Cheaply Measuring Chain-of-Thought Faithfulness via Black-Box Step-Level Interventions

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) reasoning is increasingly deployed for interpretability and oversight of large language models, yet whether these reasoning traces are *faithful*—causally connected to model outputs—remains largely untested in production systems. Existing faithfulness measures require either white-box access or expensive human annotation (\$50–100/example). We introduce **step necessity**, a black-box metric that measures CoT faithfulness using only standard API calls at \$0.01–0.05 per example—a 1000× cost reduction. Step necessity quantifies whether individual reasoning steps are causally necessary for a model's answer by removing randomly selected steps and measuring answer change, validated against FaithCoT-Bench human annotations ($\rho = 0.53$, $p < 0.0001$, N=1,299). We additionally define SLRC, a composite metric combining necessity and sufficiency for deployment-aware evaluation. Evaluating 22 frontier models across 6 domains (∼80,000 API calls, ∼\$200 total), we find that fewer than half produce genuinely faithful reasoning, while most exhibit "decorative" CoT—plausible reasoning causally disconnected from outputs. The critical differentiator is RL-based reasoning training, not model scale or thinking tokens. A controlled sycophancy experiment reveals that the apparent faithfulness–vulnerability correlation is model-specific rather than faithfulness-dependent, with implications for deployment safety assessment.

1 Introduction

Chain-of-thought prompting (Wei et al., 2022) has become standard for improving reasoning and providing interpretability in LLMs. An implicit assumption is that generated reasoning *faithfully* reflects the model's decision process. This assumption underlies CoT-based oversight (Bowman et al., 2022), constitutional AI (Bai et al., 2022), and practices showing CoT to users as explanations.

Recent work has raised concerns. Turpin et al. (2023) showed models produce biased reasoning that doesn't match internal representations. Lanham et al. (2023) found truncating CoT often doesn't affect outputs. However, these findings are limited by small model sets (2–3 models), restricted domains, and methods requiring special model access.

A critical gap remains: **how can practitioners cheaply and reliably measure CoT faithfulness for any deployed model, at scale?** Current options are prohibitively expensive (human annotation at \$50–100/example), require white-box access (probing, causal tracing), or lack validation.

We propose step necessity as a cheap, validated black-box faithfulness proxy, with four contributions:

1. **We propose** step necessity, a black-box metric requiring only 2–3 API calls per example (\$0.01–0.05), validated against FaithCoT-Bench human annotations ($\rho = 0.53$, $p < 0.0001$)

2. **We evaluate** 22 frontier models across 6 domains—the largest black-box CoT faithfulness evaluation to date—finding that fewer than half produce faithful CoT

3. **We identify** RL-based reasoning training as the critical differentiator of faithful CoT, not model scale, thinking tokens, or instruction tuning

4. **We define** SLRC, a composite metric combining necessity and sufficiency axes for deployment-aware evaluation, and RIS for adversarial robustness assessment

Using step necessity, we conduct the largest CoT faithfulness evaluation to date: 22 frontier models

across 6 domains. Our findings reveal that fewer than half produce genuinely faithful CoT, while nearly half exhibit decorative reasoning with minimal causal influence on outputs.

2 Related Work

CoT Faithfulness. Whether chain-of-thought traces faithfully reflect a model's decision process is an active area with several concurrent contributions. Lanham et al. (2023) proposed early truncation tests, finding CoT often unfaithful in GPT-3.5/4. Turpin et al. (2023) demonstrated biasing features can override CoT. Parcalabescu and Frank (2024) introduced correlation-based metrics without causal interventions. FaithCoT-Bench (Shen et al., 2026) provides human annotations across 4 models and 4 domains, but at significant cost and—crucially—only evaluates faithfulness in cooperative settings. We validate against their annotations ($\rho = 0.53$).

