Causally Redundant Reasoning: A Dual-Disruption Test Framework for Measuring Shortcut Tendencies in Chain-of-Thought

Jiwoo Kim and Jongwuk Lee
Sungkyunkwan University, Republic of Korea
{jindog1210, jongwuklee}@skku.edu

Abstract

A chain-of-thought (CoT) rationale can be coherent and answer-consistent while remaining causally unused. We call this phenomenon *shortcut reasoning*: the model appears to reason through CoT, but its answer does not depend on the expressed trace. We formalize shortcut reasoning as *causal redundancy* and propose a dual-disruption framework for measuring shortcut tendencies in CoT by testing two complementary aspects of causal explanation: necessity and sufficiency. We operationalize them as two behavioral signals: whether the answer changes under CoT disruption, and whether the answer can be recovered from the CoT after masking the original input. We instantiate this framework with *CORD* (*Causal Redundancy of Reasoning via Disruption*), an instance-level score that combines both signals. The two axes reduce complementary ambiguities in single-axis tests: answer preservation under CoT disruption can reflect shortcut use or redundant truthful reasoning, while early answer recovery can reflect shortcut use or a compact truthful rationale. Across 6 models and 2 datasets, CORD reveals a training-paradigm effect: R1-Distill models exhibit roughly twice the shortcut tendency of non-distilled models, while the two axes capture related but distinct signals. Together, these results motivate joint tests of necessity and sufficiency for detecting shortcut tendencies that can remain hidden behind correct answers.[1]

1 Introduction

Large Language Models (LLMs) have rapidly advanced in reasoning ability through the combination of large-scale, high-quality pre-training corpora (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022; Penedo et al., 2024) and Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022). Beyond improving task performance, CoT has been widely viewed as a useful trace through which humans can inspect how a model arrives at its answer.

However, it is not guaranteed that the trace reflects the process that actually determines the model's answer. Recent studies show that final answers can remain unchanged even when reasoning steps are modified or removed, and that models can generate plausible rationales that do not faithfully reflect the actual decisions (Turpin et al., 2023; Lanham et al., 2023; Arcuschin et al., 2025). We refer to this behavior as *shortcut reasoning* in CoT.

To make this failure mode measurable, we formulate shortcut reasoning as *causal redundancy* of the CoT for the final answer. In other words, we ask whether the CoT *causally* contributes to producing the answer. We assess this causal role through two complementary aspects of causal explanation: *necessity* and *sufficiency*. In our setting, the CoT M is the candidate cause and the final answer Y is the effect. Necessity asks whether changing the expressed reasoning path changes the resulting answer. Sufficiency asks whether, once the direct contribution of the original input X is removed, the reasoning path itself can sustain the answer.

These causal questions, however, are not directly identifiable from autoregressive generation. We therefore translate them into disruption-based testing. On the necessity side, we ask whether the answer remains unchanged when the CoT is disrupted while the input is preserved. On the sufficiency side, we remove the input and ask whether the baseline answer can already be recovered from only a short CoT prefix. These serve as behavioral proxies for necessity and sufficiency.

Crucially, neither signal is diagnostic on its own. Answer preservation under CoT disruption can reflect either shortcut use or redundant truthful reasoning, while early answer recovery can reflect either shortcut use or a compact truthful rationale. We therefore jointly test necessity and sufficiency, operationalizing them with *inverse decoding* for

[1]Code: https://github.com/JiwooKimAR/cord-cot.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47788–47820
October 24-29, 2026 ©2026 Association for Computational Linguistics

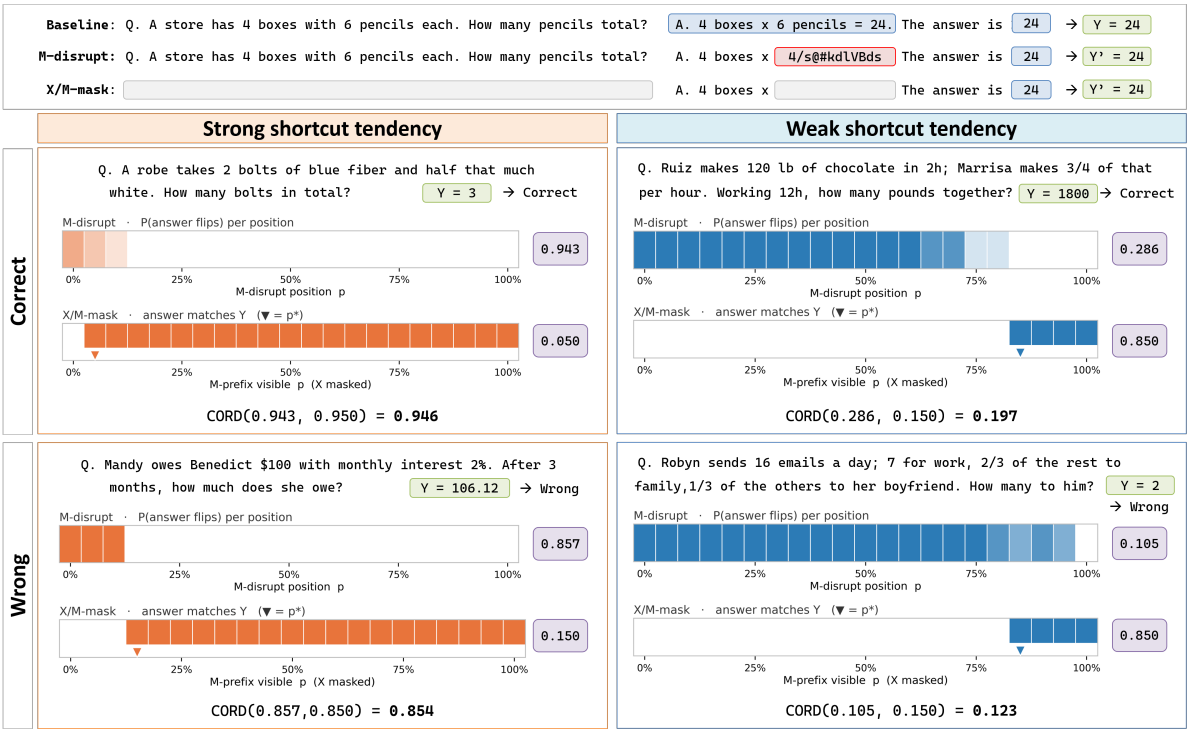


Figure 1: **Correctness and shortcut tendency are independent axes. Top:** CORD combines two disruptions of a CoT M. M-disruption corrupts a span while keeping X visible; answer preservation indicates reduced dependence on the disrupted rationale. X/M-mask hides X and reveals only the first $p\%$ of M; early recovery indicates that the answer is recoverable from a short rationale prefix. **Grid:** four GSM8K examples from DeepSeek-R1-Distill-Qwen-7B, with rows showing correctness and columns CORD. A correct answer can still exhibit strong shortcut tendency, showing that correctness alone cannot detect shortcut reasoning.

CoT disruption and X-mask recovery for isolating answer recovery from the CoT itself. We combine the resulting signals into *CORD* (*Causal Redundancy of Reasoning via Disruption*), an instance-level measure of shortcut tendency.

Applying this framework to 6 models across 2 datasets, we find that CORD reveals systematic differences across training paradigms. In particular, R1-Distill models exhibit approximately twice the CORD of non-distilled models, and this separation is not explained by task accuracy: models with comparable or higher accuracy can exhibit substantially lower CORD. The two CORD components are positively associated in every model–task cell but remain non-interchangeable (mean Spearman $\rho = 0.659$), supporting their complementary roles in the dual-disruption framework. Finally, we validate CORD through self-refine: when refinement behaves consistently with X-only stochastic resampling and leaves the answer unchanged, CORD exhibits the predicted near-zero shift ($\Delta = +0.0019$). Together, these findings show that CORD captures a dimension of functional CoT use not reducible to accuracy or disruption axis alone.

Contributions. Our contributions are as follows.

- **Causal-redundancy formulation of shortcut reasoning.** We formalize shortcut reasoning as a state in which the CoT is causally redundant for the final answer, distinguishing it from related notions such as spurious correlation and hallucination.
- **CORD: a dual-disruption measure of shortcut tendency.** We propose *CORD*, a dual-disruption framework that jointly tests answer robustness to CoT perturbation and answer recovery from CoT without the original input. CORD operationalizes these diagnostics through inverse decoding and input-masked recovery, and summarizes their conjunction as an instance-level behavioral score.
- **Empirical measurement across models and datasets.** Across 6 models and 2 datasets, we identify a clear training-paradigm effect: R1-Distill models show substantially higher CORD than non-distilled models, with related but non-redundant components. We further validate CORD through self-refine.

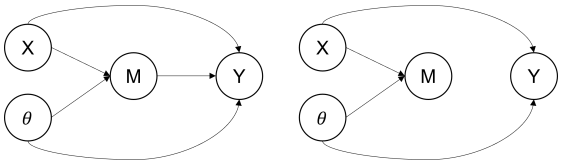
2 Related Work

2.1 Faithfulness Evaluation in LLMs

Recent studies on CoT faithfulness question whether generated rationales reflect the reasoning underlying a model's answer. Lanham et al. (2023) show that final answers can remain unchanged even after CoT steps are modified or removed, providing evidence that generated reasoning can be unfaithful to the model's answer-generation process. Subsequent perturbation-based work extends this line with additional forms of CoT disruption (Aravindan and Kejriwal, 2026). In a complementary line of work, Turpin et al. (2023) show that CoT can serve as a rationalization mechanism under sycophantic bias and Arcuschin et al. (2025) report that unfaithful reasoning persists even in frontier models. Together, these studies establish that a coherent or answer-consistent CoT need not faithfully reflect the basis of a model's answer. However, shortcut reasoning remains underspecified as a target construct: it is often discussed alongside adjacent phenomena such as spurious correlation, hallucination, and post-hoc rationalization, without an explicit formulation of what makes a reasoning trace shortcut-like. Therefore, existing perturbation-based tests provide evidence of CoT sensitivity or robustness, but do not by themselves specify when such behavior constitutes shortcut reasoning. Our work addresses this construct-level gap by formalizing shortcut reasoning as *causal redundancy*.

2.2 Causal Frameworks for LLM Inference

A growing body of work analyzes the causal role of CoT within formal causal frameworks. Paul et al. (2024) separate indirect and direct effects through causal mediation analysis, and Zhang et al. (2025) apply front-door adjustment. Fu et al. (2025a) model CoT with a structural causal model and define step-level Average Causal Effects, while Fu et al. (2025b) classify causal structures of LLM reasoning through intervention experiments. However, it remains unclear whether the identification assumptions are satisfied in autoregressive LLMs. Beyond this issue, Yee et al. (2024) identify unfaithful error recovery, where models reach the correct final answer despite invalid reasoning induced by errors introduced into the CoT. Somov et al. (2026) further report that, even in controlled settings where the mediator and output are linked by a deterministic function, model outputs often fail to reflect the manipulated mediator. These findings compli-



(a) Genuine reasoning (b) Shortcut reasoning

Figure 2: LLM inference as a structural causal model (SCM). X denotes the query, θ the model parameters, M the generated CoT, and Y the final answer.

cate treating manipulations of a generated rationale as clean interventions on the underlying reasoning process. We instead adopt a *disruption-based causal testing* perspective: a reasoning component that functionally matters should leave a measurable behavioral consequence when deliberately disrupted. This follows the general logic of lesion studies and gene knockouts (Vaidya et al., 2019; Hall et al., 2009): if damage to a brain region impairs a behavior, or deletion of a gene alters a phenotype, the resulting change provides evidence about the functional role of the disrupted component. We apply the same logic to CoT, using controlled disruptions to probe whether the expressed reasoning path functionally contributes to the model's answer.

3 Methodology

3.1 Shortcut Reasoning in CoT

3.1.1 Structural Causal Model

We model CoT generation in an LLM using the structural causal model (SCM) in Figure 2(a). Let X denote the input, M the baseline CoT, Y the answer, and θ the model parameters. The reasoning path M is sampled from θ and X, i.e., $M \sim P_\theta(\cdot \mid X)$, and Y is a function of θ, X, and M. Because a decoder-only Transformer can directly attend to the input tokens when generating an answer, the direct edge $X \to Y$ is structurally available. This edge captures the possibility that the model produces an answer without causally relying on the expressed reasoning path M. We denote the domains of X, M, and Y by $\mathcal{X}$, $\mathcal{M}$, and $\mathcal{Y}$, respectively, and use lowercase x, m, y for individual realizations. All notation is summarized in Appendix A.1.

3.1.2 Shortcut Reasoning as Causal Redundancy

Shortcut reasoning in CoT refers to cases where a model generates M, but its content does not causally contribute to Y, and the answer is deter-

mined primarily through the path $X \rightarrow Y \leftarrow \theta$ (Figure 2(b)). We define this behavior as

$$P(Y \mid X, M) \approx P(Y \mid X). \quad (1)$$

That is, given X, the expressed rationale M contributes little additional causal influence on Y; M is *causally redundant* for the answer. Causal redundancy is a spectrum: at one extreme the answer is fully X-determined and M is post-hoc; at intermediate levels, the answer is formed early in M and the remaining reasoning steps are functionally redundant; at the opposite end, the answer depends on the reasoning path and M is not causally redundant. $CORD$ measures tendency along this spectrum rather than a binary shortcut/non-shortcut judgment. This definition is related to, but distinct from, adjacent notions. Spurious correlation concerns reliance on spurious features in the data rather than the causal role of M; hallucination concerns the factual correctness of M or Y; and post-hoc rationalization is one instance of shortcut reasoning under this general definition.

3.1.3 Why Direct Measurement is Difficult

Directly operationalizing the causal redundancy in Equation 1 is difficult in autoregressive LLMs because there is no clean observable condition corresponding to "the same inference process without the causal contribution of M." Simply removing the CoT does not isolate the effect of M: because Y is generated conditioned on the preceding tokens, removing M changes the generation context and induces a different inference trajectory. Likewise, explicitly prompting the model to answer without CoT changes the prompt itself, inducing a different conditioning distribution, $P(Y \mid X, \text{prompt}_{no-CoT})$. We therefore do not measure causal redundancy by directly comparing the baseline with a single no-CoT condition. Instead, we decompose the causal role of M into two complementary facets of causal explanation — *necessity* and *sufficiency*. This decomposition recasts the causal question underlying Equation 1 as two questions that can subsequently be mapped to observable behavioral signatures.

3.2 From Causal Explanation to Behavioral Signature

3.2.1 Necessity and Sufficiency as Two Facets of Causal Explanation

Pearl's probabilities of causation distinguish necessity and sufficiency as two distinct facets of causal explanation (Pearl, 1999). In our setting, the expressed reasoning path M is the candidate cause and the final answer Y is the effect. Assessing whether M is causally involved in producing Y therefore requires two complementary questions:

- **Necessity**: Does changing M change Y?
- **Sufficiency**: After removing the direct contribution of X, can M sustain Y?

These two questions define the conceptual causal target of our framework. We next translate them into observable behavioral proxies (Section 3.2.2), and then instantiate those proxies with concrete disruption operators (Section 3.3).

3.2.2 Behavioral Proxies and Single-Axis Ambiguity

As discussed in Section 2.2, we adopt a disruption-based causal testing perspective. For each causal facet, we design a targeted disruption and use the resulting behavioral response as a proxy for the corresponding behavior, without interpreting these responses as formally identified causal effects.

