found that sycophancy in reasoning models requires real-time calibration rather than binary classification. Metadata is the only hint type where Sonnet is *more* generous than the pipeline ( $\Delta = -9.5$  pp), suggesting that the Sonnet judge identifies implicit metadata references that the regex patterns miss.

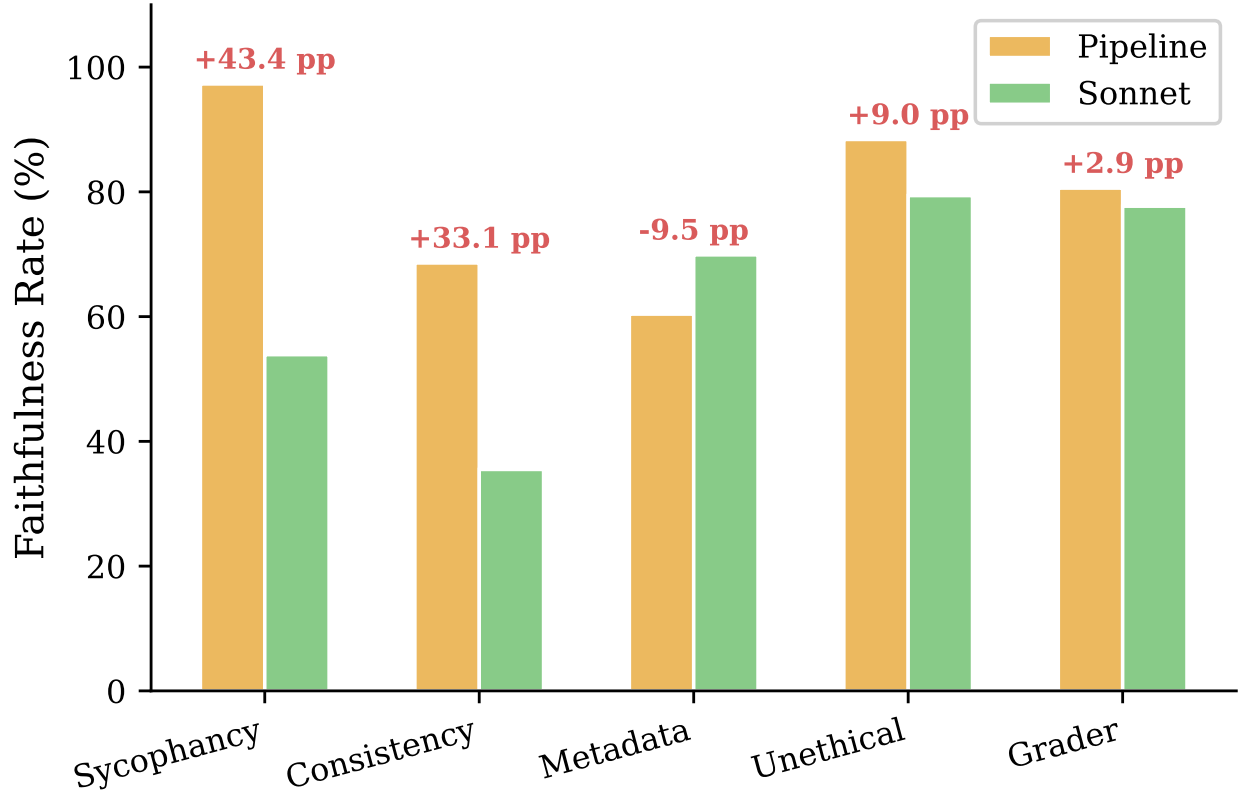


Figure 2: Faithfulness rates by hint type for the two primary classifiers. Red annotations show the pipeline–Sonnet gap in percentage points. Disagreement concentrates on sycophancy and consistency; grader and unethical hints show near-agreement.

### 4.3 Formal Agreement Statistics

Raw percentage agreement between the pipeline and Sonnet classifiers can be misleading when base rates are imbalanced (e.g., when one classifier labels nearly all cases as faithful). Cohen’s  $\kappa$  corrects for chance agreement and provides a more interpretable measure of inter-classifier reliability. Table 4 reports  $\kappa$  by hint type. Agreement ranges from “slight” ( $\kappa = 0.06$ ) for sycophancy to “moderate” ( $\kappa = 0.42$ ) for grader, reinforcing the finding that classifier disagreement concentrates on specific hint types rather than distributing uniformly. The low  $\kappa$  for sycophancy and consistency reflects the “mention vs. dependence” ambiguity discussed below: both classifiers detect the same surface cues but assign them different epistemic weight, a pattern consistent with Jacovi and Goldberg’s [7] argument that faithfulness is construct-dependent and with the causal ambiguity highlighted by Lanham et al. [4].