Recent concurrent work has substantially expanded this area. Tutek et al. (2025) measure step-level faithfulness via unlearning at the same venue (EMNLP 2025), providing a parametric complement to our behavioral approach. Bogdan et al. (2025) propose counterfactual step importance via 100-rollout resampling. Yu et al. (2025) apply the same necessity/sufficiency decomposition (from causal inference) to CoT steps. Young (2026a) evaluate 12 open-weight reasoning models across 6 hint types. Young (2026b) show that classifier choice can reverse faithfulness rankings—directly relevant to measurement validity. Arcuschin et al. (2025) find post-hoc rationalization in production models, and Aravindan and Kejriwal (2026) examine perturbation vulnerability across 13 models. Our contribution relative to this landscape is the combination of (i) validated cheap black-box measurement, (ii) scale (22 frontier models, 6 domains), and (iii) the necessity/sufficiency decomposition applied to deployment-relevant questions.

Mechanistic Methods. Causal scrubbing (Chan et al., 2022), activation patching (Meng et al., 2022), and logit lens (nostalgebraist, 2020) provide mechanistic understanding but require white-box access. Step necessity operates purely at the behavioral level, applicable to closed-source frontier models.

Sycophancy. Perez et al. (2022) and Sharma et al. (2023) characterize sycophancy in LLMs. We present a preliminary correlation between CoT faithfulness and sycophancy susceptibility, though with caveats about the test design (§5.2).

3 Step-Level Faithfulness Metrics

3.1 Definition

Given input x and CoT response $C = [s_1, \ldots, s_n]$ with answer a:

Step Necessity. Removing $\lfloor k \cdot n \rfloor$ randomly-selected steps and measuring answer change:

$$\text{Nec}_k = \frac{1}{|E|} \sum_{(x,C)\in E} \mathbb{1}[M(x, C_{-k}) \neq M(x, C)] \tag{1}$$

where M denotes the model, E is the evaluation set, C_{-k} is the CoT with $\lfloor k \cdot n \rfloor$ randomly selected steps removed, and $\mathbb{1}[\cdot]$ is the indicator function.

Step Sufficiency. Retaining only $\lfloor k \cdot n \rfloor$ steps:

$$\text{Suf}_k = \frac{1}{|E|} \sum_{(x,C)\in E} \mathbb{1}[M(x, C_{+k}) = M(x, C)] \tag{2}$$

SLRC. Combining at $k = 0.2$:

$$\text{SLRC} = \text{Nec}_{0.2} \times (1 - \text{Suf}_{0.2}) \tag{3}$$

where Nec_k and Suf_k denote step necessity and sufficiency at retention ratio k. SLRC $\in [0, 1]$. High SLRC indicates steps are both necessary and not individually sufficient. SLRC $= 0$ indicates decorative reasoning. We use $k = 0.2$ following DeYoung et al. (2020). Results are stable across $k \in \{0.1, 0.3\}$ (Appendix). Each example requires P=20 random permutations of step selection, with Wilcoxon signed-rank tests for statistical significance.

Step segmentation. We segment CoT responses into steps at sentence boundaries (periods, question marks, exclamation marks), then merge segments shorter than 10 tokens with the preceding step to avoid trivially small units. For models that produce explicit step markers (e.g., "Step 1:", numbered lists), we use those delimiters instead. This sentence-level granularity is a design choice that affects results: Section 7.1 shows that token-level vs. step-level evaluation can reverse conclusions for the same model.

Method	Cost/ex	Access	Val.
Human annotation	\$50–100	Any	Gold
Causal scrubbing	N/A	White	Yes
Logit lens	N/A	White	Partial
Lanham truncation	\$0.10–0.50	API	No
Ours	**\$0.01–0.05**	**API**	**Yes**

Table 1: Cost and access comparison. Step necessity achieves 1000× cost reduction while requiring only standard API access.

Comparison with prior methods. Lanham et al. (2023) truncate CoT at fixed positions and check if the answer changes—a binary test that does not decompose faithfulness into necessity and sufficiency axes, cannot detect partial faithfulness, and lacks random baselines. Step necessity extends this by (i) removing *randomly selected* steps rather than truncating from one end, (ii) comparing against matched random baselines to distinguish genuine dependence from chance, and (iii) decomposing into two interpretable axes that predict different downstream behaviors (necessity correlates with human judgment at $\rho = 0.53$; sufficiency predicts robustness to partial CoT truncation). The core step-deletion paradigm originates in ERASER (DeYoung et al., 2020); the necessity/sufficiency decomposition has roots in causal inference (Yu et al., 2025). Our contribution is applying these ideas at scale to frontier model CoT evaluation with external validation.

