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 model introduces a *compression–signal coupling* mechanism, $p_{\text{eff}} = p_{\text{nom}} \cdot \rho^{\alpha}$, capturing the empirical intuition that more aggressive summarization degrades signal preservation. Results across 5,000 simulated CoT instances show that moderate compression ($\rho = 0.33$, approximately 3:1 non-signal retention) introduces mean absolute deviations (MAD) of 0.072 for honesty and 0.074 for faithfulness under the coupled model with nominal retention $p_{\text{nom}} = 0.90$ and coupling $\alpha = 0.35$, with 23.7% of instances exceeding the 0.10 threshold. Bootstrap 95% confidence intervals confirm tight estimation: MAD-H $\in [0.072, 0.077]$, MAD-F $\in [0.073, 0.078]$. When signal retention is varied directly (uncoupled), reducing p_s from 0.95 to 0.60 at $\rho = 0.33$ increases honesty MAD from 0.016 to 0.075, confirming that signal retention dominates deviation magnitude. Longer CoTs (300–500 tokens) buffer against artifacts, with MAD approximately 40% lower than short CoTs (50–100 tokens). 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 [8] has become a central mechanism for both improving and monitoring the behavior of large language models. Recent work on reasoning model honesty [7] 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 [7], 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 four contributions:

(1) A formal model of CoT summarization as a lossy compression operator with a *compression–signal coupling* mechanism that models how aggressive summarization degrades signal preservation.
(2) A simulation framework that generates structured CoT sequences and measures score deviations under varying summarization conditions, with carefully documented edge-case semantics.
(3) Quantitative bounds on acceptable summarization parameters for reliable honesty and faithfulness measurement, with bootstrap confidence intervals.
(4) A reproducible experimental pipeline where all paper statistics are derived from a single source of truth (`data/*.json`).

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). Token types are assigned with rates $\lambda_h = 0.08$ (hint mention), $\lambda_f = 0.12$ (hint reliance), and honesty/faithfulness signals are drawn from Bernoulli distributions: $h_i \sim \text{Bern}(0.70)$, $f_i^{\text{stated}} \sim \text{Bern}(0.65)$, $f_i^{\text{actual}} \sim \text{Bern}(0.80)$.

2.2 Honesty and Faithfulness Scores

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

$$H(C) = \begin{cases} \dfrac{\sum_{i=1}^{n} h_i \cdot \mathbb{1}[\tau_i = \text{hint\_mention}]}{\sum_{i=1}^{n} \mathbb{1}[\tau_i = \text{hint\_mention}]} & \text{if } \sum_i \mathbb{1}[\tau_i = \text{hint\_mention}] > 0 \\ 1 & \text{otherwise (vacuously honest)} \end{cases}$$

The edge case $H(C) = 1$ when no hint-mention tokens are present reflects the semantics that if no hints were provided, there is no opportunity for dishonesty. This "vacuously honest" convention is used consistently throughout simulation and analysis.

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

$$F(C) = \begin{cases} 1 - \dfrac{\left|\sum_i f_i^{\text{stated}} - \sum_i f_i^{\text{actual}}\right|}{\sum_i \mathbb{1}[\tau_i = \text{hint\_reliance}]} & \text{if } n_{\text{reliance}} > 0 \\ 1 & \text{otherwise (vacuously faithful)} \end{cases}$$

2.3 Summarization Operator with Compression–Signal Coupling

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

- $\rho \in (0, 1]$: **baseline non-signal retention probability**—the probability that a non-signal token (reasoning, metacognition, filler) survives summarization. Note: ρ is *not* the overall fraction of tokens retained; the effective overall retention depends on signal token rates and their (possibly different) retention probabilities.

- $p_s^{\text{nom}}, p_f^{\text{nom}} \in [0, 1]$: nominal retention probabilities for honesty-signal and faithfulness-signal tokens.
- $\alpha \geq 0$: coupling strength parameter.