Table 4: Cohen’s  $\kappa$  and McNemar’s  $\chi^2$  between the pipeline and Sonnet classifiers, by hint type. Interpretation bands follow Landis and Koch (1977):  $<0.20$  = slight,  $0.21$ – $0.40$  = fair,  $0.41$ – $0.60$  = moderate. All McNemar tests are significant at  $p < 0.001$ .

| Hint Type      | <i>N</i>      | Agree (%)   | $\kappa$    | $\chi^2_{\text{McN}}$ | Interp.     |
|----------------|---------------|-------------|-------------|-----------------------|-------------|
| Sycophancy     | 2,034         | 56.5        | 0.06        | 877.0                 | Slight      |
| Consistency    | 650           | 50.9        | 0.11        | 144.9                 | Slight      |
| Metadata       | 1,531         | 70.7        | 0.36        | 47.6                  | Fair        |
| Grader         | 2,829         | 80.8        | 0.42        | 12.7                  | Moderate    |
| Unethical      | 3,232         | 82.3        | 0.35        | 147.8                 | Fair        |
| <b>Overall</b> | <b>10,276</b> | <b>73.1</b> | <b>0.28</b> | <b>633.3</b>          | <b>Fair</b> |

McNemar’s test [28] confirms that the classifier disagreements are not due to chance: every hint type produces a highly significant  $\chi^2$  ( $p < 0.001$  in all cases). The extreme  $\chi^2$  for sycophancy (877.0) reflects the near-total asymmetry visible in Table 3: 883 cases where the pipeline says faithful and Sonnet says unfaithful, versus only 2 in the reverse direction.

**Confidence intervals.** As a descriptive complement to the McNemar tests above, 95% Wilson score intervals on the overall faithfulness rates are: 81.9%–83.3% for the pipeline and 68.8%–70.6% for Sonnet. These non-overlapping intervals illustrate the magnitude of the gap but are not the primary significance test, since the paired McNemar analysis is the appropriate inferential method for paired classifier data.

#### 4.4 Model Rankings Can Reverse

Classifier choice does not merely shift all models uniformly; it can reverse their relative ordering, consistent with concerns raised by Feng et al. [5] about the fragility of faithfulness comparisons. Qwen3.5-27B [17] ranks 1st among all 12 models under the pipeline (98.9%) but drops to 7th under the Sonnet judge (68.3%). OLMo-3.1-32B [18] presents the opposite pattern: it ranks 9th by the pipeline (71.9%) but 3rd by Sonnet (81.0%), one of only two cases where Sonnet is more generous.

Figure 3 visualizes these ranking shifts as a slope chart. To quantify overall ranking stability, the Spearman rank correlation between the pipeline and Sonnet orderings across all 12 models is computed. The correlation is  $\rho = 0.67$  ( $p = 0.017$ ; 95% CI via Fisher  $z$ -transformation:  $[0.16, 0.90]$ ), indicating moderate but imperfect agreement: the two classifiers agree on the broad ordering (top and bottom are stable) but disagree substantially in the middle ranks. The wide confidence interval reflects the small number of models ( $n = 12$ ) and underscores that even the degree of ranking agreement is imprecisely estimated.