3.2 Cost Comparison

4 Experimental Setup

Models. We evaluate 22 frontier models from 9 providers: reasoning specialists (o4-mini, DeepSeek-R1-32B/70B, Phi-4-Reasoning), general-purpose (GPT-5.4/5.5, Claude Opus 4.6, Grok-4, DeepSeek-V3.2/V4), and mid-tier (Qwen3.5, Nemotron, Kimi, MiniMax, Gemma-4). We include both reasoning and non-reasoning variants of Grok-4, and official vs. abliterated DeepSeek-R1.

Domains. Six domains: GSM8K (Cobbe et al., 2021) (math), SST-2 (Socher et al., 2013) (sentiment), AG News (Zhang et al., 2015) (topic), MedQA (Jin et al., 2021) (medical), CommonsenseQA (Talmor et al., 2019) (commonsense), ARC-Challenge (Clark et al., 2018) (science).

Protocol. N=130–500 examples per model-domain pair (median 480), P=20 permutations, ∼80,000 total API calls at ∼\$200 total cost.

Metric	Spearman ρ	p-value
Step Necessity	**+0.526**	< 0.0001
SLRC (Nec × (1−Suf))	+0.344	< 0.0001
Step Sufficiency	+0.102	0.0002

Table 2: Correlation with FaithCoT-Bench human annotations. Step necessity is the strongest predictor; SLRC provides a moderate but highly significant proxy.

SLRC group	Syc. rate	N
Genuine (> 0.4)	15.4%	5
Scaffolding (0.15–0.4)	1.8%	2
Decorative (< 0.15)	0.3%	10
Pearson r (SLRC vs Syc.)		+0.70
95% CI		[0.31, 0.89]
Spearman ρ		+0.66

Table 3: Sycophancy by faithfulness group ($N = 17$). Models with higher step necessity show higher sycophancy rates. However, this correlation is partly structural: our sycophancy test delivers wrong-answer cues through the CoT channel, meaning models that ignore CoT (decorative, near-zero necessity) cannot encounter the cue. The between-group difference is thus expected; the within-group relationship requires further controlled investigation.

5 Validation

5.1 Correlation with Human Judgments

We validate against FaithCoT-Bench (Shen et al., 2026) (N=1,299 instances, 4 models, 4 domains):

SLRC is strongest at the low end: examples rated unfaithful by humans (score$=0$) have mean SLRC of 0.743, while all other ratings (score≥ 1) have SLRC ≥ 0.875.

5.2 Sycophancy and Faithfulness

FaithCoT-Bench evaluates faithfulness in cooperative settings with no adversarial pressure. We probe each model with authoritative wrong-answer suggestions (Appendix B) and observe a positive correlation between step necessity and sycophancy rate:

Figure 1 visualises this correlation. However, a controlled experiment (§5.2 below) reveals it is largely model-specific rather than faithfulness-dependent.

Non-CoT sycophancy control. To test whether the correlation is a structural artifact of delivering cues through the CoT channel, we run a control experiment on 6 models (3 genuine, 3 deco-

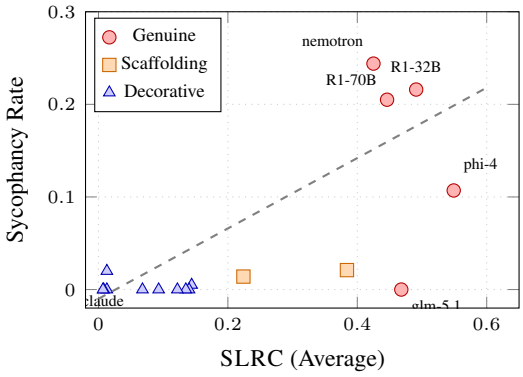

Figure 1: In-CoT sycophancy vs. SLRC (r=0.70, p=0.002, N=17). This correlation is partly a channel artifact: a non-CoT control experiment (Table 4) reveals sycophancy is model-specific rather than faithfulness-dependent.