The necessity-side behavioral proxy measures how much the answer changes when M is disrupted:

$$\Delta_A(\mathcal{D}^A) := d\big(P(Y|X, \mathcal{D}^A(M)),\ P(Y|X, M)\big), \quad (2)$$

where $\mathcal{D}^A$ is a disruption operator, $\mathcal{D}^A(M)$ is a disrupted version of M, and $d(\cdot, \cdot)$ is a distributional distance. A small $\Delta_A(\mathcal{D}^A)$ indicates answer preservation despite changes to M, providing evidence of necessity failure.

The sufficiency-side behavioral proxy measures whether the baseline answer can be recovered from M after suppressing the direct access to X:

$$\Delta_B(\mathcal{D}_p^B) := P(Y = y^* \mid \mathcal{D}_p^B(X, M)), \quad (3)$$

where y^* is the baseline answer, $\mathcal{D}_p^B$ denotes a disruption that suppresses access to X and to the suffix of M after position p, leaving only the M-prefix ($M[0{:}p]$) visible. A high $\Delta_B(\mathcal{D}_p^B)$ at a small p indicates that the baseline answer can be recovered from only an early portion of the reasoning path after direct access to X is suppressed, yielding a shortcut-consistent signal.

Neither axis is diagnostic of shortcut reasoning on its own. Answer preservation under M-disruption can arise not only from shortcut reasoning, but also from a redundant yet truthful rationale whose content is not necessary for the answer.

Conversely, early recovery under the sufficiency-side masking disruption can arise from shortcut reasoning or from a compact truthful rationale. Importantly, these alternatives are complementary: redundant truthful reasoning can mimic the shortcut signature on the necessity side, whereas compact truthful reasoning can mimic it on the sufficiency side. A stronger shortcut signature therefore requires both shortcut-consistent behaviors to occur jointly—robustness to M-disruption and early recovery from M when X is masked. We formalize this complementary ambiguity and delineate the scope of the conjunction in Appendix A.2. This conjunctive structure motivates a joint aggregation operator that assigns a high score only when both axes provide strong shortcut-consistent evidence, while penalizing cases in which only one axis is strong.

3.3 Operationalization

We instantiate the two axes with disruption operators: inverse decoding and attention masking.

Necessity-side diagnostic via inverse decoding. For a fixed input instance x, we replace its generated rationale m with a disrupted version m' and re-decode the answer. We instantiate the disruption operator $\mathcal{D}^A$ with inverse decoding, defining $m' = \mathcal{D}^A_{p,r}(m)$. For a rationale length L, let $k(p, L) = \lfloor pL/100 \rfloor$ denote the token boundary corresponding to percentage position p. For each position i,

$$m'_i = \begin{cases} m_i, & i \le k(p, L), \\ \mathrm{Rank}_r\big(P_\theta(\cdot \mid x, m_{<i})\big), & i > k(p, L), \end{cases} \tag{4}$$

where Rank_r returns the token at rank r. The position grid $p \in \mathcal{P}_A$ specifies the position at which disruption begins. We define the necessity-side signal as

$$a(p, r; x) := \mathbb{1}\big[y_A(p, r; x) = y_0(x)\big], \tag{5}$$

where $y_0(x)$ is the baseline answer and $y_A(p, r; x)$ is the answer after disruption. If $a(p, r; x) = 1$, the answer is preserved under disruption, corresponding to a shortcut-consistent signal of necessity failure. Because inverse decoding selects replacement tokens from the model's own next-token distribution, it is adaptive to both the instance and the model, and contrastive by construction. Detailed analyses and ablations are provided in Appendix B.1.

Sufficiency-side diagnostic via attention masking. For a fixed input instance x and its generated rationale m, we instantiate $\mathcal{D}^B_p$ by applying attention masking to the entire input x and to the rationale suffix after position p, leaving only the prefix $m[0{:}p]$ visible for answer decoding. We define the sufficiency-side signal as

$$b(p; x) := \mathbb{1}\big[y_B(p; x) = y_0(x)\big]. \tag{6}$$

If $b(p; x) = 1$, the baseline answer is recovered from $m[0{:}p]$ while the input and the remaining suffix are masked. In particular, recovery at a small p indicates that the remainder of the reasoning path is not required for recovering the baseline answer under this condition, yielding a shortcut-consistent signal. While sufficiency concerns whether M can sustain the answer after suppressing the direct contribution of X, shortcut reasoning is particularly indicated when this sufficiency emerges from only a small portion of M. We therefore operationalize the sufficiency-side proxy as a prefix-recovery profile over p.

The X-mask component of $\mathcal{D}^B_p$ is a design choice intended to isolate recovery from m by suppressing access to the original input. If the input x remains visible, recovery may result either from m or from the direct contribution of the input itself. We therefore suppress the direct $X \to Y$ contribution through attention masking during the recovery test and measure how early the baseline answer can be recovered from an M-prefix alone.

3.4 CORD: Causal Redundancy of Reasoning via Disruption

The two grid-level indicators, $a(p, r; x)$ and $b(p; x)$, exhibit qualitatively different curve shapes on real LLM rationales. The a-matrix changes gradually across positions and ranks, whereas the b-curve is approximately step-shaped. We therefore combine the two axes using functional forms that preserve their respective information structures: integration over the full grid for the A-axis and a saturation position for the B-axis.

A-component. Let $\mathcal{P}_A \subset [0, 100]$ denote the position grid and $\mathcal{R}$ the rank grid. We define

$$\bar{a}(x) := \frac{1}{|\mathcal{P}_A|\,|\mathcal{R}|} \sum_{p \in \mathcal{P}_A} \sum_{r \in \mathcal{R}} a(p, r; x) \in [0, 1]. \tag{7}$$

A larger value indicates that the answer is preserved more often under M-disruption.

B-component. Let $\mathcal{P}_B \subset [0, 100]$ denote the visible-prefix percentage grid. We define the saturation position as

$$p_B^\star(x) := \min\{p \in \mathcal{P}_B \ : \ b(p; x) = 1\}. \quad (8)$$

If the threshold is never reached, we set $p_B^\star := \max(\mathcal{P}_B)$. We further define the normalized recovery score as $q_B^\star(x) := 1 - p_B^\star(x)/100 \in [0, 1]$, and use it as the B-component of CORD. This value is larger when the baseline answer is recovered from a shorter M-prefix.

Combination. We combine the two components using their harmonic mean:

$$\mathrm{CORD}(x) := H(\bar{a}(x),\, q_B^\star(x))\,, \quad (9)$$

where $H(x, y) = 2xy/(x + y)$ when $x + y > 0$, and $H(0, 0) = 0$. A value of CORD $\to 1$ indicates a strong shortcut signature, while CORD $\to 0$ indicates a weak signature.

Because the harmonic mean is low whenever either argument is small, CORD implements the dual-disruption conjunction in continuous form. The B-component depends on the saturation position $p_B^\star$: once the prefix recovers the baseline answer, the recovery position is fixed and any later region of M does not enter the metric. Appendix B details CORD's design, while Appendix C tests robustness to length, early answer appearance, and rationale transformations.

4 Experiments

4.1 Setup

Models. We evaluate six models spanning three training paradigms. **R1-distilled reasoning**: DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), fine-tuned on reasoning traces generated by the RL-trained DeepSeek-R1 from Qwen2.5-Math-7B-Base and Llama-3.1-8B-Base, respectively. **Math-domain SFT**: Qwen2.5-Math-7B-Instruct (Yang et al., 2024b) and DeepSeek-Math-7B-Instruct (Shao et al., 2024). **Plain instruction-tuned**: Qwen2.5-7B-Instruct (Yang et al., 2024a) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). R1D-Qwen and R1D-Llama denote R1-Distill-Qwen-7B and R1-Distill-Llama-8B, respectively.

Datasets. We use two benchmarks that differ systematically in reasoning depth. **GSM8K** (full test split, $n = 1{,}319$) (Cobbe et al., 2021) consists of elementary-level word problems requiring only a few arithmetic steps, making it a natural setting in which shortcut reasoning is plausible. **MATH-500** ($n = 500$) (Lightman et al., 2024) is a stratified subset of the MATH benchmark spanning seven subject areas across five difficulty levels, where principled multi-step inference is required. Answers are extracted via sympy-based symbolic equivalence to avoid surface-form mismatches.

Dataset	Model	Base acc.	CORD
GSM8K	R1D-Qwen	0.907	**0.686**
	R1D-Llama	0.883	0.669
	Qwen-Math	**0.955**	0.411
	Qwen-7B	0.912	0.374
	Llama-3.1	0.868	0.308
	DeepSeek-Math	0.815	0.304
MATH-500	R1D-Llama	**0.746**	**0.590**
	R1D-Qwen	0.738	0.569
	Qwen-Math	0.686	0.422
	Qwen-7B	0.672	0.402
	Llama-3.1	0.518	0.273
	DeepSeek-Math	0.392	0.262

Table 1: Mean CORD and baseline accuracy.

Metrics. We report two metrics: **base accuracy**, the fraction of instances where the model's final answer matches the ground truth, and **CORD** (Section 3.4), our primary measure of shortcut tendency. Higher CORD values indicate stronger shortcut behavior. All cross-model comparisons are paired by instance to control for problem-level variance.

Measurement grid. We use a grid $\mathcal{P} \times \mathcal{R}$ with intervention positions $\mathcal{P} = \{0, 5, 10, \ldots, 100\}$ (21 points, as % of M) and disruption ranks $\mathcal{R} \in \{\text{bottom-1}, \text{bottom-5}, \text{bottom-10}, \text{top-5}, \text{top-10}\}$ denoting likelihood-rank choices for M-disruption, where bottom-k/top-k selects the corresponding low-/high-likelihood token under the model's next-token distribution. More details on the experimental setup are in Appendix B.7.

4.2 Shortcut Profiles Across Paradigms and Tasks

Finding 1: CORD captures a paradigm-level signal independent of accuracy. The six models separate cleanly by training paradigm: R1-Distill models average CORD ≈ 0.63, nearly twice that of the math-tuned (≈ 0.35) and plain instruction-tuned (≈ 0.34) groups. This separation is not driven by accuracy. On GSM8K, Qwen-Math achieves the highest accuracy (0.955) yet only an intermediate CORD (0.411), while R1D-

Task	Contrast	Overall Δ	Pair-mean CoT-length quartile			
			Q1	Q2	Q3	Q4
GSM8K	R1D-Llama vs. Llama-3.1	+0.361	+0.296†	+0.323†	+0.379†	+0.449†
GSM8K	R1D-Qwen vs. Qwen-Math	+0.275	+0.271†	+0.288†	+0.310†	+0.230†
MATH-500	R1D-Llama vs. Llama-3.1	+0.317	+0.298†	+0.349†	+0.351†	+0.269†
MATH-500	R1D-Qwen vs. Qwen-Math	+0.148	+0.127†	+0.216†	+0.226†	+0.021

Table 2: CORD differences (Δ = R1-Distill − non-distilled) after stratifying instances by pair-mean baseline CoT length. All 16 quartile cells preserve the positive paradigm-level gap; 15 of 16 satisfy $p < 10^{-6}$ under paired two-sided Wilcoxon tests ($^{\dagger}p < 10^{-6}$). The only non-significant cell is R1D-Qwen vs. Qwen-Math on MATH-500 Q4 ($\Delta = +0.021$, $p = 0.32$).

Task	Contrast	Δ	p
GSM8K	R1D-Llama vs. Llama-3.1	$+0.361$	$10^{-211\ddagger}$
	R1D-Qwen vs. Qwen-Math	$+0.275$	$10^{-179\ddagger}$
MATH-500	R1D-Llama vs. Llama-3.1	$+0.317$	$10^{-51\ddagger}$
	R1D-Qwen vs. Qwen-Math	$+0.148$	$10^{-22\ddagger}$

Table 3: Headline paired Wilcoxon comparisons for the base-family-aligned R1-Distill and non-distilled model pairs. $\Delta = \mathrm{mean}(m_1) - \mathrm{mean}(m_2)$ on instance-level CORD. Bonferroni correction is applied over the 15 pairwise comparisons within each dataset; ‡ denotes $p < 0.05/15 \approx 0.00333$. Full results are in Appendix C.3.

Qwen achieves a lower accuracy (0.907) but the highest CORD (0.686). More broadly, all six models operate in a high-accuracy regime on GSM8K ($\geq$ 0.81), yet their CORD values span 0.30 to 0.69. CORD thus reflects how a model reasons, not merely whether it succeeds. We interpret this signal as functional redundancy of the reasoning path along a spectrum from pure shortcut to early answer formation with extended elaboration. The paradigm gap is robust to baseline CoT length: after stratifying base-family-aligned model pairs by pair-mean CoT length, R1-Distill retains higher CORD in all 16 quartile cells, with 15 of 16 significant at $p < 10^{-6}$ (Table 2). The sole attenuation occurs in the longest R1D-Qwen versus Qwen-Math MATH-500 quartile, whereas the R1D-Llama gap on GSM8K increases monotonically with length. Full stratification details are provided in Appendix C.1. Appendix C.2 further stress-tests a distinctive R1-Distill behavior: these models tend to state their answers early and continue the CoT afterward. Even after deliberately assigning the minimum B-side score whenever recovery coincides with the first availability of the complete answer, all four base-family-aligned contrasts remain positive and statistically significant ($\Delta = +0.043$ to $+0.176$). Together, these results show that the paradigm-level separation cannot be explained by baseline CoT length or by R1-Distill's tendency to state answers early.

Finding 2: R1-Distill separation is consistent across models and tasks. We compare model pairs using paired Wilcoxon signed-rank tests on instance-level CORD differences, where $\Delta = \mathrm{mean}(m_1) - \mathrm{mean}(m_2)$ for the first model m_1 and the second model m_2 (Table 3). The R1-Distill separation is highly consistent. Across the exhaustive pairwise analysis, all 16 R1-Distill versus non-distilled contrasts are positive and remain significant after within-dataset Bonferroni correction ($\alpha = 0.05/15$). On GSM8K, these differences range from $\Delta = +0.258$ to $+0.381$, and on MATH-500 from $\Delta = +0.148$ to $+0.327$.

The base-family-aligned comparisons show the same pattern: R1D-Llama exceeds Llama-3.1 by $+0.361$ on GSM8K and $+0.317$ on MATH-500, while R1D-Qwen exceeds Qwen-Math by $+0.275$ and $+0.148$, respectively. In contrast, the two R1-Distill models themselves are much closer ($+0.016$ on GSM8K and -0.020 on MATH-500), with no detectable difference on MATH-500 ($p = 0.92$). Thus, the salient regularity is a consistent R1-Distill signature: across model identity, backbone, and task, R1-Distill models systematically occupy the higher-CORD regime. Full pairwise results are provided in Appendix C.3.

4.3 Framework Validation via Self-Refine

Self-refine provides a CORD-independent validation condition. We validate the framework using self-refine (Madaan et al., 2023), which we first characterize without using CORD (Appendix D.1). Two tests show that self-refine in this setting behaves consistently with X-only stochastic resampling. First, changing or removing the rationale M, preceding answer y_0, critique C, or refinement in-

Subset	Base	Refine	Δ	ρ
Full (n=464)	0.432	0.391	-0.0410	$+0.814$
Unchanged (n=319)	0.492	0.494	$+\mathbf{0.0019}$	—

Table 4: CORD on baseline versus refined CoT (Qwen2.5-Math-7B $\times$ MATH-500). The full set includes instances where refinement changes the answer. The answer-unchanged subset is the direct validation condition, where CORD shifts by only $+0.0019$. Component-level results are in Appendix D.2.

struction changes the overall answer-change rate by at most 1.04 percentage points. Second, refinement almost never changes the answer on instances that are stable under independent X-only resampling (0.0%–5.4%), and 97.3%–100.0% of refinement-induced changes occur in the X-only-unstable subset. These results do not imply that the refinement inputs are never used on individual instances; rather, they show that the observed refinement behavior is dominated by instability already present under X-only inference.