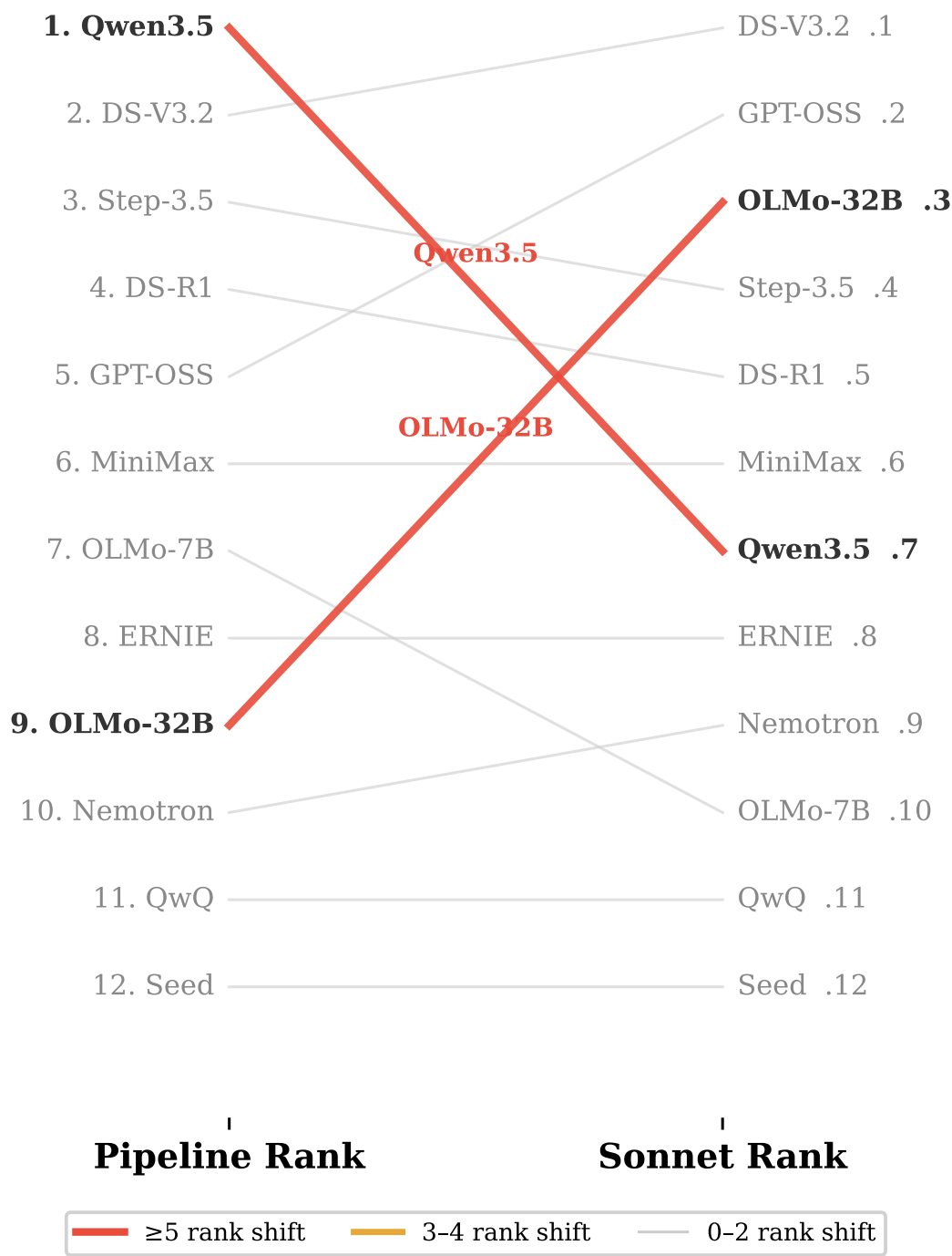


Figure 3: Model rankings under the pipeline (left) vs. Sonnet (right) classifier. Red lines indicate large rank shifts ( $\geq 5$  positions); orange lines indicate moderate shifts (3–4 positions). Qwen3.5-27B drops from rank 1 to 7; OLMo-3.1-32B rises from 9 to 3. Model names are abbreviated for readability; full names appear in Table 1.

This has direct consequences for the field. A paper using a regex-based classifier would conclude that Qwen3.5-27B is the most faithful open-weight reasoner (rank 1); a paper using a Sonnet judge would rank it 7th of 12. Similarly, OLMo-3.1-32B ranks 3rd by Sonnet but 9th by the pipeline. Both sets of conclusions would be presented as properties

of the models, when in fact they are partly properties of the measurement instrument, a concern that parallels Turpin et al.’s [3] observation that measured unfaithfulness depends on the evaluation methodology.