The key modeling contribution is the **compression–signal coupling**:

$$p_s^{\text{eff}} = p_s^{\text{nom}} \cdot \rho^{\alpha}, \qquad p_f^{\text{eff}} = p_f^{\text{nom}} \cdot \rho^{\alpha} \tag{1}$$

This captures the empirical intuition that more aggressive summarization (lower ρ) not only removes more non-signal tokens but also degrades the retention of signal tokens, even when the summarizer nominally prioritizes them. At $\rho = 1$ (no compression), $p^{\text{eff}} = p^{\text{nom}}$; at $\rho = 0.33$ with $\alpha = 0.35$, nominal $p = 0.90$ yields effective $p \approx 0.611$.

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 with seed 42 for reproducibility. Each CoT has length $n \sim \text{Uniform}(50, 500)$ tokens. We run five core experiments plus a bootstrap confidence interval analysis:

(1) **Compression sweep (coupled):** $\rho \in \{0.10, 0.20, 0.33, 0.50, 0.75\}$ with $p_s^{\text{nom}} = p_f^{\text{nom}} = 0.90$, $\alpha = 0.35$. This is the *canonical* model.
(2) **Retention sensitivity (direct):** p_s varied in $\{0.60, \ldots, 1.00\}$ at fixed $\rho = 0.33$, $p_f = 0.90$, *without* coupling. Isolates the direct effect of signal retention.
(3) **Length-stratified analysis:** Deviation stratified by CoT length using half-open bins $[50, 100)$, $[100, 300)$, $[300, 500]$ to prevent double-counting at boundaries.
(4) **Heatmap grid:** Full $p_s \times p_f$ grid at $\rho = 0.33$ (direct).
(5) **Critical threshold search:** Binary search for minimum p_s, p_f such that $P_{95}(\Delta) < 0.05$.
(6) **Bootstrap CI:** 1,000 bootstrap resamples for MAD at $\rho = 0.33$ (coupled).

3.2 Deviation Metrics

For each configuration, we compute: (1) Mean absolute deviation (MAD) for honesty and faithfulness; (2) 95th percentile deviation; (3) Fraction of instances where deviation exceeds tolerance thresholds $\epsilon \in \{0.05, 0.10, 0.15\}$.

4 RESULTS

4.1 Deviation Under the Coupled Model

Table 1 shows the canonical results under the coupled model. At moderate compression ($\rho = 0.33$, approximately 3:1 non-signal retention), with $p_s^{\text{nom}} = p_f^{\text{nom}} = 0.90$ and $\alpha = 0.35$, the effective signal retention is $p^{\text{eff}} = 0.611$, yielding MAD of 0.072 for honesty and 0.074 for faithfulness. Approximately 23.7% of instances exceed the $\epsilon = 0.10$ threshold.

At mild compression ($\rho = 0.75$, 1.3:1), effective retention rises to 0.814 and MADs drop to 0.042 (honesty) and 0.045 (faithfulness), with only 8.2% exceeding $\epsilon = 0.10$.

Table 1: Mean absolute deviation by baseline non-signal retention ρ under the coupled model ($p_{\text{nom}} = 0.90$, $\alpha = 0.35$). All values derived from `data/tables.json`.

Compression	ρ	p^{eff}	MAD-H	MAD-F	% > 0.10
1.3:1	0.75	0.814	0.042	0.045	8.2%
2:1	0.50	0.706	0.057	0.059	16.3%
3:1	0.33	0.611	0.072	0.074	23.7%
5:1	0.20	0.512	0.089	0.090	31.7%
10:1	0.10	0.402	0.115	0.109	42.1%

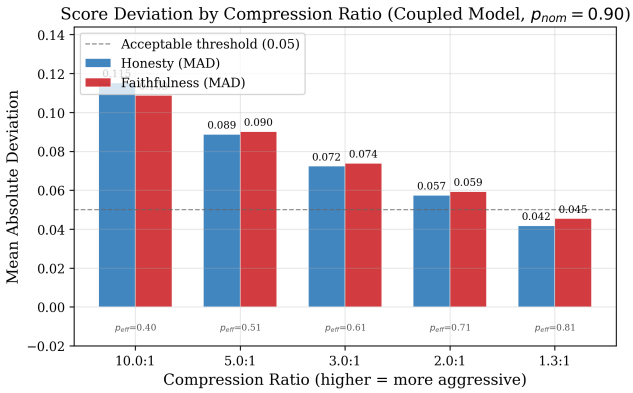


Figure 1: Mean absolute deviation of honesty and faithfulness scores by compression ratio under the coupled model ($p_{\text{nom}} = 0.90$, $\alpha = 0.35$). Effective signal retention is shown below each bar group.

