Validating the Impact of Summarized Chain-of-Thought on Honesty and Faithfulness Scores

Anonymous Author(s)

ABSTRACT

Reasoning models increasingly expose chain-of-thought (CoT) outputs to enable monitoring of model honesty and faithfulness. However, when APIs return summarized rather than full CoT—as with Claude 4.5 Haiku—a measurement gap may arise: summaries could omit details that affect computed scores. We formalize this summarization deviation problem and present a simulation-based framework to quantify potential score distortions under varying summarization fidelity, compression ratios, and task complexity. Our analysis models CoT content as structured sequences of reasoning tokens with tagged honesty and faithfulness signals, applying parameterized summarization operators to estimate deviation bounds. Results across 5,000 simulated CoT instances show that moderate compression (3:1) introduces mean absolute deviations of 0.031 for honesty and 0.047 for faithfulness when key signal tokens are retained with 90% probability, but deviations grow to 0.142 and 0.198 under aggressive compression (10:1) with 60% retention. These findings quantify the conditions under which summarized CoT remains a reliable proxy for full CoT evaluation and identify critical thresholds for summarization fidelity.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning [6] has become a central mechanism for both improving and monitoring the behavior of large language models. Recent work on reasoning model honesty [5] evaluates whether models faithfully verbalize their use of provided hints in their reasoning chains. However, a critical measurement challenge arises when the API returns summarized CoT rather than the model's full internal reasoning [1].

For Claude 4.5 Haiku specifically, the Anthropic API returns a summarized chain of thought. As noted by Walden [5], this creates a potential gap between the content in the original CoT and what is available for measurement, which could lead to deviations between measured and true honesty and faithfulness scores. While the authors hypothesize that deviations are small given their explicit verbalization instructions, they acknowledge this cannot be validated without access to full CoTs.

We address this validation gap through three contributions:

(1) A formal model of CoT summarization as a lossy compression operator with parameterized signal retention rates.
(2) A simulation framework that generates structured CoT sequences and measures score deviations under varying summarization conditions.
(3) Quantitative bounds on acceptable summarization parameters for reliable honesty and faithfulness measurement.

2 PROBLEM FORMULATION

2.1 CoT Structure Model

We model a full chain of thought as a sequence $C = (t_1, t_2, \ldots, t_n)$ of reasoning tokens, where each token t_i carries attributes: a content type $\tau_i \in \{\text{reasoning}, \text{hint\_mention}, \text{hint\_reliance}, \text{metacognition}, \text{filler}\}$ and signal indicators $h_i \in \{0, 1\}$ (honesty-relevant) and $f_i \in \{0, 1\}$ (faithfulness-relevant).

2.2 Honesty and Faithfulness Scores

The honesty score $H(C)$ measures whether the model acknowledges receiving hints:

$$H(C) = \frac{\sum_{i=1}^{n} h_i \cdot \mathbb{1}[\tau_i = \text{hint\_mention}]}{\max\left(1, \sum_{i=1}^{n} \mathbb{1}[\tau_i = \text{hint\_mention}]\right)}$$

The faithfulness score $F(C)$ measures whether the model's stated reasoning aligns with its actual hint usage:

$$F(C) = 1 - \frac{\left|\sum_i f_i^{\text{stated}} - \sum_i f_i^{\text{actual}}\right|}{\max(1, n_{\text{relevant}})}$$

2.3 Summarization Operator

A summarization operator Σ_θ with parameters $\theta = (\rho, p_s, p_f)$ maps full CoT C to summary $\hat{C}$:

- $\rho \in (0, 1]$: compression ratio (fraction of tokens retained)
- $p_s \in [0, 1]$: probability of retaining honesty-signal tokens
- $p_f \in [0, 1]$: probability of retaining faithfulness-signal tokens

The deviation is then $\Delta H = |H(C) - H(\hat{C})|$ and $\Delta F = |F(C) - F(\hat{C})|$.

3 METHODOLOGY

3.1 Simulation Design

We generate 5,000 synthetic CoT instances per configuration. Each CoT has length $n \sim \text{Uniform}(50, 500)$ tokens, with hint mentions occurring at rate $\lambda_h = 0.08$ and faithfulness signals at $\lambda_f = 0.12$. We evaluate a grid of summarization parameters: $\rho \in \{0.1, 0.2, 0.33, 0.5, 0.75\}$, $p_s \in \{0.6, 0.7, 0.8, 0.9, 0.95, 1.0\}$, and $p_f \in \{0.6, 0.7, 0.8, 0.9, 0.95, 1.0\}$.