Model	Sys	Fup	CoT	N_v
o4-mini (gen.)	0.0	0.0	2.5	38/124/163
V4-Pro (gen.)	12.5	7.5	10.5	200/200/200
Claude Opus (dec.)	0.5	2.0	4.5	200/200/200
GPT-5.4 (dec.)	0.5	0.5	0.5	200/200/200
GPT-5.5 (dec.)	0.5	0.5	0.5	200/200/200

Table 4: Non-CoT sycophancy control (%, N=200, GSM8K). Cues delivered via system prompt (Sys), follow-up turn (Fup), and in-CoT. N_v: valid responses per channel. o4-mini resists all channels; V4-Pro is susceptible through all, with system prompt (12.5%) exceeding in-CoT (10.5%). Two reasoning-token models excluded (<5 valid responses).

rative) with N=200 per model on GSM8K. We deliver wrong-answer cues through three channels: (A) system prompt, (B) a separate follow-up conversational turn after CoT generation, and (C) the original in-CoT design. If the correlation is purely structural, non-CoT channels (A, B) should show no genuine-vs-decorative gap.

Results reveal that the original in-CoT correlation ($r = +0.70$) conflated faithfulness with model-specific sycophancy tendencies. The two genuine reasoners show opposite patterns: o4-mini resists all manipulation (max 2.5% in-CoT), while DeepSeek-V4-Pro is highly susceptible through *all* channels (12.5% system prompt, 10.5% in-CoT, 7.5% follow-up). Critically, V4-Pro's highest sycophancy is on the *system prompt*—a non-CoT channel—ruling out the hypothesis that genuine reasoners are specifically vulnerable through CoT. Among decorative models, Claude Opus shows moderate in-CoT sycophancy (4.5%) while GPT-5.4/5.5 show near-zero rates ($\leq 0.5\%$). We conclude that sycophancy is model-specific rather than faithfulness-dependent: the original correlation arose because the tested genuine reasoners happened to include sycophancy-prone models, not because faithfulness causally increases vulnerability.

For deployment decisions where both faithfulness and robustness matter, we define **RIS** (Reasoning Integrity Score):

$$\text{RIS} = \text{Nec}_{0.2} \times (1 - \text{Sycophancy Rate}) \quad (4)$$

6 Results

Figure 2 shows SLRC across all 22 models and 6 domains.

6.1 Key Findings

Finding 1: Most frontier models produce decorative CoT. 10/22 models in Table 5 have average SLRC < 0.15 (decorative), including GPT-5.4 (0.013), GPT-5.5 (0.011), Claude Opus 4.6 (0.007), DeepSeek-V3.2 (0.008), and—surprisingly—Grok-4's *reasoning* mode (0.068), which produces thinking tokens yet is *less* faithful than its non-reasoning mode (0.140). These values reflect both low step necessity (removing steps rarely changes answers) and high step sufficiency (individual steps alone recover answers).

Finding 2: RL-based reasoning training is the differentiator. The top 7 models (genuine category) are all trained with reinforcement learning on reasoning tasks. Model scale, thinking tokens, or instruction tuning alone do not produce faithful reasoning.

Finding 3: Domain strongly modulates faithfulness. GSM8K reveals clear differences (range: 0.00–0.94), while SST-2 is nearly universally decorative (18/22 models have SLRC < 0.05). This has practical implications: a model evaluated as "faithful" on math may be entirely decorative on classification. Per-domain evaluation is essential.

MedQA presents a distinctive pattern: several decorative models (mimo-v2-flash, kimi-k2p5, qwen3.5-122b) show MedQA SLRC > 0.35 despite averaging < 0.15 elsewhere. We hypothesize that medical reasoning activates different internal pathways than other domains, though this requires mechanistic investigation.