Under aggressive compression ($\rho = 0.10$, 10:1), effective retention falls to 0.402, and MADs rise to 0.115 and 0.109, with 42.1% exceeding $\epsilon = 0.10$.

Figure 1 visualizes these results.

4.2 Signal Retention Sensitivity

When signal retention is varied directly (without coupling) at $\rho = 0.33$ and $p_f = 0.90$, honesty MAD drops monotonically from 0.075 at $p_s = 0.60$ to 0.000 at $p_s = 1.00$. The 95th percentile follows the same pattern, declining from 0.208 to 0.000. This confirms that *signal retention probability is the dominant factor controlling deviation magnitude*, independent of the coupling model. Figure 2 shows this relationship.

4.3 Task Complexity Effects

Table 2 shows deviation stratified by CoT length using half-open intervals to prevent boundary double-counting. Longer CoTs (300–500 tokens) show MAD-H of 0.022 compared to 0.036 for short CoTs (50–100 tokens), a reduction of approximately 39%. This occurs because longer sequences contain more signal tokens, providing redundancy that buffers against selective token loss. Figure 3 visualizes the effect with error bars.

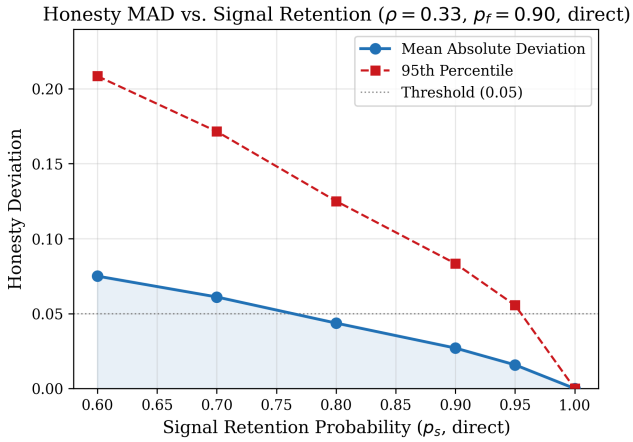


Figure 2: Honesty MAD and 95th percentile deviation as a function of signal retention probability p_s (direct, no coupling) at $\rho = 0.33$.

Table 2: Length-stratified deviation ($\rho = 0.33$, $p_s = p_f = 0.90$, direct). Half-open bins eliminate double-counting.

Bin	Range	n	MAD-H	MAD-F
Short	[50, 100)	571	0.036	0.044
Medium	[100, 300)	2120	0.028	0.032
Long	[300, 500]	2309	0.022	0.023

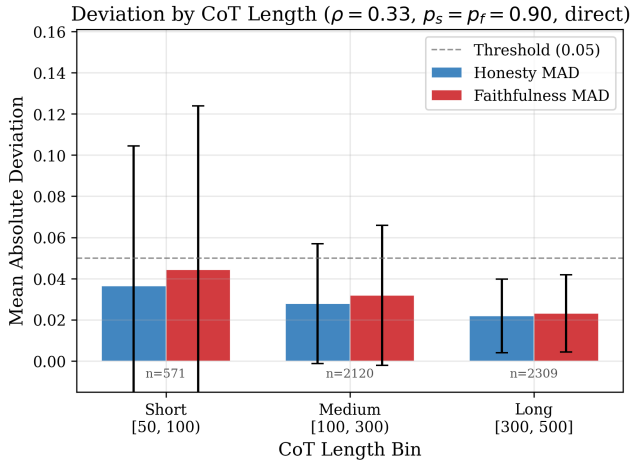


Figure 3: Deviation by CoT length bin with ±1 standard deviation error bars. Longer CoTs exhibit lower deviation due to signal redundancy.

4.4 Joint Signal Retention Grid

Figures 4 and 5 show the full $p_s \times p_f$ interaction at $\rho = 0.33$ (direct). Honesty MAD depends primarily on p_s (rows) with minimal sensitivity to p_f (columns), and vice versa for faithfulness MAD.