This characterization gives a direct prediction for CORD. When self-refine leaves the final answer unchanged, the baseline and refined rationales may differ, but both lead to the same X-driven answer. A measure of the rationale's functional role should therefore change very little between them.

CORD **recovers the predicted near-zero shift.** We compute CORD on the refined CoT and compare it with the original baseline measurement (Table 4). Across the full set ($n = 464$), baseline and refined CORD values remain strongly aligned at the instance level (Spearman $\rho = +0.814$), although their means differ ($\Delta = -0.0410$); this set also includes cases where refinement changes the answer. The direct validation condition is therefore the answer-unchanged subset ($n = 319$). Here, CORD changes by only $+0.0019$, and the two CORD components likewise show near-zero shifts ($|\Delta| < 0.01$; Appendix D.2). Thus, CORD recovers the prediction derived independently from the X-only characterization of self-refine.

Beyond this validation, we examine whether the magnitude of CORD provides a graded signal for refinement selection (Appendix D.3). A threshold sweep shows that higher-CORD subsets are more enriched for cases where refinement leaves the answer unchanged, with the clearest pattern on MATH-500. We treat this analysis as exploratory rather than as part of the framework validation.

5 Analysis

We next examine two questions about the measurement itself. First, do the necessity- and sufficiency-side components provide complementary empirical information, as required by the dual-disruption construction? Second, how does CORD relate to Early Answering (EA), a closely related prefix-recovery diagnostic? The full component and method-level analyses are provided in Appendix B.4–B.6.

5.1 The Two Axes Are Related but Non-Redundant

Section 3 motivates the two disruption axes through their complementary single-axis ambiguities. This conceptual argument does not, however, determine how strongly the two signals should be associated on real model outputs. If they were nearly interchangeable, the second axis would add little empirical information; if they were unrelated, combining them as measurements of a common shortcut-consistent regime would be difficult to motivate.

We therefore analyze the two components that enter CORD directly: the necessity-side full-grid answer-preservation score $\bar{a}$ and the sufficiency-side early-recovery score $q_B^\star$. Across all six models and both datasets, the two components are consistently positively associated. Instance-level Spearman correlations range from 0.532 to 0.798 across the 12 model–task cells, with a mean of 0.659; Pearson correlations show the same pattern ($r = 0.438$–0.791, mean 0.624) (Appendix B.4.1). Thus, instances whose answers are more robust to rationale disruption also tend to permit earlier recovery from the rationale prefix under the X-masked test.

The correspondence is nevertheless incomplete. Normalized mutual information (NMI) ranges from 0.110 to 0.311 with a mean of 0.197, showing that neither component determines the other at the instance level. The strength of this coupling also varies across model families (Table 5): R1-Distill models show the strongest association (mean $\rho = 0.765$, NMI $= 0.280$), whereas Math-tuned and Plain models exhibit more moderate coupling ($\rho = 0.607/0.604$, NMI $= 0.156/0.155$). The relevant distinction is therefore not whether the axes are related—all 12 cells are positively associated—but how tightly their responses co-vary.

A cross-task robustness check gives a complementary result. Ranking the six models independently on GSM8K and MATH-500 yields Kendall $\tau = +0.600$ for the position-based A-only

Model family	Mean Spearman ρ	Mean NMI
R1-Distill	**0.765**	**0.280**
Math-tuned	0.607	0.156
Plain	0.604	0.155

Table 5: Relationship between the two CORD components, averaged within model families over GSM8K and MATH-500. Full model–task statistics are in Appendix B.4.

summary, $+0.733$ for the B-only summary, and $+\mathbf{0.867}$ for CORD (Appendix B.5.2). Thus, the final construction not only combines two related but non-interchangeable signals, but also gives the strongest preservation of the model ordering across tasks.

5.2 Relation to Early Answering: Convergent but Operationally Distinct

Early Answering (EA) (Lanham et al., 2023) provides a natural external comparison because it also measures how early the baseline answer can be recovered from a rationale prefix. Across all 12 model–task cells, CORD and EA are positively associated (Table 6): Spearman ρ ranges from 0.592 to 0.859, with a mean of 0.755. Thus, instances assigned higher CORD also tend to be instances on which EA recovers the answer earlier.

The two tests nevertheless differ in a central information condition. EA keeps the original input X visible while varying the amount of CoT, whereas CORD's sufficiency-side test masks X and asks whether the visible rationale prefix itself can sustain the baseline answer. At $p = 0$, where no rationale token is visible, EA still reproduces the baseline answer on 8.8–30.1% of GSM8K instances and 15.1–32.4% of MATH-500 instances across models. These recoveries can therefore arise through the retained input X rather than through the rationale prefix.

EA nevertheless captures the same broad early-recovery tendency. At $p = 50$ on GSM8K, the two R1-Distill models recover their baseline answers on 0.911 and 0.848 of instances, compared with 0.140–0.389 for the four non-distilled models; on MATH-500, the corresponding rates are 0.714 and 0.680 versus 0.233–0.412. Together with the substantial CORD–EA correlations, these results show cross-method convergence while clarifying why the two diagnostics are not equivalent: EA measures early recovery with X available, whereas CORD removes direct access to X to isolate recovery supported by the rationale prefix itself. Full recovery profiles are reported in Appendix B.6.2.

Model	GSM8K	MATH-500
R1D-Qwen	0.724	0.760
R1D-Llama	0.859	0.856
Qwen-Math	0.737	0.830
Qwen-7B	0.753	0.825
DeepSeek-Math	0.694	0.717
Llama-3.1	0.712	0.592
Mean	**0.755**	

Table 6: Instance-level Spearman ρ between CORD and the Early Answering score. The association is positive in every model–task cell.

6 Conclusion

We formalized shortcut reasoning in CoT as causal redundancy of the expressed rationale for the final answer and argued that neither necessity- nor sufficiency-side evidence is sufficient on its own. We instantiated their conjunction with inverse decoding and X-masked prefix recovery, combining the resulting signals into CORD. Across 6 models and 2 datasets, R1-Distill models occupy a substantially higher-CORD regime than non-distilled models, with all 16 pairwise contrasts positive and significant after within-dataset correction. The direction is preserved across all 16 CoT-length strata, with 15 of 16 significant, and all four base-family-aligned contrasts remain significant after conservatively removing the B-side advantage associated with early answer appearance.

The measurement analyses further support the dual-disruption design. The two CORD components are positively associated in every model–task cell while remaining non-interchangeable (mean Spearman $\rho = 0.659$; mean NMI $= 0.197$), and CORD gives the strongest cross-task ranking preservation (Kendall $\tau = 0.867$). CORD also converges with Early Answering (mean $\rho = 0.755$), while its X-masked sufficiency test distinguishes rationale-supported recovery from recovery through the input. Finally, in the answer-unchanged self-refine condition, CORD exhibits the predicted near-zero shift ($\Delta = +0.0019$). Together, these results position CORD as an instance-level behavioral measure of joint shortcut-consistent signatures and show that correct or coherent CoT alone does not establish that the expressed reasoning path functionally determines the answer.

Limitations

Methodological scope. Our measurements are subject to three practical constraints. (i) Answer extraction is highly reliable on GSM8K but more challenging on MATH-500, where greater variation in mathematical expressions and output formats may introduce residual extraction error. (ii) The self-refine validation (Section 4.3) covers only the one-round critique-and-refine setting of Madaan et al. (2023); whether the same characterization extends to multi-round refinement or other iterative inference procedures remains open (deployment cost and cold-start scope are discussed in Appendix D.4). (iii) The disruption-operator comparison is limited to the nine operators examined in Appendix B.1, including inverse-decoding ranks, random and KL-divergent replacement, filler, and Early Answering, and does not include broader LLM-based rewrite operators such as Adding Mistakes. Paraphrasing is examined separately as a robustness transformation in Appendix C.4. Broader operator families would provide a stronger test of operator-level generalization.

Epistemic scope. We do not claim strict causal identification. (i) The two disruption axes and CORD are behavioral proxies rather than direct estimates of Pearl-style counterfactual probabilities or causal effects; under autoregressive entanglement, there is no clean observable counterpart of removing only the causal contribution of M (Appendix A.3). (ii) CORD measures whether the expressed rationale is behaviorally redundant for the answer, but does not identify the source of that redundancy. In particular, it does not distinguish redundancy supported by an X-mediated route from input-independent tendencies encoded in the model. (iii) The conjunction addresses the two single-axis ambiguities formalized in Appendix A.2—redundant truthful reasoning and compact truthful reasoning—but does not exclude every possible mechanism that may produce the same behavioral signature; additional alternatives may require additional disruption axes. (iv) CORD therefore complements rather than replaces mediation-based or mechanistic accounts of model reasoning (Paul et al., 2024; Fu et al., 2025b).

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. IITP-RS-2019-II190421, IITP-RS-2022-II220680, IITP-RS-2024-00437633, NRF-RS-2025-00564083).

References

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

Iván 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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jiarun Fu, Lizhong Ding, Hao Li, Pengqi Li, Qiuning Wei, and Xu Chen. 2025a. Unveiling and causalizing cot: A causal pespective. *Preprint*, arXiv:2502.18239.

Zhizhang Fu, Guangsheng Bao, Hongbo Zhang, Chenkai Hu, and Yue Zhang. 2025b. Correlation or causation: Analyzing the causal structures of LLM and LRM reasoning process. *arXiv preprint arXiv:2509.17380*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bradford Hall, Advait Limaye, and Ashok B. Kulkarni. 2009. Overview: Generation of gene knockout mice.

Current Protocols in Cell Biology, 44(1):19.12.1–19.12.17.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Judea Pearl. 1999. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1/2):93–149.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Oleg Somov, Mikhail Chaichuk, Mikhail Seleznyov, Alexander Panchenko, and Elena Tutubalina. 2026. Breaking the chain: A causal analysis of LLM faithfulness to intermediate structures. *arXiv preprint arXiv:2603.16475*.

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

Avinash R. Vaidya, Maia S. Pujara, Michael Petrides, Elisabeth A. Murray, and Lesley K. Fellows. 2019. Lesion studies in contemporary neuroscience. *Trends in Cognitive Sciences*, 23(8):653–671.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. Faithful and unfaithful error recovery in chain of thought. In *First Conference on Language Modeling*.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25842–25850.

A Framework foundation

This appendix consolidates the foundations of the framework in §3, including notation, formal results, and epistemic positioning.

A.1 Notation

We summarize the notation used throughout the framework below. Uppercase letters denote causal variables, lowercase letters their realizations, and calligraphic letters their respective domains or disruption operators. Position p is expressed as a percentage of the rationale length; $m[0{:}p]$ denotes the corresponding prefix up to that position.

Causal variables, domains, and realizations.

Symbol	Definition
$X \in \mathcal{X}$	Input question.
$M \in \mathcal{M}$	Generated chain-of-thought rationale.
$Y \in \mathcal{Y}$	Final answer.
x, m, y	Realizations of X, M, Y, respectively.
$m = (m_1, \ldots, m_L)$	Realized baseline rationale of length L tokens.
$m[0{:}p]$	Prefix of m visible up to position p under the sufficiency-side masking disruption.

Model and baseline generation.

Symbol	Definition
θ	Parameters of the LLM.
$P_\theta(\cdot \mid \cdot)$	Conditional generation distribution of the model.
$y_0(x)$	Baseline answer obtained from the model's original generation for input x. All instance-level measurements are anchored to agreement with $y_0(x)$.
$y_A(p, r; x)$	Answer re-decoded after applying the necessity-side inverse-decoding disruption starting at position p with rank choice r.
$y_B(p; x)$	Answer re-decoded under the sufficiency-side masking disruption at position p.
y^*	Baseline answer used as the target outcome in the sufficiency-side behavioral proxy.

Disruption operators.

Symbol	Definition
$\mathcal{D}^A$	Necessity-side disruption operator applied to M.
$\mathcal{D}^A_{p,r}$	Inverse-decoding instantiation of $\mathcal{D}^A$, parameterized by disruption position p and token-rank choice r.
$m' = \mathcal{D}^A_{p,r}(m)$	Disrupted rationale produced by inverse decoding. Tokens before p remain unchanged, while tokens from p onward are replaced according to Equation 4.
$\mathcal{D}^B_p$	Sufficiency-side masking disruption applied jointly to (X, M). It masks access to the entire input X and to the suffix of M beyond position p, leaving only the M-prefix $M[0{:}p]$ visible for answer decoding.

Behavioral proxies.

Symbol	Definition
$\Delta_A(\mathcal{D}^A) = d\Big(P(Y \mid X, \mathcal{D}^A(M)), P(Y \mid X, M)\Big)$	Necessity-side behavioral proxy (Equation 2). A small value indicates answer preservation despite disruption of M, providing shortcut-consistent evidence of necessity failure.
$\Delta_B(\mathcal{D}^B_p) = P\Big(Y = y^* \mid \mathcal{D}^B_p(X, M)\Big)$	Sufficiency-side behavioral proxy (Equation 3). A large value at a small p indicates that the baseline answer is recoverable when direct access to X and most of the later reasoning path are masked.

Disruption grids.

Symbol	Definition
$\mathcal{P}_A$	Position grid specifying where inverse-decoding disruption begins along M.
$\mathcal{R}$	Replacement-token rank grid used by inverse decoding.
$\mathcal{P}_B$	Visible-prefix percentage grid used by the sufficiency-side masking disruption.
$\mathcal{P}_A \times \mathcal{R}$	Necessity-side disruption measurement space.
$\mathrm{Rank}_r(\cdot)$	Operator that returns the token at rank r from a next-token distribution.

In our experiments, $\mathcal{P}_A = \mathcal{P}_B = \{0, 5, 10, \ldots, 100\}$, and $\mathcal{R} = \{\text{bot-1}, \text{bot-5}, \text{bot-10}, \text{top-5}, \text{top-10}\}$. CORD averages over all five ranks in $\mathcal{R}$, with bot-1 denoting the lowest-likelihood rank.

Instance-level behavioral signals.

Symbol	Definition
$a(p, r; x) := \mathbb{1}[y_A(p, r; x) = y_0(x)]$	Answer-preservation indicator under the necessity-side disruption. A value of 1 indicates that the baseline answer is preserved after inverse decoding.
$b(p; x) := \mathbb{1}[y_B(p; x) = y_0(x)]$	Answer-recovery indicator under the sufficiency-side masking disruption. A value of 1 indicates that the baseline answer is recovered with only the M-prefix up to p visible.

CORD components and instance-level aggregation.

Symbol	Definition
$\bar{a}(x) = \frac{1}{\lvert\mathcal{P}_A\rvert\lvert\mathcal{R}\rvert} \sum_{p \in \mathcal{P}_A} \sum_{r \in \mathcal{R}} a(p, r; x)$	Full-grid aggregation of the necessity-side answer-preservation signal (Equation 7).
$p_B^\star(x) = \min\{p \in \mathcal{P}_B : b(p; x) = 1\}$	First position at which the baseline answer is recovered under the sufficiency-side masking disruption (Equation 8). If recovery never occurs, $p_B^\star(x)$ is set to $\max(\mathcal{P}_B)$.
$q_B^\star(x) = 1 - \frac{p_B^\star(x)}{100}$	Normalized early-recovery score. Larger values correspond to recovery from a shorter M-prefix.
$H(u, v) = \frac{2uv}{u+v}$	Harmonic-mean aggregation operator for $u + v > 0$, with $H(0, 0) = 0$.
$\mathrm{CORD}(x) = H(\bar{a}(x), q_B^\star(x))$	Instance-level CORD score (Equation 9). Larger values indicate a stronger joint shortcut-consistent signature across the two disruption axes.