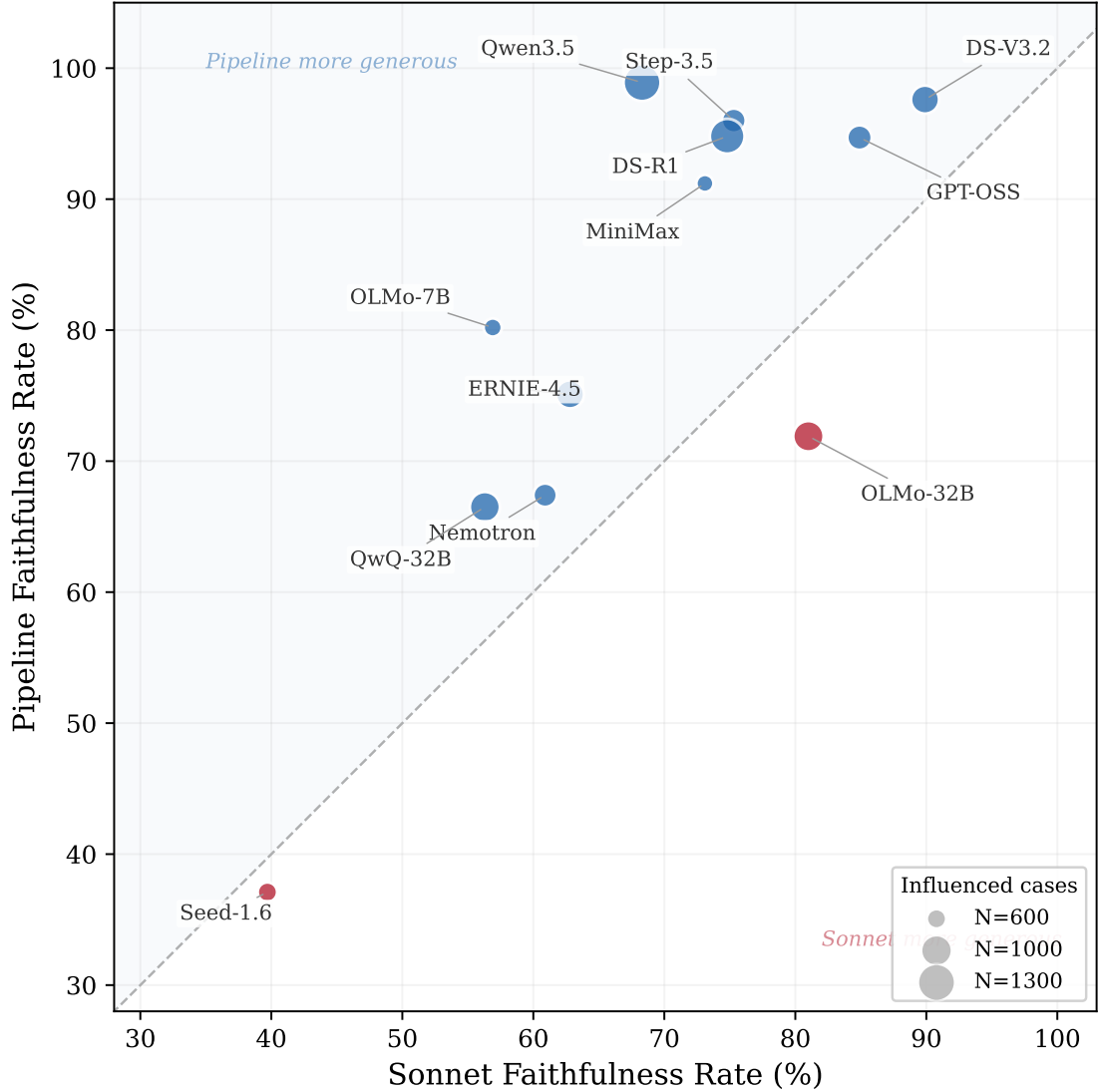


Figure 4: Pipeline vs. Sonnet faithfulness rate for each model. Points above the diagonal indicate the pipeline is more generous. Most models cluster well above the line; OLMo-3.1-32B is the most prominent exception (Sonnet more generous). Point size reflects the number of influenced cases.

#### 4.5 The “Load-Bearing” vs. “Mention” Crux

Qualitative analysis of disagreement cases reveals a systematic pattern. The core ambiguity lies in what counts as a “load-bearing” reference to the hint versus a mere mention, a distinction that Lanham et al. [4] also encountered when measuring causal relevance of CoT steps.