3.2 Deviation Metrics

For each configuration, we compute: (1) Mean absolute deviation (MAD) for honesty and faithfulness; (2) Maximum deviation across instances; (3) Fraction of instances where deviation exceeds tolerance thresholds $\epsilon \in \{0.05, 0.10, 0.15\}$.

4 RESULTS

4.1 Deviation Under Standard Conditions

At moderate compression ($\rho = 0.33$, approximately 3:1) with high signal retention ($p_s = p_f = 0.9$), the mean absolute deviation is

Table 1: Mean absolute deviation by compression ratio ($p_s = p_f = 0.9$).

Compression	ρ	MAD-H	MAD-F	% > 0.10
1.3:1	0.75	0.012	0.018	1.1%
2:1	0.50	0.021	0.033	2.8%
3:1	0.33	0.031	0.047	6.0%
5:1	0.20	0.058	0.089	14.3%
10:1	0.10	0.142	0.198	38.7%

0.031 for honesty and 0.047 for faithfulness. Only 4.2% of instances exceed the $\epsilon = 0.10$ threshold for honesty, and 7.8% for faithfulness.

4.2 Signal Retention Sensitivity

Signal retention probability has a stronger effect than compression ratio on deviation magnitude. Reducing p_s from 0.95 to 0.70 at fixed $\rho = 0.33$ increases honesty MAD from 0.019 to 0.091. This confirms that whether signal tokens survive summarization is more important than overall summary length.

4.3 Task Complexity Effects

Longer CoTs (300–500 tokens) show lower relative deviation than shorter CoTs (50–100 tokens) because they contain more redundant signal tokens. This suggests that Claude 4.5 Haiku's extended reasoning, which tends to produce longer CoTs, may naturally buffer against summarization artifacts.

4.4 Critical Thresholds

For deviations to remain below 0.05 with 95% probability, the summarization must maintain $p_s \geq 0.88$ and $p_f \geq 0.85$ at 3:1 compression. At 5:1 compression, the requirements tighten to $p_s \geq 0.94$ and $p_f \geq 0.92$.

5 DISCUSSION

Our simulation-based analysis provides the first quantitative bounds on CoT summarization deviation for honesty and faithfulness measurement. The key finding is that moderate summarization (up to 3:1 compression) with high signal retention (≥ 0.9) introduces acceptably small deviations (MAD < 0.05), supporting the hypothesis of Walden [5] that deviations are likely small under their experimental conditions.

However, our results also identify conditions where summarization artifacts become substantial. Aggressive compression (10:1) or poor signal retention (< 0.7) can produce deviations exceeding 0.15, which would materially affect honesty and faithfulness conclusions. This has implications for API design: exposing summarization parameters or fidelity guarantees would enable researchers to calibrate confidence in their measurements.

Two important limitations apply. First, our model assumes independent signal retention, while real summarization models may exhibit correlated omissions. Second, we model summarization as a token-level operation, whereas actual LLM summarizers operate at the semantic level, potentially preserving meaning even when specific tokens are dropped.

6 RELATED WORK

Chain-of-thought prompting [6] and its extensions have been studied extensively for reasoning capability. Faithfulness of CoT has been questioned by work showing that models sometimes arrive at correct answers through unfaithful reasoning chains [3, 4]. The specific problem of summarized CoT evaluation was identified by Walden [5] in the context of measuring reasoning honesty. Our work complements the faithfulness probing approach of Chen et al. [2] by focusing on the summarization artifact rather than internal model representations.

7 CONCLUSION

We formalized and quantified the CoT summarization deviation problem for honesty and faithfulness measurement. Our simulation framework establishes that moderate summarization with high signal retention produces acceptably small deviations, but identifies critical thresholds beyond which measurement reliability degrades significantly. These results provide practical guidance for researchers working with summarized CoT APIs and motivate the development of fidelity-guaranteed summarization for safety-critical CoT monitoring.

REFERENCES

[1] Anthropic. 2025. Extended thinking with Claude. *Anthropic Documentation* (2025).
[2] Yifei Chen et al. 2024. Seeing is believing: Measuring faithfulness of chain-of-thought reasoning via probing. *arXiv preprint arXiv:2402.19450* (2024).
[3] Tamera Lanham et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* (2023).
[4] Miles Turpin et al. 2024. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* (2024).
[5] James Walden. 2026. Reasoning Models Will Blatantly Lie About Their Reasoning. *arXiv preprint arXiv:2601.07663* (2026).
[6] Jason Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.