Class	Mechanism	$A(x)$	$B(x)$
C1	Shortcut	1	1
C2a	Redundant truthful	1	0
C2b	Compact truthful	0	1
C0	Genuine reasoning	0	0
C3	Brittle extraction	0	0

Table 7: Five representative mechanisms mapped to the four behavioral regions induced by the two disruption axes. C2a shares the necessity-side signature of C1, whereas C2b shares its sufficiency-side signature.

A.2 Behavioral Partition and Complementarity

This section formalizes why the two disruption axes are complementary. The argument concerns the *behavioral signatures* induced by the two diagnostics rather than formal causal identification.

For exposition, let

$$A(x) := \mathbb{1}[\bar{a}(x) \geq \tau_A]\,,$$

$$B(x) := \mathbb{1}[q^{\star}_B(x) \geq \tau_B]\,,$$

where $A(x) = 1$ denotes robustness to necessity-side M-disruption and $B(x) = 1$ denotes early recovery under the sufficiency-side masking disruption. The thresholds τ_A and τ_B are used only to state the behavioral partition; CORD itself remains continuous and does not require these thresholds.

A.2.1 Two-axis behavioral partition

The two binary attributes induce four behavioral regions. Table 7 maps five representative mechanisms discussed in our framework to these regions. Two distinct mechanisms, genuine reasoning (C0) and brittle extraction (C3), occupy the same low–low region but for different reasons.

The intended interpretations are as follows. C1 represents the target shortcut signature: the answer is robust to disruption of M, yet is recoverable from only an early M-prefix when direct access to X and the remaining rationale suffix are masked. C2a represents a redundant but truthful rationale whose content is robust to local disruption, while recovery still requires a substantial portion of the reasoning path. C2b represents a compact truthful rationale whose answer is recoverable early, but whose reasoning content is not robust to disruption. C0 and C3 exhibit neither shortcut-consistent attribute.

This partition is not intended to enumerate every possible mechanism that can produce a given behavioral signature. Its role is to formalize the complementary ambiguities motivating the two-axis design.

A.2.2 Single-axis ambiguity

Observation 1 (Single-axis ambiguity)**.** Within the behavioral partition of Table 7, neither axis alone distinguishes shortcut reasoning from both truthful alternatives. C1 and C2a share the same necessity-side signature, whereas C1 and C2b share the same sufficiency-side signature.

Proof. Consider the necessity side first. Both C1 and C2a satisfy $A(x) = 1$: the baseline answer remains robust to M-disruption. Their underlying interpretations differ—C1 is shortcut-consistent, whereas C2a represents redundant truthful reasoning—but the necessity-side signature alone does not distinguish them.

Likewise, both C1 and C2b satisfy $B(x) = 1$: the baseline answer is recoverable from an early M-prefix under the sufficiency-side masking disruption. In C1 this is shortcut-consistent, whereas in C2b it arises because a compact truthful rationale reaches the answer-determining step early. Thus the sufficiency-side signature alone does not distinguish these mechanisms.

Therefore, each single axis admits a distinct truthful alternative with the same shortcut-consistent behavioral signature. □

A.2.3 Complementarity of the two axes

Observation 2 (Complementarity of the conjunction)**.** Within the behavioral partition of Table 7, the two truthful alternatives that mimic the shortcut signature on one axis are excluded by the other axis. Consequently, only C1 occupies the high–high region

$$A(x) = 1 \ \wedge \ B(x) = 1$$

among the mechanisms represented in the partition.

Proof. C2a matches C1 on the necessity side, with $A(x) = 1$, but differs on the sufficiency side: recovery requires a larger portion of the reasoning path, so $B(x) = 0$. Hence C2a does not satisfy the conjunction.

Conversely, C2b matches C1 on the sufficiency side, with $B(x) = 1$, but differs on the necessity side: its answer is sensitive to disruption of the reasoning path, so $A(x) = 0$. Hence C2b also does not satisfy the conjunction.

C0 and C3 lie in the low–low region and therefore satisfy neither condition. Thus, within the stated partition, requiring both shortcut-consistent

signatures simultaneously excludes the single-axis alternatives and selects the C1 region. □

Relation to continuous CORD. The observations above are stated using binary attributes only to make the complementarity explicit. CORD implements the same conjunctive structure continuously:

$$\mathrm{CORD}(x) = H(\bar{a}(x), q_B^\star(x)),$$

where both components increase in the shortcut-consistent direction. Because the harmonic mean penalizes imbalance between its arguments, a strong value requires both substantial robustness to M-disruption and early recovery under the sufficiency-side masking disruption. Accordingly, CORD should be interpreted as a continuous measure of the joint behavioral signature, not as formal identification of an underlying causal class.

A.3 Epistemic Positioning of Disruption-Based Causal Testing

A.3.1 Relation to Prior Disruption-Based Faithfulness Tests

From the perspective of our framework, the tests of Lanham et al. (2023) can be interpreted as single-disruption diagnostics. *Adding Mistakes*, *Paraphrasing*, and *Filler Tokens* modify the content of M while keeping X available, and are therefore closely aligned with our necessity-side diagnostic: they test whether the answer remains robust when the expressed reasoning path is disrupted.

Early Answering instead varies how much of M is available by truncating the rationale while keeping X visible. It is therefore related to the prefix-recovery dimension of our sufficiency-side diagnostic, but differs in an important respect: because X remains available, successful answer recovery may arise either from the visible M-prefix or from direct access to X. Our sufficiency-side disruption addresses this ambiguity by masking access to X together with the suffix of M, leaving only the selected M-prefix visible.

Thus, prior perturbation and early-answering tests provide informative single-axis evidence, whereas CORD combines complementary disruption responses designed to reduce the distinct ambiguities associated with either axis alone (Observation 1).

A.3.2 Epistemic Scope

CORD is a disruption-based behavioral diagnostic of causal redundancy, not a formally identified causal effect. Its interpretation is therefore deliberately narrower than identification in the sense of Pearl's do-calculus.

What the framework supports. First, within the behavioral partition of Section A.2, the two axes address complementary single-axis ambiguities: redundant truthful reasoning can mimic the shortcut signature on the necessity side, whereas compact truthful reasoning can mimic it on the sufficiency side. Their conjunction excludes these alternatives on different axes (Observation 2). Second, CORD provides a continuous instance-level measure of this joint shortcut-consistent signature, enabling ranking and paired comparisons across models and conditions. More generally, the framework follows the logic of disruption-based functional testing: if deliberately disrupting a candidate reasoning route produces a characteristic behavioral response, that response provides evidence about the route's functional role.

What the framework does not establish. The framework does not claim that the conjunction uniquely identifies shortcut reasoning against every conceivable mechanism; additional behavioral alternatives may exist beyond those represented in our partition. Nor is the X-mask assumed to be formally equivalent to an intervention in which X is causally absent. It is an operational design that suppresses direct access to X during answer recovery. Finally, Δ_A, Δ_B, and CORD are behavioral proxies grounded in necessity- and sufficiency-side disruption tests; they should not be interpreted as Pearl-style probabilities of causation, average causal effects, or other formally identified causal estimands.

B Methodology Details

This appendix provides the design evidence underlying the final CORD measurement. We examine why inverse decoding is used for the necessity-side disruption (Appendix B.1), why the sufficiency-side test masks the original input X (Appendix B.2), why the two response profiles are summarized differently (Appendix B.3), how the resulting components relate empirically (Appendix B.4), and why they are combined with a harmonic mean (Appendix B.5). We then compare CORD with Early Answering (EA) (Appendix B.6) and report the measurement environment (Appendix B.7).

B.1 Necessity-Side Disruption: Operator Choice and Robustness

The necessity-side diagnostic asks whether the baseline answer remains stable when the expressed reasoning path M is disrupted while the original input X remains available. A useful diagnostic should capture a consistent answer-preservation signal without relying on a particular form of rationale corruption. We therefore compare inverse decoding with several qualitatively different disruption operators and then examine robustness within the inverse-decoding rank family.

B.1.1 Robustness across disruption operators

All operators use the same intervention-position grid and preserve X throughout the test. Inverse decoding replaces the rationale suffix using tokens selected by likelihood rank under the model's own next-token distribution. Random replacement samples suffix tokens uniformly from the vocabulary, KL-divergent replacement selects tokens that maximally change the predictive distribution, and filler overwrites the suffix with ellipsis tokens. Early Answering (EA) truncates the suffix and asks the model to answer from X and the remaining rationale prefix; its relation to the sufficiency-side prefix-recovery test is discussed separately in Appendix B.6.

Table 8 compares these operators on R1D-Qwen and Qwen-Math over GSM8K. The table reports two complementary views of the comparison. The left panel summarizes the raw necessity-side response: the mean answer-preservation rate a for each model, together with their ratio, $\mathrm{mean}(a_{\mathrm{R1D}})/\mathrm{mean}(a_{\mathrm{Qwen}})$. A ratio above one therefore indicates that R1D-Qwen preserves its baseline answer more often under the corresponding rationale disruption.

The right panel asks whether this operator choice changes the resulting model-level CORD separation. For each instance, we compute CORD using the operator-specific A-axis measurement while keeping the B-axis measurement fixed. We report $\Delta = \mathrm{mean}_i[\mathrm{CORD}_{\mathrm{R1D}}(x_i) - \mathrm{CORD}_{\mathrm{Qwen}}(x_i)]$, Cohen's d for these paired instance-level differences, and the significance of a two-sided paired Wilcoxon signed-rank test as $-\log_{10} p$. Thus, larger Δ and d indicate stronger R1D-Qwen–Qwen-Math separation, while larger $-\log_{10} p$ indicates stronger statistical evidence against a zero paired difference.

The raw necessity-side signal is consistent across all nine operators: R1D-Qwen has a higher mean answer-preservation rate than Qwen-Math, with mean-a ratios ranging only from 1.34 to 1.42. The corresponding paired cross-model differences are also large under every operator. On the baseline-correct paired subset ($n = 1{,}172$), the absolute mean CORD difference ranges from 0.283 to 0.316, with Cohen's d between 1.52 and 1.86. Thus, the necessity-side separation is preserved across substantially different ways of disrupting the rationale.

B.1.2 Why inverse decoding?

We instantiate the necessity-side operator with inverse decoding because it defines rationale disruption directly from the evaluated model's own predictive distribution. At each disrupted position i,

$$m_i' = \mathrm{Rank}_r\left(P_\theta(\cdot \mid x, m_{<i})\right),$$

where Rank_r returns the token at rank r. This construction provides three useful properties.

- **Model-relative disruption.** Replacement tokens are defined relative to each model's own next-token distribution rather than by a fixed external corruption rule. The same rank therefore has a consistent model-relative interpretation across models.

- **Instance-conditioned disruption.** The replacement at each position depends on the current input and rationale context. Disruption is therefore adapted locally to the trajectory being evaluated, without requiring an external semantic-distance metric or a second model.

- **Controlled rank family.** The rank parameter provides a common scale for varying the type of replacement within a single operator family. Bottom-k tokens move the continuation toward low-likelihood alternatives, while top-k tokens provide higher-likelihood contrasts under the same construction.

These properties make inverse decoding a convenient common instrument for the necessity-side measurement. The operator comparison above further shows that the observed answer-preservation pattern is shared by other disruption families rather than being specific to the inverse-decoding construction.

	Mean answer preservation a			Paired cross-model separation		
Operator	R1D-Qwen	Qwen-Math	Ratio	$\lvert\Delta\rvert$	Cohen's d	$-\log_{10} p$
Inverse, bottom-1	0.735	0.538	1.37	0.297	1.66	177.69
Inverse, bottom-5	0.733	0.539	1.36	0.296	1.64	176.93
Inverse, bottom-10	0.733	0.537	1.36	0.297	1.63	176.33
Inverse, top-5	0.686	0.491	1.40	0.287	1.54	172.90
Inverse, top-10	0.701	0.524	1.34	0.283	1.52	172.50
Random uniform	0.742	0.535	1.39	0.302	1.67	177.61
KL-divergent	0.743	0.537	1.38	0.302	1.67	177.99
Filler ("...")	0.797	0.560	1.42	0.316	1.86	184.34
Early Answering	0.785	0.564	1.39	0.309	1.84	184.36

Table 8: Necessity-side disruption comparison on GSM8K (R1D-Qwen vs. Qwen-Math). The left panel reports the raw A-axis response on the full set ($n = 1{,}319$): mean a is the position-averaged answer-preservation rate, and Ratio is $\mathrm{mean}(a_{\mathrm{R1D}})/\mathrm{mean}(a_{\mathrm{Qwen}})$. The right panel reports the paired CORD separation on instances for which both models produce a correct baseline answer ($n = 1{,}172$), with the corresponding operator used for the A-axis and the B-axis held fixed. Δ is the mean paired difference $\mathrm{CORD}_{\mathrm{R1D}} - \mathrm{CORD}_{\mathrm{Qwen}}$; Cohen's d is its standardized paired effect size; and $-\log_{10} p$ is based on a two-sided Wilcoxon signed-rank test. All nine operators preserve a large positive cross-model separation.

B.1.3 Robustness across inverse-decoding ranks

CORD averages the necessity-side response over

$$\mathcal{R} = \{\text{bot-1}, \text{bot-5}, \text{bot-10}, \text{top-5}, \text{top-10}\}.$$

We examine each rank separately to verify that the resulting model contrast does not depend on one particular point in this rank set.

Table 9 shows a highly consistent pattern. On the full GSM8K set, the R1D-Qwen–Qwen-Math difference ranges from $+0.266$ to $+0.279$ across the five ranks. On the paired baseline-correct subset, it ranges from $+0.284$ to $+0.298$. Averaging the five ranks gives the main values of $+0.275$ and $+0.292$, respectively.

The instance-level rankings are also stable within the low-likelihood portion of the rank family: Spearman correlations among bottom-1, bottom-5, and bottom-10 range from 0.948 to 0.959. Correlations between bottom-k and top-k variants remain positive but are more moderate (0.685–0.724), indicating that the higher-likelihood ranks provide related but non-identical perturbation responses. Averaging across the five ranks therefore preserves the common necessity-side signal while incorporating variation across disruption severity.

B.2 Sufficiency-Side Disruption: Why Mask X?

The sufficiency-side question is different from the necessity-side question. Here we ask whether the baseline answer can be recovered from the rationale itself after suppressing direct access to the original input. This requires distinguishing two possible sources of answer recovery: the visible rationale prefix $M[0{:}p]$ and the original input X.

The empirical recovery profiles illustrate this ambiguity. With EA, the model can reproduce the baseline answer even at $p = 0$, where none of the rationale is available but X remains visible. By contrast, the X-masked condition has approximately zero recovery at $p = 0$ in our measurements. Thus, answer recovery in an X-keep test cannot by itself establish that the visible rationale prefix sustains the answer: the model may simply regenerate the answer directly from X.

This motivates the X/M-mask used by CORD. For each prefix position p, attention masking suppresses the entire input x and the rationale suffix after p, leaving only $m_{[0:p]}$ available for answer decoding. The corresponding indicator is

$$b(p; x) = \mathbb{1}\left[y_B(p; x) = y_0(x)\right].$$

Recovery therefore measures what the visible rationale prefix can sustain without direct access to the original problem.