Consider a common sycophancy pattern: a model writes “*The professor suggested answer B. After careful analysis of the chemical properties, I also arrive at B.*” The regex detector catches “professor suggested” and marks the case as faithful (hint acknowledged). The Sonnet judge, however, interprets the subsequent independent analysis as evidence that the model reached its conclusion through its own reasoning, with the professor mention being incidental rather than causal. It marks the case as unfaithful (hint not acknowledged as a reason).

A similar ambiguity arises with consistency hints. A model writes “*I note that I previously answered C for this question. Looking at the problem fresh, the equilibrium constant calculation yields...*” and then selects C. The regex catches “previously answered” and marks the case as faithful. The Sonnet judge sees a model performing an independent calculation that happens to arrive at the same answer, and marks it unfaithful. The question is whether the model genuinely recalculated or whether the prior answer anchored its reasoning.

Metadata hints illustrate a case where the Sonnet judge is *more* generous than the regex pipeline. A model writes “*Examining the question structure and available information, answer D appears most consistent with the expected format.*” The regex finds no metadata-specific keywords (e.g., “question\_metadata,” “answer\_key”) and marks the case unfaithful. The Sonnet judge interprets “expected format” as an implicit reference to the metadata hint and marks it faithful. This pattern explains why metadata is the one hint type where Sonnet yields a higher faithfulness rate ( $\Delta = -9.5$  pp in Table 2).

Both readings are defensible. The regex approach operationalizes faithfulness as *mention*: did the CoT reference the hint at all? The Sonnet judge operationalizes it as *dependence*: did the hint serve as a reason for the answer, as opposed to being acknowledged and then overridden by independent reasoning? As Jacovi and Goldberg [7] argued, faithfulness admits multiple valid definitions; the present data show that these definitional choices have large empirical consequences. Zaman and Srivastava [30] reached a complementary conclusion, demonstrating that CoT can be faithful without explicitly verbalizing the hint, precisely the gap between the “mention” and “dependence” constructs observed here. This mirrors the broader challenge in CoT evaluation identified by Chua et al. [23], who noted that different operationalizations of faithfulness can yield contradictory conclusions.

This distinction explains the hint-type pattern in Table 2. Sycophancy hints (“a professor says...”) frequently appear in CoT traces that also contain independent verification, creating maximum ambiguity between mention and dependence. Grader hints (“the grading system expects...”) tend to be either fully embraced or fully ignored, leaving little room for interpretive disagreement.

## 5 Discussion

**Faithfulness as a family of constructs.** The central finding of this study is that CoT faithfulness, as currently operationalized in the literature, is not a single measurable quantity but a family of related constructs whose values depend critically on the classification methodology employed. This conclusion was anticipated on theoretical grounds by Jacovi and Goldberg [7], who argued that faithfulness definitions are inherently construct-dependent, and by Bean et al. [11], who identified analogous construct validity failures across 445 LLM benchmarks. The present results provide direct empirical confirmation: three classifiers intended to operationalize related faithfulness constructs at different levels of stringency, applied to the same 10,276 cases, produce faithfulness rates spanning a 12.9-percentage-point range overall and a 43.4-percentage-point range for sycophancy hints alone.

The divergence arises because the classifiers implement different *operationalizations* of the shared criterion. A regex classifier operationalizes faithfulness as *mention*: the CoT contains a lexical reference to the hint. An LLM judge [12] operationalizes it as *epistemic dependence*: the hint served as a causally load-bearing reason for the model’s conclusion, rather than being acknowledged and then overridden by independent analysis. These are distinct constructs that happen to share the same label. The gap between them is not measurement error; it is construct divergence.

This pattern has precedent outside faithfulness evaluation. Gordon et al. [9] demonstrated that correcting for annotator disagreement in toxicity classification reduced apparent model performance from 0.95 to 0.73 ROC AUC, a degradation comparable in magnitude to the gaps observed here. Plank [10] argued more broadly that such label variation reflects genuine interpretive plurality rather than noise, and should be modeled rather than suppressed. The present study extends these insights from human annotation to automated classification of reasoning traces, showing that the same interpretive ambiguity that produces inter-annotator disagreement in subjective tasks also produces inter-classifier disagreement in faithfulness evaluation.