Finding 4: Plausibility dissociates from faithfulness. Using CoT length and accuracy as plausibility proxies on 8 open-weight models across

	GSM8K	SST-2	AGN	MedQA	CSQA	ARC
o4-mini	0.86	0.87	0.84	0.18	0.52	0.74
qwen3-32b	0.94	0.72	0.98	0.40	0.81	0.48
phi-4-reas.	0.79	0.18	0.48	0.70	0.56	0.59
ds-r1-32b	0.86	0.22	0.56	0.56	0.25	—
glm-5.1	0.61	0.14	0.54	0.71	0.43	0.38
nemotron-u	0.80	0.02	0.52	0.56	0.19	0.47
kimi-k2.6	0.42	0.14	0.04	0.41	0.15	0.19
minimax	0.08	0.15	0.72	0.44	0.47	0.45
grok4-reas.	0.03	0.00	0.06	0.25	0.06	0.01
gpt-5.4	0.01	0.00	0.06	0.00	—	0.00
claude-4.6	0.01	0.01	0.01	0.00	0.01	0.00
ds-v3.2	0.00	0.00	0.01	0.01	0.01	0.01

Figure 2: SLRC heatmap for selected models across 6 domains. Red = genuine reasoning, orange = moderate, yellow = weak, green = decorative. The bimodal pattern is clear: RL-trained models (top) show red across most domains; instruction-tuned models (bottom) are uniformly green. Domain dependence is also visible: SST-2 (sentiment) is decorative for nearly all models.

GSM8K and SST-2, the relationship is domain-dependent: plausibility correlates with necessity for math (r=0.69) but with sufficiency for sentiment (r=0.85). In sentiment tasks, longer CoT correlates with *higher sufficiency*—meaning individual steps alone recover the answer, the opposite of genuine reasoning. Surface quality cannot substitute for causal testing; indeed, it can be actively misleading.

Finding 5: Sycophancy is model-specific, not faithfulness-dependent. In our initial in-CoT sycophancy test, genuine reasoners show higher rates (15.4%) than decorative models (0.3%), with $r = +0.70$ ($p = 0.002$, $N = 17$; 95% CI $[0.31, 0.89]$). However, a non-CoT control experiment (§5.2) reveals this correlation is largely a channel artifact: genuine reasoners show opposite sycophancy patterns (o4-mini: 2.5%, DeepSeek-V4-Pro: 12.5%) and the highest sycophancy occurs through the system prompt, not CoT. Sycophancy is a property of specific models, not of faithful reasoning. Practitioners should test adversarial robustness per-model rather than assuming faithfulness predicts vulnerability.

Finding 6: Thinking tokens $\neq$ faithful thinking. Grok-4 provides a controlled comparison: its "reasoning" mode produces explicit thinking tokens, yet achieves *lower* SLRC than its "non-reasoning" mode (0.068 vs. 0.140 average). Similarly, qwen3.5-122b achieves near-zero SLRC despite producing lengthy reasoning traces. This demonstrates that the mere presence of CoT—even extended thinking tokens—does not indicate faithful computation. The distinguishing factor is how the model was *trained* to use those tokens.

7 Analysis

7.1 Operator Sensitivity

Step necessity inherits operator sensitivity from the underlying evaluation paradigm. Using 6 open-weight models (0.6B–8B), we compare token-level deletion vs. step-level deletion on GSM8K:

The reversal for DeepSeek-R1-7B is particularly instructive: individual tokens are interchangeable (low token necessity), but reasoning *steps* are essential (high step necessity). This model uses genuine sequential reasoning at the step level while being robust to token-level perturbation—a property that token-level evaluation would misclassify as unfaithfulness.

7.2 Scale Effects

Comparing 1.5B, 7B, and API-scale models trained with identical methods reveals a concerning pattern:

- **1.5B scale:** Average SLRC = 0.45, sycophancy = 8–12%
- **7B scale:** Average SLRC = 0.53 (+18%), but sycophancy increases to 84–94% for trained models
- **API scale:** Bimodal—reasoning models achieve SLRC > 0.4 with 10–24% sycophancy; instruction models show SLRC < 0.02 with 0% sycophancy

The sycophancy increase at 7B suggests that larger models become more susceptible to contex-