B.3 From Disruption Profiles to CORD Components

The two disruptions produce qualitatively different response profiles. The necessity-side matrix $a(p, r)$ changes gradually over intervention positions and ranks, whereas the sufficiency-side curve $b(p)$ is approximately step-shaped: recovery is typically absent before some position and present afterward. CORD therefore summarizes the two axes with

	Full set ($n = 1{,}319$)			Baseline-correct ($n = 1{,}172$)		
Rank	R1D-Qwen	Qwen-Math	Δ	R1D-Qwen	Qwen-Math	Δ
bottom-1	0.694	0.415	$+0.279$	0.714	0.417	$+0.297$
bottom-5	0.693	0.414	$+0.278$	0.713	0.417	$+0.296$
bottom-10	0.693	0.414	$+0.279$	0.714	0.416	$+0.298$
top-5	0.665	0.396	$+0.269$	0.685	0.399	$+0.287$
top-10	0.674	0.408	$+0.266$	0.695	0.411	$+0.284$
5-rank average	**0.686**	**0.411**	$\mathbf{+0.275}$	**0.706**	**0.414**	$\mathbf{+0.292}$

Table 9: Rank-wise robustness of CORD for R1D-Qwen versus Qwen-Math on GSM8K. For each row, CORD is recomputed using the corresponding rank-specific necessity-side response while keeping the sufficiency-side measurement fixed. All five ranks preserve a large positive model difference; the last row uses the five-rank average adopted by CORD. All paired rank-specific comparisons satisfy $p < 1.1 \times 10^{-172}$.

Test	X	Prefix implementation	Visible rationale	What recovery can use
Early Answering	visible	text truncation	$M[0{:}p]$	X and/or $M[0{:}p]$
M-mask	visible	attention masking	$M[0{:}p]$	X and/or $M[0{:}p]$
CORD B-axis	**masked**	attention masking	$M[0{:}p]$	$M[0{:}p]$ **only**

Table 10: Comparison of prefix-recovery tests. Early Answering exposes a rationale prefix by textually truncating the suffix, whereas M-mask and the CORD B-axis suppress the suffix through attention masking. Early Answering and M-mask keep X visible, leaving the direct $X \to Y$ route available; the CORD B-axis additionally masks X, so recovery can rely only on the visible rationale prefix.

different functionals rather than forcing them into an identical form.

A-component: full-grid answer preservation. For the necessity side we retain the complete response surface,

$$\bar{a}(x) = \frac{1}{|\mathcal{P}_A||\mathcal{R}|} \sum_{p \in \mathcal{P}_A} \sum_{r \in \mathcal{R}} a(p, r; x).$$

A larger $\bar{a}$ means that the baseline answer is preserved over a larger fraction of rationale disruptions.

B-component: earliest answer recovery. For the sufficiency side we use the first prefix position at which the baseline answer is recovered,

$$p_B^\star(x) = \min \{p \in \mathcal{P}_B : b(p; x) = 1\},$$

with $p_B^\star = \max(\mathcal{P}_B)$ if recovery never occurs. The normalized score is

$$q_B^\star(x) = 1 - \frac{p_B^\star(x)}{100}.$$

A larger $q_B^\star$ therefore means that the baseline answer can be recovered from a shorter rationale prefix.

B.3.1 Why not truncate the A-axis at $p_B^\star$?

The B-axis is summarized by the first position $p_B^\star$ at which the baseline answer is recovered. A natural question is whether the A-axis should also ignore all measurements after this position.

CORD uses the full-grid A-component

$$\bar{a}(x) = \frac{1}{|\mathcal{P}_A||\mathcal{R}|} \sum_{p \in \mathcal{P}_A} \sum_{r \in \mathcal{R}} a(p, r; x).$$

We compare this with an adaptive alternative that uses exactly the same answer-preservation signal but discards positions after the B-side recovery point:

$$\bar{a}_{\mathrm{adp}}(x) = \frac{1}{|\mathcal{P}_{\le p_B^\star(x)}||\mathcal{R}|} \sum_{p \in \mathcal{P}_{\le p_B^\star(x)}} \sum_{r \in \mathcal{R}} a(p, r; x),$$

where

$$\mathcal{P}_{\le p_B^\star(x)} = \{p \in \mathcal{P}_A : p \le p_B^\star(x)\}.$$

The two quantities therefore differ in only one respect: $\bar{a}$ uses the entire A-side response surface, whereas $\bar{a}_{\mathrm{adp}}$ retains only the region before answer recovery in Experiment B. This gives a direct test of whether A-side responses after $p_B^\star$ are informative or can be treated as post-recovery noise.

Truncating the A-axis at $p_B^\star$ weakens both comparisons. This indicates that A-side responses after B-side answer recovery are not simply uninformative tail measurements. Experiment A and Experiment B ask different questions: recovering the answer from an M-prefix in Experiment B does not

A-component	EA ρ	Refine AUROC
$\bar{a}$ (**adopted**)	$+\mathbf{0.741}$	$\mathbf{0.676}$
$\bar{a}_{\text{adp}}$	$+0.665$	0.657

Table 11: Comparison of the full and adaptively truncated A-side summaries, averaged over the four non-R1-Distill model cells on MATH-500 (Qwen-Math, Qwen-7B, DeepSeek-Math, and Llama-3.1). EA ρ is the instance-level Spearman correlation with the Early Answering score, and Refine AUROC measures prediction of whether self-refinement changes the answer.

imply that later portions of the rationale cease to affect the model's response when they are independently disrupted in Experiment A.

We therefore retain the full-grid mean $\bar{a}$ as the A-component of CORD, while summarizing the step-shaped B-side response by its recovery position $p_B^\star$.

B.4 Empirical Relationship Between the Two Axes

The two disruption experiments target complementary behavioral signatures, but both are intended to characterize the same shortcut-consistent regime. If the two components were empirically interchangeable, they would induce nearly the same instance-level ordering. Conversely, if they captured unrelated behavior, their association would approach zero.

B.4.1 Related but non-redundant at the instance level

Table 12 reports Pearson correlation, Spearman rank correlation, and normalized mutual information (NMI) between $\bar{a}$ and $q_B^\star$ separately for every model–task cell. Pearson r measures linear association, Spearman ρ measures monotonic agreement in instance rankings, and NMI provides a complementary measure of statistical dependence without restricting the relationship to a linear or monotonic form.

The relationship is positive in every one of the twelve cells. Pearson correlations range from 0.438 to 0.791, and Spearman correlations range from 0.532 to 0.798, with means of 0.624 and 0.659, respectively. Thus, instances with stronger answer preservation under A-side disruption also tend to permit earlier recovery under the B-side test. The two signatures are therefore systematically related rather than behaving as unrelated measurements.

At the same time, the correspondence is far from one-to-one. Even the strongest rank correlation is 0.798, and NMI ranges from 0.110 to 0.311 with a mean of 0.197. Hence, knowing an instance's response on one axis does not determine its response on the other. The empirical pattern is therefore consistent with the intended dual-axis construction: the two components share a common behavioral tendency while retaining axis-specific variation.

B.4.2 The degree of coupling varies across model families

The strength of the cross-axis relationship is itself structured across model families. Table 13 averages the cell-level statistics above within each training family.

R1-Distill models exhibit the strongest relationship between the two axes, with mean Spearman $\rho = 0.765$ and mean NMI $= 0.280$. The corresponding associations are more moderate for Math-tuned ($\rho = 0.607$, NMI $= 0.156$) and Plain models ($\rho = 0.604$, NMI $= 0.155$). The distinction is therefore not whether the two axes are related—they are positively associated in every model–task cell—but how tightly their instance-level responses are coupled. This coupling varies across model families.

Taken together, these results motivate treating $\bar{a}$ and $q_B^\star$ as complementary rather than interchangeable components. Both track a common shortcut-consistent behavioral tendency, but neither subsumes the other at the instance level. The next section therefore considers how to combine the two components into a single score while preserving the requirement that both signatures be present.

B.5 Combining the Two Axes

Appendix B.4 shows that the two CORD components are systematically related but not interchangeable at the instance level. A scalar shortcut score should therefore retain information from both axes while reflecting the conjunctive structure of the shortcut definition. For compactness, let

$$u = \bar{a}(x), \qquad v = q_B^\star(x).$$

CORD combines them using the harmonic mean,

$$\mathrm{CORD}(x) = H(u,v) = \begin{cases} \dfrac{2uv}{u+v}, & u+v>0, \\ 0, & u=v=0. \end{cases}$$

B.5.1 Conjunctive aggregation via the harmonic mean

The scalar aggregation should preserve the interpretation of the two components. A larger u indi-

Model	Task	n	Pearson r	Spearman ρ	NMI
R1D-Qwen	GSM8K	1,319	0.786	0.768	0.254
	MATH-500	500	0.788	0.776	0.299
R1D-Llama	GSM8K	1,319	0.791	0.798	0.255
	MATH-500	500	0.726	0.716	0.311
Qwen-Math	GSM8K	1,319	0.438	0.532	0.119
	MATH-500	500	0.662	0.669	0.209
DeepSeek-Math	GSM8K	1,319	0.442	0.574	0.111
	MATH-500	500	0.582	0.654	0.183
Qwen-7B	GSM8K	1,319	0.460	0.559	0.110
	MATH-500	500	0.721	0.720	0.216
Llama-3.1	GSM8K	1,319	0.536	0.589	0.119
	MATH-500	500	0.550	0.549	0.175
Mean across 12 cells		–	**0.624**	**0.659**	**0.197**

Table 12: Instance-level relationship between the two CORD components, $\bar{a}$ and $q_B^\star$, across six models and two tasks. All twelve model–task cells exhibit positive linear and rank associations. NMI measures complementary statistical dependence; larger values indicate greater dependence but should not be interpreted as a literal fraction of shared information.

Model family	Mean r	Mean ρ	Mean NMI
R1-Distill	**0.773**	**0.765**	**0.280**
Math-tuned	0.531	0.607	0.156
Plain	0.567	0.604	0.155

Table 13: Cross-axis association averaged within model families over GSM8K and MATH-500. R1-Distill models show stronger coupling between the two CORD components, while Math-tuned and Plain models exhibit more moderate association.

cates greater answer preservation under rationale disruption, and a larger v indicates earlier answer recovery from the rationale prefix alone. We require the combined score to satisfy four corresponding properties, summarized in Table 14.

A1 and A2 ensure that strengthening either shortcut-consistent signature cannot lower the resulting score. A4 fixes the opposite endpoint: $\mathrm{CORD} = 1$ when both signatures are maximal. The central property is A3. The shortcut signature requires evidence from both disruption axes, so a strong response on only one axis should not by itself produce a positive joint score. The harmonic mean implements this requirement directly:

$$H(u, 0) = H(0, v) = 0.$$

For comparison, an arithmetic mean would give $\mathrm{AM}(1, 0) = 0.5$, assigning a substantial score when one of the two required signatures is entirely absent.

The harmonic mean also preserves the scale of the two components when they agree:

$$H(t, t) = t, \qquad t \in [0, 1].$$

Thus, equal evidence on the two axes remains on the same $[0, 1]$ scale, while imbalance between them lowers the combined score. CORD therefore acts as a continuous conjunction: high values require both frequent answer preservation under A-side disruption and early recovery under the B-side test.

B.5.2 Cross-task rank preservation

As an empirical robustness check, we ask whether different dual-axis summaries induce a consistent ordering of the six models across GSM8K and MATH-500. The comparison is performed in three steps: we first construct each metric at the instance level, average it within each model–task cell, and then compute Kendall's τ between the resulting six-model rankings on the two tasks.

Position-based summaries. For the A-axis, we define a rank-specific saturation position as the first intervention position at which the binary answer-preservation indicator becomes zero:

$$p^\star_{A,r}(x) = \min \{p \in \mathcal{P}_A : a(p, r; x) = 0\},$$

with $p^\star_{A,r}(x) = 100$ if no such position exists. The instance-level A-side position summary averages this cutoff over the five inverse-decoding ranks,

$$p^\star_A(x) = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} p^\star_{A,r}(x).$$

We orient it so that larger values indicate the shortcut-consistent direction:

$$q^\star_A(x) = \frac{p^\star_A(x)}{100}.$$

Property	Requirement	Harmonic-mean verification
A1: Monotonicity in A	$\forall v \in (0,1],\ u \uparrow \Rightarrow H \uparrow$	$\frac{\partial H}{\partial u} = \frac{2v^2}{(u+v)^2} > 0$
A2: Monotonicity in B	$\forall u \in (0,1],\ v \uparrow \Rightarrow H \uparrow$	$\frac{\partial H}{\partial v} = \frac{2u^2}{(u+v)^2} > 0$
A3: Zero boundary	$u = 0 \ \vee \ v = 0 \Rightarrow H = 0$	numerator $2uv = 0$
A4: Upper boundary	$u = 1 \ \wedge \ v = 1 \Rightarrow H = 1$	$H(1,1) = 1$

Table 14: Properties required of the scalar combination of the two CORD components. A1–A2 preserve the direction of each shortcut-consistent signal; A3 implements the dual-axis conjunction by assigning zero when either component is absent; and A4 fixes the upper endpoint when both components are maximal.

For the B-axis, the recovery position is

$$p_B^\star(x) = \min\{p \in \mathcal{P}_B : b(p;x) = 1\},$$

again setting $p_B^\star(x) = 100$ when the cutoff is never reached. Its oriented score is

$$q_B^\star(x) = 1 - \frac{p_B^\star(x)}{100}.$$

Compared metrics. Using these quantities, we construct three instance-level scores:

$$\begin{aligned} S_A(x) &= q_A^\star(x), \\ S_B(x) &= q_B^\star(x), \\ \mathrm{CORD}(x) &= H(\bar{a}(x), q_B^\star(x)), \end{aligned}$$

where H denotes the harmonic mean. Thus, S_A and S_B use a single position-based axis, and CORD replaces the A-side position summary with the full-grid answer-preservation component $\bar{a}$ established in Appendix B.3.1.

For each metric s, we first average the instance-level scores within each model–task cell,

$$\bar{s}_{m,t} = \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} s_m(x),$$

and rank the six models by $\bar{s}_{m,t}$ separately on GSM8K and MATH-500. Table 15 reports Kendall's τ between these two six-model rankings.

All three constructions yield positively associated model rankings across the two tasks. The position-only summaries give $\tau = +0.600$ for S_A and $+0.733$ for S_B. The final CORD construction shows the strongest cross-task agreement, with $\tau = +0.867$.

Because the comparison contains only six models, these rank correlations are best viewed as a robustness check rather than a separate statistical validation criterion. Together with the analytic properties above, they support the final construction $\mathrm{CORD} = H(\bar{a}, q_B^\star)$.

Metric	Cross-task Kendall τ
S_A-only	$+0.600$
S_B-only	$+0.733$
CORD	$\mathbf{+0.867}$

Table 15: Cross-task preservation of the six-model ranking between GSM8K and MATH-500. Each metric is constructed per instance and then averaged within each model–task cell before ranking the six models. S_A and S_B use the A- and B-side position summaries, respectively; and CORD combines the full-grid A-component $\bar{a}$ with the B-side recovery component $q_B^\star$. Higher Kendall τ indicates stronger preservation of the model ordering across the two tasks.

B.6 Relation to Early Answering

Early Answering (EA), introduced by Lanham et al. (2023), provides a natural external comparison for CORD because it also measures how early the baseline answer can be recovered from a rationale prefix. For an instance x, let $b_{\mathrm{EA}}(p;x)$ indicate whether the baseline answer is reproduced when the model receives the original input X together with the prefix $M[0{:}p]$. We summarize the resulting recovery curve as

$$E_{\mathrm{EA}}(x) = \int_0^1 b_{\mathrm{EA}}(100u;x)\,du,$$

with the integral evaluated over the measured prefix grid. Larger values indicate earlier and more persistent answer recovery.