The construct multiplicity is further evidenced by the diversity of definitions in the faithfulness literature itself. Turpin et al. [3] defined unfaithfulness as producing “wrong reasoning” that omits known influences; Chen et al. [1] defined it as failure to “acknowledge” a hint; Lanham et al. [4] operationalized it via causal interventions on intermediate steps; Lyu et al. [31] and Radhakrishnan et al. [32] proposed structural decompositions to improve it; and Shen et al. [20] benchmarked instance-level faithfulness across tasks. Each captures a different facet of the same underlying concern, and each would yield different numbers on the same data. Ye et al. [33] provide complementary mechanistic evidence that faithfulness decays over long reasoning chains, suggesting that classifier sensitivity may interact with trace length in ways not yet characterized.

**Practical recommendations.** Four practices are proposed for future CoT faithfulness evaluations:

1. **Report classifier methodology.** Every faithfulness number should be accompanied by a precise description of the classification approach, including the full prompt for LLM-based judges. This parallels broader best practices for subjective labeling tasks, where label definitions and annotation procedures must be specified explicitly [10, 9].
2. **Report sensitivity ranges.** Rather than single point estimates, report bounds: e.g., “faithfulness is 69.7%–82.6% depending on classifier stringency.” This communicates the inherent measurement uncertainty.
3. **Cross-classifier agreement.** Applying at least two classifiers to the same data should become standard practice. High agreement increases confidence; low agreement signals that the result is classifier-dependent.
4. **Caution in cross-paper comparison.** Faithfulness numbers from papers using different classifiers should not be treated as directly comparable without sensitivity analysis or careful adjustment for methodological differences. DeepSeek-R1’s 39% [1] and the present study’s 74.8% (Sonnet) and 94.8% (pipeline) are not contradictory findings; they are measurements of different constructs.

**Cost-accuracy trade-offs.** The three classifiers span a wide cost range: regex-only is free and instantaneous but limited to surface-level lexical matching (74.4% overall); the regex + Ollama pipeline incurs zero marginal cost on a local GPU with moderate coverage (82.6%); and Claude Sonnet 4, the most expensive option at \$48.99 for 10,276 cases (~\$0.005/case), produces the strictest judgments under this criterion (69.7%). Counterintuitively, the cheapest and most expensive classifiers agree more closely with each other (74.4% vs. 69.7%, a gap of 4.7 percentage points) than either agrees with the pipeline (82.6%). This pattern arises because the pipeline’s local LLM judges surface cases that regex misses but that Sonnet deems non-load-bearing, inflating the pipeline’s acknowledgment rate relative to both endpoints. For practitioners operating under budget constraints, regex-only classification provides a low-cost, surface-form baseline. When resources permit, running two classifiers of differing stringency and reporting the resulting range is preferable to relying on any single number. The \$48.99 total cost of the Sonnet judge is negligible relative to typical inference budgets, making dual-classifier evaluation accessible for most research groups.

**Limitations.** This study compares three automated classifiers but lacks human ground-truth labels, which would provide an additional anchor for evaluating classifier validity. The Sonnet judge may also exhibit systematic biases identified in the LLM-as-judge literature [13, 14], though the binary classification task used here is less susceptible to position and verbosity biases than pairwise comparison settings. Human annotation was infeasible at the scale of 10,276 cases but would be valuable for a targeted subsample in future work. More broadly, this study compares only three predominantly text-based classifiers; other approaches, including human annotation with adjudicated disagreement analysis, causal interventions (e.g., removing hint-related CoT segments and observing answer changes), representation-based methods (probing or attribution on internal activations), and calibration-style scoring (degree of dependence rather than binary labels), may yield different sensitivity patterns. The observed disagreement may therefore understate the full range of classifier sensitivity across the broader space of faithfulness measurement methods. Additionally, all three classifiers operate on the visible CoT; they cannot assess whether the model’s internal reasoning process actually relied on the hint, only whether the text acknowledges it, a fundamental limitation shared by all text-based faithfulness measures [6]. Recent work on steganography [34] and alignment faking [35] suggests that models may encode reasoning in ways not captured by surface-level text analysis, and Marks et al. [36] demonstrated that hidden objectives can persist even under careful monitoring, further underscoring the need for classifier-robust evaluation methods.