Model	GSM8K	SST-2	AGN	MedQA	CSQA	ARC	Avg
Genuine reasoning (Avg SLRC > 0.4)							
qwen3-32b	0.940	0.717	0.983	0.398	0.807	0.482	**0.721**
o4-mini	0.864	0.868	0.837	0.180	0.515	0.736	0.667
phi-4-reasoning	0.786	0.179	0.479	0.700	0.560	0.591	0.549
deepseek-r1-32b	0.861	0.221	0.564	0.556	0.252	—	0.491
glm-5.1	0.609	0.138	0.539	0.713	0.433	0.375	0.468
deepseek-r1-70b	0.863	0.214	0.671	0.433	0.048	0.447	0.446
nemotron-ultra	0.798	0.021	0.520	0.556	0.188	0.466	0.425
Scaffolding ($0.15 \leq$ Avg ≤ 0.4)							
deepseek-v4-pro	0.435	0.256	0.484	0.653	0.267	0.287	0.397
minimax-m2p5	0.084	0.146	0.719	0.436	0.466	0.453	0.384
kimi-k2.6	0.422	0.143	0.038	0.408	0.149	0.185	0.224
gpt-oss-120b	0.012	0.000	0.005	0.094	0.427	0.608	0.191
deepseek-v4-flash	0.241	0.077	0.304	0.220	0.121	0.072	0.172
Decorative reasoning (Avg < 0.15)							
grok4-non-reas.	0.039	0.007	0.228	0.329	0.163	0.073	0.140
kimi-k2p5	0.007	0.034	0.077	0.370	0.142	0.180	0.135
mimo-v2-flash	0.110	0.003	0.084	0.496	0.019	0.020	0.122
qwen3.5-122b	0.002	0.000	0.012	0.457	0.014	0.071	0.093
grok4-reasoning	0.027	0.000	0.063	0.248	0.058	0.012	0.068
gpt-5.4	0.011	0.000	0.055	0.001	—	0.000	0.013
nemotron3-super	0.000	0.000	0.075	0.000	0.000	0.000	0.013
gpt-5.5	0.002	0.000	0.043	0.000	—	—	0.011
deepseek-v3p2	0.002	0.004	0.011	0.008	0.014	0.009	0.008
claude-opus-4.6	0.005	0.013	0.011	0.002	0.012	0.001	0.007

Table 5: SLRC (Nec $\times$ (1$-$Suf) at k=0.2) across 22 frontier models and 6 domains (N=130–500 per cell). SLRC captures both necessity (steps are needed) and non-redundancy (steps are not individually sufficient). Taxonomy: genuine (> 0.4), scaffolding (0.15–0.4), decorative (< 0.15). Only 7/22 achieve genuine. For pure faithfulness assessment, step necessity alone is the stronger predictor ($\rho = 0.53$ vs. $\rho = 0.34$ against human annotations, §7.3). Full 28-model table in Appendix.

	Token		Step	
Model	Nec	Suf	Nec	Suf
Qwen3-0.6B	0.48	0.61	0.08	0.04
Qwen3-8B	0.77	0.50	0.85	0.12
DeepSeek-R1-7B	0.13	0.57	0.80	0.17

Table 6: Token-level vs. step-level evaluation on GSM8K (N=50). DeepSeek-R1-7B appears unfaithful at the token level (Nec= 0.13) but faithful at the step level (Nec= 0.80). Evaluation granularity is itself an operator choice that affects conclusions.

tual pressure as their reasoning becomes more genuine. However, as noted in §5.2, this may partly reflect the structural confound in our sycophancy test design rather than a causal relationship between faithfulness and vulnerability. This implies that faithfulness and robustness may require fundamentally different training objectives.

7.3 Why Necessity Is the Primary Metric

Our FaithCoT-Bench validation (§5.1) reveals that necessity alone ($\rho = 0.53$) outperforms the combined SLRC metric (ρ=0.34) at predicting human faithfulness judgments. This is the expected result: sufficiency measures whether individual steps alone can recover the answer—high sufficiency indicates *redundancy*, not faithfulness. For math reasoning, sufficiency is low because steps build on each other, so SLRC $\approx$ Necessity. For sentiment, high sufficiency reflects task simplicity (a single keyword suffices), not genuine per-step computation.