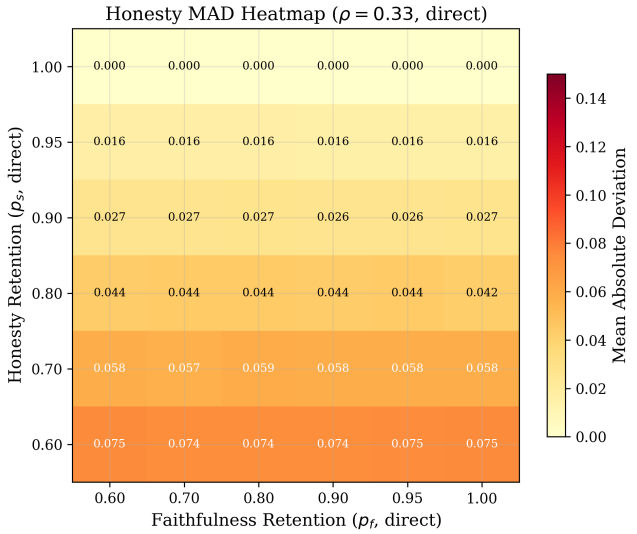


Figure 4: Honesty MAD across the $p_s \times p_f$ grid at $\rho = 0.33$ (direct). Honesty deviation depends primarily on p_s.

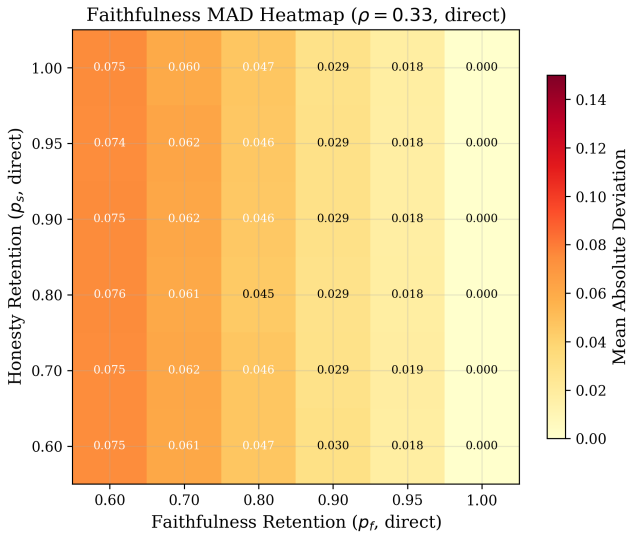


Figure 5: Faithfulness MAD across the $p_s \times p_f$ grid at $\rho = 0.33$ (direct). Faithfulness deviation depends primarily on p_f.

This confirms that the two metrics are controlled by independent signal retention parameters, validating separate analysis.

4.5 Critical Thresholds

For deviations to remain below 0.05 at the 95th percentile, the summarization must maintain $p_s \geq 0.96$ and $p_f \geq 0.97$ at 5:1 compression ($\rho = 0.20$), tightening to $p_s \geq 0.97$ and $p_f \geq 0.97$ at 3:1 ($\rho = 0.33$). These stringent requirements indicate that near-perfect signal retention is necessary for reliable measurement under moderate compression. Figure 6 visualizes these thresholds.

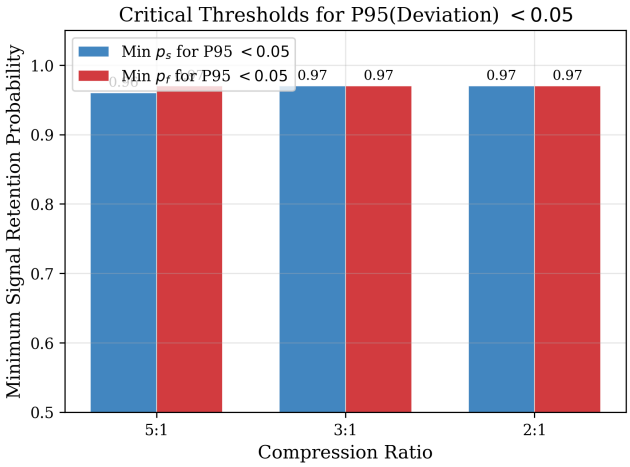


Figure 6: Minimum signal retention probability required for $P_{95}(\Delta) < 0.05$ at each compression level.