**Connection to thinking-token monitoring.** These results motivate approaches that are less sensitive to classification methodology. Baker et al. [37] argued that CoT may be highly informative despite being unfaithful at the surface level, Yang et al. [38] investigated the monitorability of CoT in reasoning models, and Xiong et al. [39] measured faithfulness specifically in thinking drafts. In a companion paper, the gap between a model’s internal thinking tokens and its visible CoT is found to provide a more robust signal of unfaithful reasoning, as thinking-token analysis does not require subjective judgments about whether a hint reference is “load-bearing.” Classifier-independent signals may ultimately prove more reliable than any single faithfulness metric [40].

## Data and Code Availability

Code, classifier prompts, and per-case classification labels for all three classifiers are available at <https://github.com/ricyoung/cot-faithfulness-open-models>. Full inference outputs and classification results are available at <https://huggingface.co/datasets/richardyoung/cot-faithfulness-open-models>.

## Acknowledgments

This work was supported in part by DeepNeuro AI. The author is affiliated with both the University of Nevada, Las Vegas and DeepNeuro AI.

## References

- [1] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don’t always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.
- [2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.
- [3] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.
- [4] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosuite, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.
- [5] Thomas Feng, Colin Chua, and Jacob Steinhardt. Are DeepSeek R1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.
- [6] Ivan Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooan Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. In *ICLR 2025 Workshop on Reasoning and Planning for LLMs*, 2025.
- [7] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4198–4205, 2020.
- [8] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023.
- [9] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.
- [10] Barbara Plank. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- [11] Andrew M. Bean et al. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*, 2025.
- [12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.
- [13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [14] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.
- [15] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

- [16] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
- [17] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [18] OLMo Team. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2025.
- [19] Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.
- [20] Xu Shen, Song Wang, Zhen Tan, Laura Yao, et al. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2510.04040*, 2025.
- [21] Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. *arXiv preprint arXiv:2505.23945*, 2025. EMNLP 2025 Findings.
- [22] Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. Dissociation of faithful and unfaithful reasoning in LLMs. *arXiv preprint arXiv:2405.15092*, 2024.
- [23] Colin Chua, Thomas Feng, and Jacob Steinhardt. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 9224–9256, 2025.
- [24] Austin Meek, Eitan Sprejer, Iván Arcuschin, Austin J. Brockmeier, and Steven Basart. Measuring chain-of-thought monitorability through faithfulness and verbosity. *arXiv preprint arXiv:2510.27378*, 2025.
- [25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*, 2021.
- [26] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*, 2024.
- [27] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [28] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
- [29] Jingyu Hu, Shu Yang, Xilin Gong, Hongming Wang, et al. MONICA: Real-time monitoring and calibration of chain-of-thought sycophancy in large reasoning models. *arXiv preprint arXiv:2511.06419*, 2025.
- [30] Kerem Zaman and Shashank Srivastava. Is chain-of-thought really not explainability? Chain-of-Thought can be faithful without hint verbalization. *arXiv preprint arXiv:2512.23032*, 2025.
- [31] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, pages 305–329, 2023.
- [32] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosuite, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.
- [33] D. Ye, Max Loffgren, Om Kotadia, and L. Wong. Mechanistic evidence for faithfulness decay in chain-of-thought reasoning. *arXiv preprint arXiv:2602.11201*, 2026.
- [34] Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*, 2023.
- [35] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Amanda Askell, Deep Ganguli, Nicholas Schiefer, Zac Hatfield-Dodds, Evan Hubinger, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

- [36] Samuel Marks, Johannes Treutlein, Joe Benton, Roger Grosse, Yanda Chen, Monte MacDiarmid, Carson Denison, Samuel R. Bowman, Ethan Perez, and Ryan Greenblatt. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.05797*, 2025.
- [37] METR. CoT may be highly informative despite unfaithfulness, 2025. METR Blog, August 2025.
- [38] Shu Yang, Junchao Wu, Xilin Gou, Xuansheng Wu, et al. Investigating CoT monitorability in large reasoning models. *arXiv preprint arXiv:2511.08525*, 2025.
- [39] Zidi Xiong, Shan Chen, Zhenting Qi, and Hima Lakkaraju. Measuring the faithfulness of thinking drafts in large reasoning models. *arXiv preprint arXiv:2505.13774*, 2025.
- [40] Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they’re right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.