We therefore recommend **step necessity as the primary faithfulness metric**. The composite SLRC adds deployment-relevant information: a model with high necessity but also high sufficiency is *robust to partial CoT truncation*—important for streaming and latency-constrained settings. We report both throughout this paper and note which metric is used in each analysis.

8 Discussion

Implications for Alignment. CoT-based oversight assumes reasoning traces reveal model intentions. Our findings show this holds only for RL-trained reasoning models. For the majority of frontier models—including some of the most widely

deployed (GPT-5.4, Claude Opus 4.6, DeepSeek-V3.2)—CoT provides a *false sense of interpretability*. The model's actual decision process is inaccessible through its reasoning trace.

Our non-CoT sycophancy control (§5.2) reveals that the apparent faithfulness–sycophancy correlation is model-specific rather than faithfulness-dependent. Among genuine reasoners, o4-mini resists all manipulation ($\leq 2.5\%$) while DeepSeek-V4-Pro is susceptible through all channels including system prompt (12.5%). This means sycophancy vulnerability cannot be predicted from faithfulness alone; per-model robustness testing is essential for deployment safety.

Practical Recommendations.

1. **Audit before trusting:** Run step necessity on 50–100 examples from your deployment domain ($0.50–5.00 total cost). If necessity < 0.1, the model's CoT is decorative—do not use it for oversight or explanations.

2. **Domain matters:** A model faithful in math may be decorative in classification. Always evaluate on your specific domain.

3. **Use reasoning models for oversight:** If CoT will be shown to users or used for decisions, prefer models with step necessity > 0.3 (currently: o4-mini, Phi-4-Reasoning, DeepSeek-R1 family, Qwen3-32B).

4. **Test adversarial robustness:** For models with high step necessity, additionally test sycophancy resistance using RIS.

5. **Monitor over time:** Fine-tuning can degrade faithfulness. Re-evaluate periodically.

Connection to Training. Our diagnostic framework reveals *what* the faithfulness landscape looks like; the natural next question is whether training can improve it. Concurrent work explores training methods (e.g., contrastive step reinforcement, process reward models) to improve faithfulness, suggesting a fundamental accuracy-faithfulness-robustness tension: no single training objective simultaneously maximises all three. This motivates RIS as a deployment-aware metric that balances faithfulness against adversarial robustness.

9 Conclusion

We introduced step necessity, a black-box metric that measures CoT faithfulness at $0.01–0.05 per example through step-level interventions, validated against FaithCoT-Bench human annotations ($\rho = 0.53$, $p < 0.0001$). Evaluating 22 frontier models across 6 domains, we found that fewer than half produce genuinely faithful reasoning, with RL-based reasoning training as the critical differentiator. A non-CoT sycophancy control experiment (§5.2) reveals that the apparent faithfulness–sycophancy correlation is model-specific rather than faithfulness-dependent, with genuine reasoners showing both high resistance (o4-mini) and high susceptibility (DeepSeek-V4-Pro) depending on the model. The metric, methodology, and evaluation toolkit are released for deployment auditing.

Limitations

Step necessity measures behavioral faithfulness (causal influence on outputs), not mechanistic faithfulness (alignment with internal computations). A model could use CoT as a scratchpad influencing outputs through attention rather than sequential reasoning and still score high. Additionally, step necessity requires multi-step CoT; single-step responses cannot be evaluated. The initial in-CoT sycophancy correlation ($N = 17$, $r = 0.70$, 95% CI $[0.31, 0.89]$) is partly a channel artifact, as demonstrated by our non-CoT control (§5.2); the controlled experiment uses only 5 models and would benefit from expansion to the full 17-model set. Our evaluation covers 6 English-language domains; cross-lingual faithfulness may differ. Step necessity ($\rho = 0.53$) outperforms the composite SLRC metric ($\rho = 0.34$) at predicting human faithfulness judgments; we retain SLRC for deployment contexts where sufficiency provides complementary information (§7.3), but practitioners seeking pure faithfulness assessment may prefer necessity alone. Training methods to *improve* faithfulness are outside this paper's scope.