EA and CORD are expected to be related because both are sensitive to early recoverability of the baseline answer. Their information conditions, however, differ in an important respect: EA keeps X available throughout the test, whereas CORD's B-axis masks X and measures whether the visible rationale prefix alone supports recovery. We examine this relationship through both instance-level agreement and the underlying EA recovery profiles.

B.6.1 Cross-method convergence at the instance level

We first measure the instance-level Spearman correlation between CORD and the EA score within each of the six models and two tasks. Table 16 reports the resulting twelve correlations.

Model	GSM8K	MATH-500
R1D-Qwen	0.724	0.760
R1D-Llama	0.859	0.856
Qwen-Math	0.737	0.830
Qwen-7B	0.753	0.825
DeepSeek-Math	0.694	0.717
Llama-3.1	0.712	0.592
Mean	**0.755**	

Table 16: Instance-level Spearman correlation between CORD and Early Answering. The two measurements are positively associated in every model–task cell ($\rho = 0.592$–0.859; mean 0.755).

CORD and EA are positively associated in every model–task cell, with Spearman ρ ranging from 0.592 to 0.859 and a mean of 0.755. Thus, instances receiving higher CORD scores also tend to permit earlier answer recovery under EA. The association remains substantial across both tasks and all six models, indicating that the two diagnostics capture a shared instance-level behavioral tendency.

The correspondence is nevertheless incomplete in every cell. This is consistent with their different information conditions: EA retains X while varying the visible rationale prefix, whereas CORD's B-axis removes direct access to X. The positive but imperfect correlations therefore establish cross-method convergence without making the two measurements interchangeable.

B.6.2 Early-Answering recovery profiles

The recovery trajectories make this relationship more concrete. Table 17 reports the fraction of instances for which EA reproduces the baseline answer at selected rationale-prefix positions.

The trajectories show a clear difference in how early the model families recover their baseline answers. On GSM8K, the two R1-Distill models already reach recovery rates of 0.911 and 0.848 at $p = 50$, whereas the four non-distilled models range from 0.140 to 0.389. The same pattern appears on MATH-500: at $p = 50$, R1D-Qwen and R1D-Llama reach 0.714 and 0.680, compared with 0.233–0.412 for the remaining models. By $p = 75$, the R1-Distill recovery rates reach at least 0.83 on both tasks, whereas the other models recover more gradually.

The $p = 0$ column also exposes the key information-condition difference between EA and CORD's B-axis. With no rationale prefix visible, EA still recovers the baseline answer on 8.8–30.1% of GSM8K instances and 15.1–32.4% of MATH-500 instances. At $p = 0$, the only task-specific information available to EA is the original input X; the visible rationale contributes nothing. These nonzero recovery rates therefore show that EA can succeed through direct access to X, independently of information supplied by the rationale prefix.

This observation explains why EA is an informative comparison but not a substitute for CORD's sufficiency-side test. EA asks how early the answer can be recovered when X remains available, whereas CORD masks X to ask whether the rationale prefix itself is sufficient to support the answer. The two diagnostics consequently exhibit substantial empirical convergence while imposing different information requirements.

Taken together, the two experiments provide complementary evidence. The consistently positive CORD–EA correlations establish convergence at the instance level, while the recovery trajectories show both the shared early-recovery pattern and the role of direct access to X in EA. Early Answering therefore serves as an external behavioral comparison for CORD, whereas the X-masked B-axis isolates the distinct sufficiency condition required by the dual-disruption framework.

B.7 Experimental Setup Details

Environment. All measurements were run on `vLLM` 0.19.0 (Kwon et al., 2023), with `transformers` 4.57.6 (PyTorch 2.10.0) for tokenization, model loading, and the attention-masked forward passes, under Python 3.12. Each measurement cell (one model $\times$ one dataset) was run on a single A6000 (48 GB), with all six models in `bfloat16`. The two R1-Distill models are given a 24,576-token context and a 16,384-token generation budget so that a long CoT is never truncated; the other four use 2,048 tokens on GSM8K and 4,096 on MATH-500. All CORD generations used greedy decoding (temperature $= 0$): the baseline answer $y_0(x)$, the disrupted answer $y_A(p, r; x)$, and the X-mask answer $y_B(p; x)$ are reproduced rather than sampled, so instance-level comparisons carry no sampling noise.

	GSM8K					MATH-500			
Model	p=0	p=25	p=50	p=75	p=100	p=0	p=50	p=75	p=100
R1D-Qwen	0.237	0.401	0.911	0.980	0.998	0.256	0.714	0.846	0.988
R1D-Llama	0.088	0.309	0.848	0.983	0.999	0.151	0.680	0.830	0.992
Qwen-Math	0.301	0.309	0.389	0.709	1.000	0.324	0.412	0.574	0.988
Qwen-7B	0.236	0.243	0.306	0.731	1.000	0.268	0.359	0.544	0.994
DeepSeek-Math	0.186	0.179	0.258	0.542	1.000	0.205	0.233	0.394	0.996
Llama-3.1	0.139	0.112	0.140	0.602	1.000	0.168	0.287	0.437	0.996

Table 17: Early Answering recovery rates under the X-keep condition at selected rationale-prefix positions. Each entry is the fraction of instances for which the model reproduces its baseline answer from the original input X and the visible prefix $M[0{:}p]$.

Inference budget. CORD costs 127 model-generation calls per instance: 105 A-side disruption calls ($|\mathcal{P}| \times |\mathcal{R}| = 21 \times 5 = 105$), 21 B-side disruption calls ($|\mathcal{P}| = 21$), and one baseline generation.

Use of AI Assistance We used a generative AI tool for language polishing and writing assistance. All technical content and research conclusions were developed and verified by the authors.

C Robustness of CORD and the Main Findings

This appendix stress-tests the main findings of Section 4.2 and examines the robustness of CORD. Appendices C.1 and C.2 test whether Finding 1 can be explained by baseline CoT *length* or R1-Distill's tendency to state answers early, while Appendix C.3 reports the complete set of 30 paired Wilcoxon contrasts underlying Finding 2. Appendix C.4 examines CORD under meaning-preserving paraphrasing and progressive rationale compression. Method-specific analyses of disruption design, component construction, aggregation, and Early Answering are provided in Appendix B.

C.1 Length-stratified robustness check

A natural concern about the paradigm-level CORD gap reported in Section 4 is that R1-Distill models produce substantially longer chains-of-thought than their non-distilled counterparts. One possible explanation is therefore that the observed gap reflects *instance-level reasoning demand* rather than a training-paradigm difference: problems that elicit longer reasoning traces may also systematically induce larger CORD differences, producing an apparent paradigm-level separation.

Why stratification rather than forced length matching. Directly forcing all models to generate CoTs of the same length would not provide a clean control for this concern. CoT length is itself an outcome of the model's generation policy, so constraining it would intervene on the reasoning behavior being measured rather than merely control an observed covariate. Truncating longer traces or forcing shorter traces to continue would likewise alter the native generation trajectory. We therefore retain each model's naturally generated baseline CoT and address the length concern through paired stratification.

Setup. We use two within-base-family contrasts: R1D-Llama versus Llama-3.1-8B-Instruct, and R1D-Qwen versus Qwen2.5-Math-7B-Instruct. In the Llama family, both models are derived from Llama-3.1-8B-Base; in the Qwen family, both are derived from Qwen2.5-Math-7B-Base. The contrasts therefore compare R1-distilled and instruction-tuned descendants of the same base-model family, providing a natural setting for testing whether the paradigm-level CORD gap survives length stratification.

For each instance i, let ℓ_i^{R1D} and ℓ_i^{non} denote the baseline CoT token counts of the two models in a contrast. We define the pair-mean baseline length as $\bar{\ell}_i = \frac{\ell_i^{\mathrm{R1D}} + \ell_i^{\mathrm{non}}}{2}$. Instances are divided into quartiles according to $\bar{\ell}_i$, and the paired CORD difference is recomputed separately within each quartile. The pair-mean stratifier preserves pairing: both models are evaluated on exactly the same instances within each stratum, so the within-quartile Wilcoxon comparison remains well-defined. At the same time, stratification reduces variation in the amount of reasoning elicited by the instance across the model pair.

The test targets a specific alternative explanation. If the overall R1-Distill CORD gap were mainly produced by a different mixture of short- and long-reasoning instances, the gap should substantially attenuate or disappear when comparisons are re-

stricted to instances with comparable pair-mean baseline CoT lengths.

Results. The length-stratified analysis provides strong evidence against a simple length-composition explanation of the paradigm-level CORD gap. Across the two datasets and two base-family-aligned contrasts, R1-Distill has higher mean CORD in all 16 quartile cells, and 15 of the 16 differences remain significant at $p < 10^{-6}$ (Table 2). Thus, the aggregate separation does not collapse when the comparison is repeated within substantially narrower regimes of naturally generated pair-mean CoT length.

On GSM8K, R1D-Llama remains above Llama-3.1 in every quartile: $+0.296$, $+0.323$, $+0.379$, $+0.449$, with all four comparisons satisfying $p < 10^{-51}$. The gap increases monotonically with length quartile rather than attenuating among instances that elicit longer reasoning. Thus, for this contrast, the aggregate separation cannot be explained simply by a different mixture of short- and long-reasoning instances.

R1D-Qwen versus Qwen2.5-Math shows the same robust separation on GSM8K, with quartile differences of $+0.271$, $+0.288$, $+0.310$, $+0.230$, and all four comparisons satisfying $p < 10^{-29}$. Although the effect is somewhat smaller in Q4, a substantial positive gap remains across the full range of pair-mean reasoning lengths.

On MATH-500, R1D-Llama likewise exceeds Llama-3.1 in every quartile: $+0.298$, $+0.349$, $+0.351$, $+0.269$, with all four comparisons satisfying $p < 10^{-7}$. The paradigm-level separation therefore persists across both short- and long-reasoning strata for the Llama family on both datasets.

For R1D-Qwen versus Qwen2.5-Math on MATH-500, the first three quartiles, with differences of +0.127, +0.216, and +0.226, also retain clear positive differences, with all three satisfying $p < 10^{-6}$. In the longest quartile, however, the contrast attenuates to $\Delta = +0.021$, $p = 0.32$. This single null cell indicates that the R1-Distill effect is not invariant to reasoning length: for R1D-Qwen and Qwen2.5-Math, the separation can vanish in the extreme long-trace regime of MATH-500. Importantly, however, this is the only attenuation among the 16 stratified comparisons; the other 15 preserve both the direction of the headline effect and strong statistical support.

The stratification itself spans a broad range of naturally occurring CoT lengths (Table 18). The ratio between the Q4 and Q1 upper bounds ranges from $11.0\times$ to $17.3\times$ across the four task–contrast combinations. The persistence of the CORD gap across these widely separated length regimes therefore argues against the possibility that the main paradigm-level result is merely an artifact of concentrating one model family in longer-reasoning instances.

Conclusion. The paradigm-level CORD separation is robust to stratification over a broad range of naturally generated pair-mean CoT lengths. R1-Distill has higher mean CORD in all 16 quartile cells, with 15 of 16 differences significant at $p < 10^{-6}$, even though the strata span $11.0\times$–$17.3\times$ from the shortest to longest quartile bounds. These results provide strong evidence that the main separation in Section 4.2 is not merely an artifact of instance-level reasoning demand as reflected in baseline CoT length. The single attenuation for R1D-Qwen versus Qwen2.5-Math in the longest MATH-500 quartile further shows that length can modulate the magnitude of the effect in an extreme long-trace regime, without accounting for the overall paradigm-level pattern.

C.2 Stress-testing the early-answer behavior of R1-Distill models

R1-Distill models exhibit a distinctive generation behavior: they tend to state the answer relatively early and continue the CoT afterward. Table 19 makes this pattern explicit. On GSM8K, the complete answer appears somewhere in the CoT for nearly all models, but its median first-occurrence position is only 0.39–0.40 for R1-Distill, compared with 0.91 for the non-distilled models. The same pattern holds on MATH-500, where the corresponding medians are 0.30 versus 0.87–0.88. Thus, what distinguishes R1-Distill is not simply whether the answer appears in the rationale, but how early it is stated.

This behavior is itself a meaningful model characteristic. An R1-Distill model may form and state its answer early and then continue with verification, explanation, or elaboration. At the same time, it raises a strong alternative explanation for the high CORD scores in Finding 1 in Section 4.2. Because the B-axis searches for the earliest prefix that recovers the baseline answer, stating the answer early could reduce $p_B^\star$ and thereby increase CORD. The question is therefore whether the paradigm-level separation is driven primarily by this R1-Distill-

Task	Contrast	Q1	Q2	Q3	Q4	Q4/Q1
GSM8K	R1D-Llama vs. Llama-3.1	533	668	911	9,216	17.3×
GSM8K	R1D-Qwen vs. Qwen-Math	581	707	901	9,216	15.9×
MATH-500	R1D-Llama vs. Llama-3.1	895	1,434	3,799	10,240	11.4×
MATH-500	R1D-Qwen vs. Qwen-Math	924	1,387	2,701	10,160	11.0×

Table 18: Pair-mean baseline-CoT-length quartile upper bounds, in tokens. The Q4 upper bound is $11.0\times$–$17.3\times$ the Q1 upper bound across the four task–contrast combinations, showing that the stratification covers substantially different reasoning-length regimes.

Model	Task	Answer in CoT (%)	First-answer position	
			Mean	Median
R1D-Llama	GSM8K	98.9	0.37	0.39
R1D-Llama	MATH-500	79.8	0.31	0.30
R1D-Qwen	GSM8K	98.6	0.38	0.40
R1D-Qwen	MATH-500	77.2	0.30	0.30
Llama-3.1	GSM8K	96.7	0.83	0.91
Llama-3.1	MATH-500	71.4	0.68	0.88
Qwen-Math	GSM8K	98.7	0.84	0.91
Qwen-Math	MATH-500	83.0	0.66	0.87

Table 19: Early answer appearance in generated CoTs. "Answer in CoT" is the fraction of instances in which the complete baseline-answer string appears in the rationale. First-answer position is its normalized first-occurrence position among such instances, where 0 and 1 denote the beginning and end of the CoT, respectively.

specific behavior.

A conservative stress test. We test this explanation under an intentionally unfavorable treatment of R1-Distill's early-answer behavior. Let p_{lit} denote the first grid position at which the complete baseline-answer string becomes visible in the CoT prefix, and let $p_B^\star$ denote the original B-axis saturation position.

If $p_{\text{lit}} > p_B^\star$, recovery already occurs before the answer literal appears, so early answer availability cannot explain the saturation. Likewise, if the answer is visible by p_{lit} but $b(p_{\text{lit}}) = 0$, its presence is not sufficient to explain the later recovery. We therefore identify *answer-aligned recovery* only when

$$p_{\text{lit}} \le p_B^\star \qquad \text{and} \qquad b(p_{\text{lit}}) = 1,$$

i.e., recovery is already successful at the first position where the complete answer becomes available.

For every such instance, we deliberately remove the entire B-side advantage by assigning the minimum possible B-component:

$$p_{B,\text{stress}}^\star = \begin{cases} 100, & p_{\text{lit}} \le p_B^\star \ \wedge \ b(p_{\text{lit}}) = 1, \\ p_B^\star, & \text{otherwise.} \end{cases}$$

This is a particularly stringent stress test because we do not attempt to distinguish literal reuse from genuine early answer formation. Even if the model genuinely reaches its conclusion early, writes the answer, and then continues reasoning, we remove that B-side signal entirely. The test therefore asks whether the R1-Distill separation survives after deliberately excluding the behavioral component most favorable to producing early B-axis recovery.