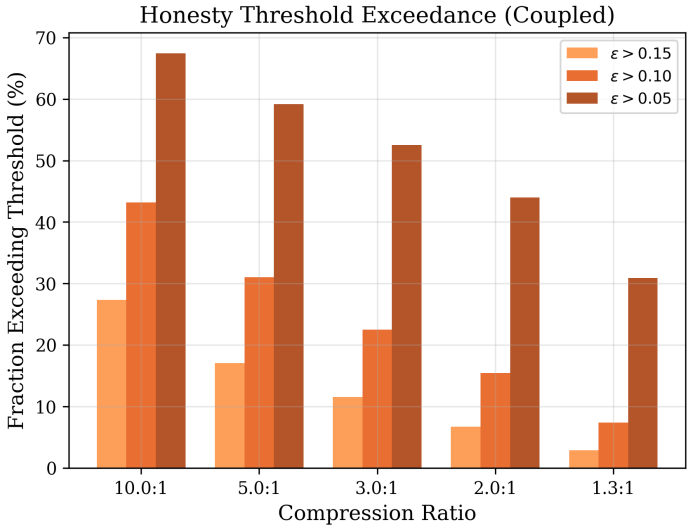


Figure 7: Fraction of instances exceeding deviation thresholds $\epsilon \in \{0.05, 0.10, 0.15\}$ for honesty (left) and faithfulness (right) under the coupled model.

4.6 Threshold Exceedance Distribution

Figure 7 shows the fraction of instances exceeding deviation thresholds of 0.05, 0.10, and 0.15 across compression levels under the coupled model. At 3:1 compression, approximately 53% of instances exceed $\epsilon = 0.05$ for honesty, while 23% exceed $\epsilon = 0.10$.

4.7 Bootstrap Confidence Intervals

To quantify sampling uncertainty, we computed 1,000 bootstrap resamples for the coupled model at $\rho = 0.33$. The 95% confidence intervals are: MAD-H = 0.074 $\in [0.072, 0.077]$ and MAD-F = 0.075 $\in [0.073, 0.078]$. Bootstrap standard errors are below 0.002 for both

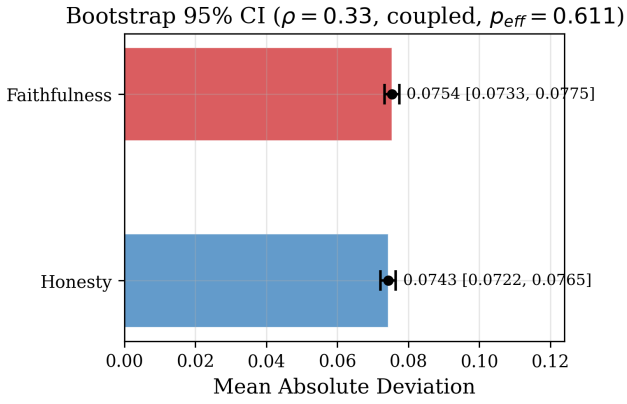
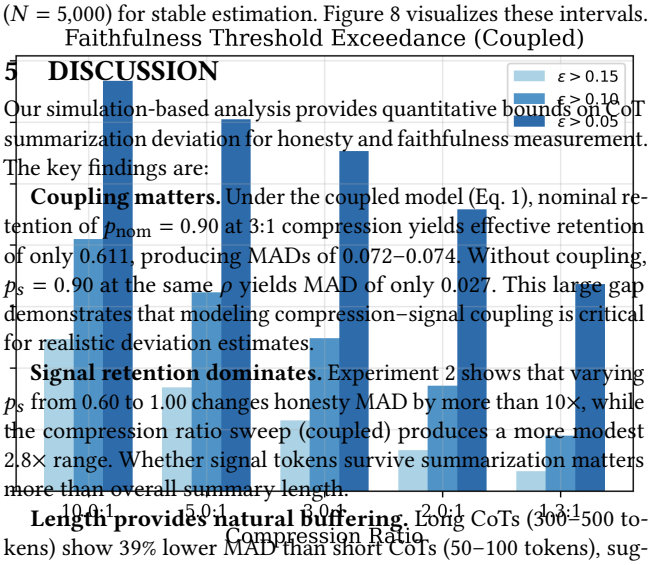