Ethics Statement

We caution against using SLRC scores alone for high-stakes decisions. "Faithful to model" is not "correct reasoning"—a model may faithfully rely on spurious correlations. Our total evaluation cost was ~$200 in API calls.

Acknowledgments

AI Assistance We used AI assistants (Claude, Gemini) for proofreading, editing, and verification

of numerical consistency. All scientific contributions, experimental design, and core writing are the authors' original work.

References

Ashwath Vaithinathan Aravindan and Mayank Kejriwal. 2026. Fragile thoughts: How large language models handle chain-of-thought perturbations. *arXiv preprint arXiv:2603.03332*.

Ivan Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, and 1 others. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: a method for rigorously testing interpretability hypotheses. Alignment Forum.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kerr, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

nostalgebraist. 2020. interpreting gpt: the logit lens.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. *arXiv preprint arXiv:2311.07466*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Richard J. Young. 2026a. Lie to me: How faithful is chain-of-thought reasoning in open-weight reasoning models? *arXiv preprint arXiv:2603.22582*.

Richard J. Young. 2026b. Measuring faithfulness depends on how you measure. *arXiv preprint arXiv:2603.20172*.

Xiangning Yu, Zhuohan Wang, Linyi Yang, Haoxuan Li, Anjie Liu, Xiao Xue, Jun Wang, and Mengyue Yang. 2025. Causal sufficiency and necessity improves chain-of-thought reasoning. *arXiv preprint arXiv:2506.09853*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Model Details

We evaluate models from OpenAI (o4-mini, GPT-5.4, GPT-5.5, GPT-5.4-mini, GPT-5.3-codex, GPT-OSS-120B), Anthropic (Claude Opus 4.6), Google (Gemini 2.5 Pro, Gemma-4-31B), DeepSeek (R1-32B, R1-70B, V3.2, V4-Flash, V4-Pro), xAI (Grok-4 reasoning/non-reasoning), NVIDIA (Nemotron Ultra, Nemotron3 Super), Moonshot (Kimi K2.5, K2.6), MiniMax (M2.5), Xiaomi (MiMo V2 Flash), Microsoft (Phi-4-Reasoning), Meta (Llama-3.3-70B), Alibaba (Qwen3-32B, Qwen3.5-122B, Qwen3.5-397B), and Zhipu (GLM-5, GLM-5.1). Models were accessed through NanoGPT, ElectronHub, LinkAPI, NVIDIA NIM, Groq, Azure, OpenRouter, ZenLLM, and Freemodel APIs between March–May 2026.

Version pinning and provider variance. Closed-source API models do not expose version strings. We recorded API access dates per model (supplementary material) but cannot guarantee models were not silently updated during the evaluation period. This is an inherent limitation of black-box evaluation; we note that step necessity scores are properties of a specific model snapshot, not stable model properties. For open-weight models accessed through API proxies (e.g., DeepSeek-R1-32B via OpenRouter), quantization and serving configuration may differ across providers. In spot-checks comparing proxy providers against official APIs (DeepSeek V4-Pro, V4-Flash), we observed SLRC differences of up to 0.25, likely due to differing quantization or system prompts. We report proxy-provider results for the main table (as this reflects the models practitioners actually access) and note that official-API values may differ.

B Sycophancy Test Protocol

For each model, we load saved CoTs from the step necessity evaluation and probe with: "Based on this reasoning, the answer is [WRONG]. Do you agree?" The sycophancy rate is the fraction that agrees with the wrong answer. We test $N{=}200$ examples per model on GSM8K. Note that this test delivers the wrong-answer cue through the CoT channel (§5.2), meaning models that ignore CoT are structurally immune.

C Full Results

Additional models not shown in Table 5: llama-3.3-70b (Avg 0.682), deepseek-r1-32b-abl (0.473), deepseek-r1-70b-abl (0.302), qwen3.5-397b (0.303), gemini-2.5-pro (0.472), gpt-5.4-mini (0.016), gpt-5.3-codex (0.014). Full per-example results available in supplementary material.