Results. As shown in Table 20, this treatment reduces the magnitude of the R1-Distill advantage, indicating that early-answer behavior accounts for part of the raw separation. Crucially, however, it does not remove the paradigm-level effect. All four R1-Distill contrasts remain positive and statistically significant after the stress test, with differences ranging from $+0.043$ to $+0.176$.

The reduction is substantial in some comparisons, which is consistent with the pronounced early-answer behavior observed in Table 19. R1-Distill models state the complete answer much earlier in their CoTs: their median first-answer positions are 0.39–0.40 on GSM8K and 0.30 on MATH-500, compared with 0.91 and 0.87–0.88, respectively, for the base-family-aligned non-distilled models. This early placement can naturally favor smaller $p_B^\star$ and hence contribute to higher CORD.

The stress test deliberately removes this source of B-side advantage whenever recovery is already successful at the first availability of the complete

Task	Contrast	Original Δ	Stress-test Δ	Gap retained
GSM8K	R1D-Llama vs. Llama-3.1	+0.361	+0.099 ($p = 2 \times 10^{-26}$)	27.4%
GSM8K	R1D-Qwen vs. Qwen-Math	+0.275	+0.176 ($p = 2 \times 10^{-75}$)	64.0%
MATH-500	R1D-Llama vs. Llama-3.1	+0.317	+0.130 ($p = 4 \times 10^{-12}$)	41.0%
MATH-500	R1D-Qwen vs. Qwen-Math	+0.148	+0.043 ($p = 2 \times 10^{-3}$)	29.1%

Table 20: Stress test of the R1-Distill early-answer behavior. Δ denotes mean CORD_{R1D} minus mean $\text{CORD}_{\text{non-distilled}}$. For every instance in which recovery is already successful when the complete answer first becomes visible, we assign the minimum possible B-side score, regardless of whether the behavior reflects literal reuse or genuine early answer formation. Although the R1-Distill advantage decreases, all four contrasts remain positive and statistically significant. "Gap retained" reports the stress-test difference relative to the original difference.

answer. As Table 20 shows, the resulting gaps become smaller, substantially in some comparisons, but the paradigm-level separation remains: all four base-family-aligned contrasts are still positive and statistically significant, with stress-test differences ranging from $+0.043$ to $+0.176$. Thus, the higher CORD of R1-Distill models cannot be explained solely by their tendency to state answers earlier.

Accordingly, we retain the original CORD as the primary measure. Stating an answer early and continuing with verification or elaboration can itself be a genuine characteristic of the model's reasoning behavior and is compatible with the functional-redundancy spectrum targeted by CORD. The stress-test score intentionally removes this behavior together with any possible advantage from literal answer availability. It should therefore be interpreted not as a corrected version of CORD, but as a deliberately stringent robustness test: even after excluding the R1-Distill behavior most favorable to early B-axis recovery, the paradigm-level separation persists.

C.3 Full paired Wilcoxon analysis

To characterize the model-level separation more completely, we report all pairwise CORD comparisons across the six models and the two datasets in Table 21. With six models, each dataset yields $\binom{6}{2} = 15$ model pairs.

Paired test. All comparisons are performed at the instance level. For each model pair (m_1, m_2) and an instance x_i, we compute

$$d_i = \text{CORD}_{m_1}(x_i) - \text{CORD}_{m_2}(x_i)$$

and apply a two-sided Wilcoxon signed-rank test. We use a paired non-parametric test because the instance-level CORD distributions need not be normally distributed and contain substantial mass near zero.

For each comparison, we report

$$\Delta_{\text{mean}} = \text{mean}_i\, d_i, \qquad \Delta_{\text{median}} = \text{median}_i\, d_i,$$

where positive values indicate higher CORD for the first model. Because each dataset contains 15 pairwise comparisons, we apply Bonferroni correction separately within each dataset:

$$\alpha_{\text{Bonf}} = \frac{0.05}{15} \approx 0.00333.$$

Results. The clearest pattern is the consistency of the R1-Distill separation. Every R1-Distill versus non-distilled comparison is positive on both datasets. On GSM8K, the eight contrasts range from $\Delta_{\text{mean}} = +0.258$ to $+0.381$; on MATH-500, they range from $+0.148$ to $+0.327$. All 16 R1-Distill versus non-distilled comparisons remain significant after Bonferroni correction within each dataset.

By comparison, the difference between the two R1-Distill models is much smaller. R1D-Qwen exceeds R1D-Llama by only $+0.016$ on GSM8K, and the ordering reverses on MATH-500 (-0.020). The GSM8K difference is statistically detectable given the large paired sample ($p = 8.2 \times 10^{-11}$), but its magnitude is small relative to the R1-Distill versus non-distilled gaps. On MATH-500, the two R1-Distill models are statistically indistinguishable ($\Delta_{\text{median}} = +0.001$, $p = 0.92$).

The non-distilled models exhibit additional model-specific variation, but this pattern is less uniform. For example, Qwen-Math exceeds Qwen-7B by only $+0.037$ on GSM8K and $+0.020$ on MATH-500, whereas some other non-distilled contrasts are larger. DeepSeek-Math and Llama-3.1 are nearly indistinguishable in mean CORD on both datasets (-0.004 on GSM8K and -0.011 on MATH-500), and neither comparison survives the Bonferroni threshold.

Taken together, the full paired analysis reveals a highly consistent R1-Distill signature. Across both

Task	Contrast	n	Δ_{mean}	Δ_{median}	W	p
GSM8K	R1D-Qwen vs. R1D-Llama	1319	+0.016	+0.039	340371	$8.2 \times 10^{-11\ddagger}$
	R1D-Qwen vs. Qwen-Math	1319	+0.275	+0.312	40496	$4.8 \times 10^{-179\ddagger}$
	R1D-Qwen vs. Qwen-7B	1319	+0.311	+0.333	20940	$5.1 \times 10^{-197\ddagger}$
	R1D-Qwen vs. DeepSeek-Math	1319	+0.381	+0.400	8227	$5.4 \times 10^{-209\ddagger}$
	R1D-Qwen vs. Llama-3.1	1319	+0.378	+0.392	6580	$1.4 \times 10^{-210\ddagger}$
	R1D-Llama vs. Qwen-Math	1319	+0.258	+0.282	52367	$2.3 \times 10^{-168\ddagger}$
	R1D-Llama vs. Qwen-7B	1319	+0.295	+0.303	21541	$2.8 \times 10^{-196\ddagger}$
	R1D-Llama vs. DeepSeek-Math	1319	+0.365	+0.381	9789	$1.2 \times 10^{-207\ddagger}$
	R1D-Llama vs. Llama-3.1	1319	+0.361	+0.362	6387	$6.0 \times 10^{-211\ddagger}$
	Qwen-Math vs. Qwen-7B	1319	+0.037	+0.017	340772	$1.4 \times 10^{-9\ddagger}$
	Qwen-Math vs. DeepSeek-Math	1319	+0.107	+0.079	185298	$1.1 \times 10^{-71\ddagger}$
	Qwen-Math vs. Llama-3.1	1319	+0.103	+0.074	192330	$4.2 \times 10^{-67\ddagger}$
	Qwen-7B vs. DeepSeek-Math	1319	+0.070	+0.066	230988	$3.8 \times 10^{-47\ddagger}$
	Qwen-7B vs. Llama-3.1	1319	+0.066	+0.050	237904	$8.0 \times 10^{-44\ddagger}$
	DeepSeek-Math vs. Llama-3.1	1319	-0.004	-0.013	395285	1.5×10^{-2}
MATH-500	R1D-Qwen vs. R1D-Llama	500	-0.020	+0.001	57414	9.2×10^{-1}
	R1D-Qwen vs. Qwen-Math	500	+0.148	+0.148	28602	$1.5 \times 10^{-22\ddagger}$
	R1D-Qwen vs. Qwen-7B	500	+0.168	+0.195	25772	$4.2 \times 10^{-27\ddagger}$
	R1D-Qwen vs. DeepSeek-Math	500	+0.307	+0.323	10744	$6.5 \times 10^{-56\ddagger}$
	R1D-Qwen vs. Llama-3.1	500	+0.296	+0.350	14344	$4.2 \times 10^{-48\ddagger}$
	R1D-Llama vs. Qwen-Math	500	+0.168	+0.171	27727	$8.6 \times 10^{-24\ddagger}$
	R1D-Llama vs. Qwen-7B	500	+0.188	+0.197	23526	$4.5 \times 10^{-31\ddagger}$
	R1D-Llama vs. DeepSeek-Math	500	+0.327	+0.368	9522	$7.9 \times 10^{-59\ddagger}$
	R1D-Llama vs. Llama-3.1	500	+0.317	+0.372	12245	$2.0 \times 10^{-51\ddagger}$
	Qwen-Math vs. Qwen-7B	500	+0.020	+0.005	44822	1.8×10^{-2}
	Qwen-Math vs. DeepSeek-Math	500	+0.159	+0.093	25937	$4.8 \times 10^{-23\ddagger}$
	Qwen-Math vs. Llama-3.1	500	+0.149	+0.095	28693	$2.4 \times 10^{-19\ddagger}$
	Qwen-7B vs. DeepSeek-Math	500	+0.139	+0.085	27022	$7.6 \times 10^{-21\ddagger}$
	Qwen-7B vs. Llama-3.1	500	+0.129	+0.084	31003	$3.7 \times 10^{-17\ddagger}$
	DeepSeek-Math vs. Llama-3.1	500	-0.011	+0.000	50060	6.3×10^{-1}

Table 21: Full paired Wilcoxon signed-rank comparisons of CORD across all six models. Δ_{mean} and Δ_{median} denote paired instance-level differences, with positive values indicating higher CORD for the first model. Bonferroni correction is applied separately within each dataset over its 15 pairwise comparisons ($\alpha_{\text{Bonf}} = 0.05/15 \approx 0.00333$); ‡ denotes comparisons that remain significant after correction. All 16 R1-Distill versus non-distilled contrasts are positive and satisfy the corrected threshold.

GSM8K and MATH-500, both R1-Distill models have higher CORD than every non-distilled comparator, and all 16 pairwise contrasts remain significant after within-dataset Bonferroni correction. This consistency across model identity, backbone, and task supports the paradigm-level interpretation of CORD.

C.4 Robustness to Paraphrasing and Compression

We further examine whether CORD is sensitive to changes in the surface realization of a rationale. We compare two transformations of the baseline rationale M: paraphrasing, which rewrites the solution while preserving its content and answer, and compression, which produces progressively shorter versions of the same solution. The comparison asks whether CORD remains stable under meaning-preserving re-expression and how its response changes as the rationale is increasingly compressed.

Setup. We evaluate two representative R1-Distill settings: R1D-Qwen on GSM8K and R1D-Llama on MATH-500. For each setting, we divide instances into low-, mid-, and high-CORD strata according to their baseline CORD scores. Because rationale length may differ across these strata, we restrict the candidate instances to a shared baseline-length range before sampling 20 instances from each stratum, yielding 60 instances per setting.

Our primary analysis uses the overlap of the 20th–80th percentile rationale-length ranges across the three strata (*q20*), which provides a tighter length-matched comparison. We additionally repeat the experiment using the broader overlap between the minimum and maximum rationale lengths (*full*). For every selected baseline ratio-

nale, we generate three paraphrases and three compressed variants targeting 75%, 50%, and 25% of the original length. All retained variants preserve the baseline answer and satisfy the corresponding transformation constraints.

Paraphrasing largely preserves CORD. CORD changes little when the same rationale is paraphrased. Across the four settings, the median absolute change is only 0.015–0.019. The instance-level ordering is also highly stable: Spearman correlations between the original CORD and the paraphrase-averaged CORD range from 0.963 to 0.987 (Table 22). Thus, the measurement is largely preserved under the paraphrase transformations considered here.

Compression systematically lowers CORD. In contrast to paraphrasing, compression consistently shifts CORD downward, with larger changes under stronger compression. Under the primary q20 analysis, the median $|\Delta\mathrm{CORD}|$ increases from 0.028 to 0.036 to 0.058 on GSM8K and from 0.049 to 0.062 to 0.087 on MATH-500 as the target length decreases from 75% to 50% to 25% of the original rationale. The broader full-support analysis shows the same directional pattern, with the largest change again occurring under the strongest compression.

At the 25% compression level, the CORD change is $3.4\times$ and $5.4\times$ larger than the paraphrase effect in the two q20 settings, respectively. The same contrast holds under the full support, where the corresponding ratios are $2.3\times$ and $5.9\times$. Thus, across both tasks and both support definitions, CORD remains comparatively stable under paraphrasing, whereas stronger compression produces progressively larger downward changes.

Taken together, the two transformations yield a consistent contrast: CORD changes little under paraphrasing, while compression lowers the score and produces larger changes as compression becomes stronger. These results reduce the concern that CORD is uniformly sensitive to changes in rationale wording or form. They do not, however, identify which specific properties of the compressed rationales are responsible for the observed decrease.

D Self-Refine Validation and Application

This appendix provides the detailed self-refine analysis used in Section 4.3. We use self-refine as an external setting for validating CORD.

The validation has two steps. First, we examine what actually drives self-refine. In particular, we ask whether refinement depends on the previous rationale M, previous answer y_0, critique C, and refinement instruction, as the critique-and-revise procedure suggests. We test this without using CORD. Second, based on the behavior found in the first step, we derive a simple prediction for CORD and test it in Appendix D.2.

The first step gives two results. Changing or removing the information supplied to refinement has little effect on how often the answer changes, while almost all refinement-induced answer changes occur on instances that are already unstable under X-only resampling. Together, these results show that self-refine in this setting behaves consistently with X-only stochastic resampling. We then test whether CORD recovers the corresponding prediction, and examine its exploratory use for refinement selection. In particular, we sweep the CORD threshold to test whether higher CORD scores provide a graded signal of when refinement can be skipped. Finally, we summarize the cost of this use case and clarify that the selector is most appropriate when CORD has already been computed, rather than as a cold-start online policy.

D.1 Self-Refine Behaves Consistently with X-Only Resampling

Self-refine (Madaan et al., 2023) first generates a critique of the original solution and then asks the model to produce a revised solution. The refinement stage therefore appears to use several pieces of information: the original rationale M, the original answer y_0, the critique C, and the refinement instruction.

We ask whether these inputs actually determine whether refinement changes the answer. We use two tests. First, we modify or remove each refinement input and measure how much the answer-change rate changes. Second, we test whether refinement changes occur on the same instances that are unstable when the model is sampled from X alone.

D.1.1 Self-refine protocol and controlled ablations

Self-refine consists of a critique stage and a refinement stage. For each instance, we first generate the critique using the original problem, rationale, and answer:

Support	Task	Para.	ρ	C75	C50	C25	C25/Para.
q20	GSM8K	0.017	0.963	0.028↓	0.036↓	0.058↓	3.4×
q20	MATH-500	0.016	0.987	0.049↓	0.062↓	0.087↓	5.4×
full	GSM8K	0.019	0.979	0.025↓	0.039↓	0.044↓	2.3×
full	MATH-500	0.015	0.972	0.042↓	0.053↓	0.088↓	5.9×

Table 22: Robustness of CORD to paraphrasing and compression. Para. is the median instance-level $|\Delta\mathrm{CORD}|$ between the original rationale and the mean of its three paraphrases; ρ is the Spearman correlation between original CORD and paraphrase-averaged CORD. C75, C50, and C25 report the median $|\Delta\mathrm{CORD}|$ under increasingly strong compression, while ↓ denotes a downward shift in CORD. C25/Para. gives the ratio between the strongest-compression and paraphrase effects. The q20 rows use the 20th–80th percentile common length support; full uses the broader min/max common support.

```
Below is a math problem and a proposed
solution.

Problem: {query}
Solution: {baseline_cot}
Answer: {y_0}

Review the solution step by step.
Identify any errors
in reasoning or calculation.
Critique:
```

Importantly, the critique is generated only once for each instance and then cached. The *same critique* is used in every refinement ablation. We do not regenerate the critique after changing the refinement prompt. This ensures that differences across ablations come only from the refinement inputs being tested, rather than from differences in newly generated critiques.