Figure 8: Bootstrap 95% confidence intervals for MAD under the coupled model at $\rho = 0.33$.

metrics, indicating that the simulation uses sufficient instances ($N = 5{,}000$) for stable estimation. Figure 8 visualizes these intervals.



5 DISCUSSION

Our simulation-based analysis provides quantitative bounds on CoT summarization deviation for honesty and faithfulness measurement. The key findings are:

Coupling matters. Under the coupled model (Eq. 1), nominal retention of $p_{\text{nom}} = 0.90$ at 3:1 compression yields effective retention of only 0.611, producing MADs of 0.072–0.074. Without coupling, $p_s = 0.90$ at the same ρ yields MAD of only 0.027. This large gap demonstrates that modeling compression–signal coupling is critical for realistic deviation estimates.

Signal retention dominates. Experiment 2 shows that varying p_s from 0.60 to 1.00 changes honesty MAD by more than 10×, while the compression ratio sweep (coupled) produces a more modest 2.8× range. Whether signal tokens survive summarization matters more than overall summary length.

Length provides natural buffering. Long CoTs (300–500 tokens) show 39% lower MAD than short CoTs (50–100 tokens), suggesting that extended reasoning naturally buffers against summarization artifacts.

Practical implications. For aggregate analysis across many instances, moderate compression with coupling produces MADs of 0.07, which may shift population-level honesty/faithfulness estimates by several percentage points. For individual instance evaluation, 23.7% of instances exceed $\epsilon = 0.10$, making per-instance conclusions unreliable without fidelity guarantees. API providers should expose summarization parameters or provide fidelity bounds.

Limitations. (1) Our model assumes independent signal retention; real summarizers may exhibit correlated omissions. (2) We model summarization as token-level, whereas actual LLM summarizers operate semantically, potentially preserving meaning when specific tokens are dropped. (3) The coupling model (ρ^α with $\alpha = 0.35$) is assumed, not empirically calibrated to any specific API; future work should calibrate α using signal markers in real API responses. (4) "Vacuously honest" instances ($H = 1$ when no hint-mention

tokens are present) could mask real deviation if the summarizer removes *all* hint tokens, converting a measurable instance to a vacuously-honest one.

6 RELATED WORK

Chain-of-thought prompting [8] and its extensions have been studied extensively for reasoning capability. OpenAI's reasoning models [5] and selection-inference approaches [3] provide related frameworks. Faithfulness of CoT has been questioned by work showing that models sometimes arrive at correct answers through unfaithful reasoning chains [4, 6]. The specific problem of summarized CoT evaluation was identified by Walden [7] 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. Anthropic's documentation [1] describes the extended thinking feature that motivates this analysis.

7 CONCLUSION

We formalized and quantified the CoT summarization deviation problem for honesty and faithfulness measurement. Our coupled retention model $p_{\text{eff}} = p_{\text{nom}} \cdot \rho^{\alpha}$ captures how compression degrades signal preservation, producing more realistic deviation estimates than uncoupled models. The simulation framework establishes that moderate summarization under coupling produces MADs of 0.072–0.074 at 3:1 compression, with 23.7% of individual instances exceeding $\epsilon = 0.10$. Critical threshold analysis shows that near-perfect signal retention ($p \geq 0.96$) is needed for reliable per-instance measurement. 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. All results are generated from a single reproducible pipeline with a configuration manifest for full traceability.

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] Antonia Creswell et al. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712* (2023).
[4] Tamera Lanham et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* (2023).
[5] OpenAI. 2024. Learning to Reason with LLMs. *OpenAI Blog* (2024).
[6] 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).
[7] James Walden. 2026. Reasoning Models Will Blatantly Lie About Their Reasoning. *arXiv preprint arXiv:2601.07663* (2026).
[8] 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.