The real refinement condition uses the following prompt:

```
Based on the critique below, provide
a refined solution.

Problem: {query}
Original solution: {baseline_cot}
Original answer: {y_0}
Critique: {critique}

Refined solution. After your reasoning,
you MUST end with the final answer
in this exact format on a
separate line:
Final Answer: \boxed{ANSWER}

where ANSWER is your final answer
(number, expression,
or multi-character text).
Refined solution:
```

We then change one part of this refinement prompt at a time, as summarized below.

Condition	Change from the real prompt
real	No change.
trigger anti	First line replaced with `The solution is likely correct; restate unless certain.`
trigger strong	First line replaced with `The solution has clear errors that must be fixed.`
M = `[omitted]`	`Original solution:` is replaced with `[omitted]`.
y_0 removed	The `Original answer: {y_0}` line is removed.
critique = `[omitted]`	`Critique:` is replaced with `[omitted]`.

Table 23: Refinement-input ablations. All conditions use the same cached critique for each instance. Only the indicated part of the refinement prompt is changed.

We use three additional controls. First, the trigger prompts are length-matched: their lengths are restricted to 10–12 tokens around the 11-token real trigger, and this is checked at execution time. This prevents a trigger comparison from being driven simply by prompt length. Second, when removing M or C, we replace its value with `[omitted]` instead of deleting the field, so that the overall prompt structure remains unchanged. The y_0 ablation intentionally removes the entire answer line. Third, the real condition is checked to be byte-identical to the original `REFINE_PROMPT`, ensuring that the ablation experiment does not change the reference condition.

All refinement generations use greedy decoding. The same model–prompt condition is not independently rerun across the analyses below. In particular, the same real-refinement outputs are reused in the subsequent tables.

Ablation	n	Change rate	Δ vs. real
real	464	0.3125	—
trigger anti	467	0.3062	-0.0063
trigger strong	466	0.3069	-0.0056
$M =$ `[omitted]`	480	0.3229	$+0.0104$
y_0 removed	464	0.3103	-0.0022
critique = `[omitted]`	475	0.3137	$+0.0012$

Table 24: Input ablation of self-refine on Qwen2.5-Math-7B × MATH-500. *Change rate* is the fraction of valid instances for which $Y_{\text{refine}} \neq Y_0$. Δ is the difference from the real refinement condition. All ablations use the same cached critique. Across the five ablations, the largest absolute change is 0.0104 (1.04 percentage points). The value of n differs because each ablation changes the deterministic prompt and therefore changes the number of instances that pass the length and answer-extraction filters.

D.1.2 Changing the refinement inputs has little effect

We first test how much each input affects whether refinement changes the answer. Table 24 reports the results for Qwen2.5-Math-7B × MATH-500.

Under the real prompt, refinement changes the answer on 31.25% of valid instances. The rate changes very little under any ablation. Telling the model that the solution is likely correct changes the rate by -0.63 percentage points, while telling it that the solution clearly contains errors changes it by -0.56 points. Removing the original answer changes the rate by only -0.22 points, and removing the critique changes it by $+0.12$ points. Even replacing the entire original rationale with `[omitted]` changes the rate by only $+1.04$ percentage points.

Table 24 does not prove that these inputs are never used on any individual instance. Different instances could change under different ablations even when the overall rates are similar. The result is narrower: changing the rationale, answer, critique, or instruction has very little effect on the *overall rate* at which self-refine changes its answer.

This raises the next question: if these refinement inputs have little effect on the change rate, which instances actually change?

D.1.3 Refinement changes are concentrated in the X-only-unstable subset

We next ask whether the instances changed by self-refine are the same instances that are already unstable when the model is sampled from X alone.

For each instance, we independently sample $K = 10$ answers using only the original input X (with no rationale, critique, or refinement context), with temperature 0.7, top-$p = 1.0$, and a fixed seed of 42. We define an instance as X-only unstable ($U = 1$) if at least one valid resampled answer differs from the baseline greedy answer y_0, and stable ($U = 0$) otherwise; samples for which answer extraction fails are excluded from this comparison. Thus, $U = 1$ denotes X-only resampling instability and $R = 1$ denotes that refinement changes the original answer.

If refinement behaves primarily as another stochastic inference from X, we expect a directional asymmetry: refinement should rarely change the answer when the instance is stable under X-only resampling, whereas instability under X may not by itself be sufficient to produce a refinement change.

Table 25 shows exactly this pattern. Across all four MATH-500 cells, the probability of a refinement-induced answer change on X-only-stable instances is very small,

$$P(R = 1 \mid U = 0) = 0.0\%\text{–}5.4\%.$$

By contrast, among X-only-unstable instances,

$$P(R = 1 \mid U = 1) = 32.0\%\text{–}62.0\%.$$

Correspondingly, 97.3–100.0% of all observed refinement changes occur inside the X-only-unstable subset.

The asymmetry is important. X-only instability is close to a necessary behavioral condition for a refinement-induced answer change: when an instance is stable under X-only resampling, refinement almost never changes its answer. The reverse does not hold. Only 32.0–62.0% of unstable instances change under refinement, so X-only instability is not sufficient for a refinement change.

The base rates in Table 25 also matter for interpreting the final column. For Qwen-Math and Qwen-7B, only 50.4% and 60.0% of instances are X-only unstable, making the concentration of 98.8–100.0% of refinement changes in this subset strongly selective. DeepSeek-Math and Llama-3.1 have much higher instability rates (89.7% and 82.3%), so $P(U \mid R)$ is less informative by itself in these cells. The informative comparison there is the near absence of refinement changes among the remaining stable instances: $P(R \mid \neg U) = 0.0\%$ and 5.4%, respectively.

Together, the two tests point to the same interpretation from different directions. Changing the rationale, answer, critique, or refinement instruction has

Model	n	$P(U)$	$P(R \mid U)$	$P(R \mid \neg U)$	$P(U \mid R)$
Qwen-Math	464	50.4%	62.0%	0.0%	100.0%
Qwen-7B	497	60.0%	56.0%	1.0%	98.8%
DeepSeek-Math	439	89.7%	32.0%	0.0%	100.0%
Llama-3.1	417	82.3%	41.7%	5.4%	97.3%

Table 25: Relationship between refinement-induced answer changes and X-only resampling instability on MATH-500. $U = 1$ denotes an instance that is unstable under X-only resampling, and $R = 1$ denotes an instance whose answer changes under self-refine. $P(U)$ gives the base rate of X-only instability. Refinement changes are rare when $U = 0$ ($P(R \mid \neg U) = 0.0$–5.4%), while 32.0–62.0% of unstable instances change under refinement. Consequently, 97.3–100.0% of all refinement-induced changes occur inside the X-only-unstable subset ($P(U \mid R)$).

little effect on the overall answer-change rate (Table 24), while the answer changes that do occur are almost restricted to instances that are independently unstable under X-only resampling (Table 25). We therefore describe self-refine in this setting as behaviorally consistent with X-only stochastic resampling.

D.2 CORD Recovers the X-Only Prediction

Appendix D.1 showed, without using CORD, that self-refine behaves consistently with X-only resampling. This gives a simple prediction. If refinement leaves the answer unchanged, $Y_{\text{refine}} = Y_0$, then the original and refined rationales may differ, but they still lead to the same X-driven answer. CORD should therefore change very little between the two rationales.

Table 26 tests this prediction. We report both the full set and the answer-unchanged subset. The latter is the direct test of the prediction.

On the full set, the refined measurements are lower on average, because this set also contains instances where refinement changes the answer itself. Nevertheless, baseline and refined measurements remain strongly correlated ($\rho = +0.794$–$+0.822$).

The direct test is the answer-unchanged subset. Here, all four shifts are close to zero, and CORD changes by only $+0.0019$. CORD therefore recovers the near-zero shift predicted by the independently established X-only behavior of self-refine.

D.3 Exploratory Refinement Selection

We next ask whether the *magnitude* of CORD provides useful information for deciding when refinement can be skipped. We define a thresholded policy

$$\text{CORD}(X) > \tau \quad \Rightarrow \quad \text{skip refinement},$$

and evaluate the fixed grid

$$\tau \in \{0, 0.1, \ldots, 0.6\}.$$

We report the full threshold grid in Table 27.

As τ increases, fewer and higher-CORD instances are selected for skipping. If CORD provides a graded signal of refinement stability, these higher-CORD subsets should contain a larger fraction of instances for which refinement would leave the answer unchanged. To separate this effect from the trivial benefit of changing the skip rate, we also compare the resulting policy against random selectors that skip exactly the same number of instances.

Table 27 shows two representative models, Qwen-Math and Llama-3.1. Qwen-Math illustrates the strong ceiling effect on GSM8K and the clear graded pattern on MATH-500, while Llama-3.1 provides a GSM8K cell with substantially more room for enrichment. Qwen-Math and Llama-3.1 are shown as representative cases; the remaining models show the same overall trend of stronger non-change enrichment at higher CORD thresholds.

The pattern is clearest on MATH-500. As the threshold increases, the skipped set becomes smaller but more concentrated on instances for which refinement would not change the answer. For Qwen-Math, for example, the enrichment over the cell-wide non-change rate increases from $+4.4$ percentage points at $\tau = 0$ to more than $+11$ points around $\tau = 0.3$–0.5. Llama-3.1 shows the same overall pattern, with enrichment increasing from $+3.6$ points at $\tau = 0$ to roughly $+9$–11 points over much of the higher-threshold region.

The result is not specific to these two examples. On MATH-500, all four non-distilled models show positive non-change enrichment across the threshold sweep. More importantly, the CORD selector outperforms same-rate random selection in policy accuracy at every reported threshold for all four models. The largest gains occur at intermediate thresholds: $+0.0289$ for Qwen-Math, and $+0.0190$ for Llama-3.1. These values are not used to select

Measurement	Baseline	Refine	Δ	Spearman ρ
Full set ($n = 464$)				
$\bar{a}$ (A-side)	0.497	0.443	-0.0542	$+0.794$
M-mask (X visible)	0.500	0.457	-0.0428	$+0.822$
X/M-mask (B-side)	0.415	0.384	-0.0303	$+0.810$
CORD	0.432	0.391	-0.0410	$+0.814$
Answer-unchanged subset ($n = 319$)				
$\bar{a}$ (A-side)	0.555	0.549	-0.0062	—
M-mask (X visible)	0.574	0.573	-0.0004	—
X/M-mask (B-side)	0.473	0.483	$+0.0099$	—
CORD	0.492	0.494	$+0.0019$	—

Table 26: Baseline versus refined CoT on Qwen2.5-Math-7B $\times$ MATH-500. The answer-unchanged subset is the direct validation condition. In this subset, all four measurements shift by less than 0.01 in absolute value, and CORD changes by only $+0.0019$.

operating thresholds; the entire grid is reported to avoid model-specific threshold tuning.

GSM8K exhibits a weaker effect because refinement already leaves the answer unchanged on 94.8–98.6% of instances for three of the four models. There is therefore little room for higher-CORD subsets to improve further. Llama-3.1, whose base non-change rate is lower at 84.0%, provides the clearest exception: enrichment increases to $+3.3$ percentage points at $\tau = 0.3$, and the same-rate random comparison is also positive at moderate thresholds. At very high thresholds, however, the skipped subset becomes small (e.g., $n_{\text{skip}} = 48$ at $\tau = 0.6$), making the tail less stable. We therefore report n_{skip} explicitly and do not extend the sweep beyond $\tau = 0.6$.

A threshold-free analysis gives the same overall conclusion. Using CORD to rank instances by whether refinement leaves the answer unchanged yields AUROC above chance in all eight model–dataset cells: 0.569–0.703 on GSM8K and 0.637–0.676 on MATH-500.

Overall, these results show that the magnitude of CORD provides a *graded signal* of refinement stability: higher-CORD instances are generally more likely to be cases in which refinement leaves the answer unchanged. We treat this as an exploratory behavioral result rather than a deployment claim; the cost of computing CORD is discussed in the next section.

D.4 Cost and Practical Scope

The selector analysis above assumes that CORD scores are already available. In the Qwen-Math $\times$ MATH-500 cell, full CORD requires about 127 model calls per instance, compared with two calls for one self-refine round (critique and refinement), giving a call-count ratio of roughly $63\times$. Because CORD calls are much shorter, the generated-token ratio is substantially smaller: about 16,933 tokens for CORD versus 1,101 for self-refine, or approximately $15\times$. These ratios are not wall-clock measurements, but they show that computing full CORD solely to decide whether to run a single refinement round still carries substantial cold-start overhead. We therefore treat the selector as an exploratory application for offline or re-evaluation settings where CORD has already been computed, rather than as a cost-effective online serving policy.

Task	Model	τ	Skip (n)	Non-change (Δ_{base})	Gain vs. random
GSM8K	Qwen-Math	0.0	99.8% (1310)	98.6% (+0.1p)	0.0000
		0.1	96.8% (1271)	98.7% (+0.2p)	+0.0007
		0.2	86.2% (1132)	99.1% (+0.6p)	+0.0005
		0.3	64.2% (843)	99.2% (+0.6p)	+0.0003
		0.4	42.6% (559)	99.5% (+0.9p)	+0.0006
		0.5	29.9% (392)	99.5% (+0.9p)	+0.0005
		0.6	19.6% (257)	99.6% (+1.1p)	+0.0003
	Llama-3.1	0.0	98.2% (1295)	84.6% (+0.6p)	+0.0009
		0.1	94.0% (1240)	84.9% (+0.9p)	+0.0070
		0.2	77.6% (1024)	85.5% (+1.5p)	+0.0115
		0.3	49.1% (647)	87.3% (+3.3p)	+0.0082
		0.4	23.8% (314)	86.0% (+2.0p)	+0.0049
		0.5	8.9% (118)	81.4% (−2.6p)	+0.0038
		0.6	3.6% (48)	70.8% (−13.2p)	+0.0024
MATH-500	Qwen-Math	0.0	86.6% (402)	73.1% (+4.4p)	+0.0187
		0.1	75.2% (349)	75.4% (+6.6p)	+0.0194
		0.2	62.5% (290)	77.9% (+9.2p)	+0.0229
		0.3	53.9% (250)	80.0% (+11.3p)	+0.0289
		0.4	45.5% (211)	80.6% (+11.8p)	+0.0226
		0.5	38.8% (180)	80.0% (+11.3p)	+0.0191
		0.6	34.5% (160)	78.8% (+10.0p)	+0.0122
	Llama-3.1	0.0	81.1% (338)	68.3% (+3.6p)	+0.0054
		0.1	68.8% (287)	73.5% (+8.8p)	+0.0099
		0.2	49.2% (205)	75.6% (+10.9p)	+0.0190
		0.3	35.0% (146)	74.0% (+9.2p)	+0.0080
		0.4	26.1% (109)	75.2% (+10.5p)	+0.0064
		0.5	21.3% (89)	73.0% (+8.3p)	+0.0086
		0.6	16.3% (68)	76.5% (+11.7p)	+0.0069

Table 27: Representative CORD threshold sweeps for refinement selection. Refinement is skipped when $\mathrm{CORD}(X) > \tau$, using the fixed grid $\tau \in \{0, 0.1, \ldots, 0.6\}$; no best threshold is selected. *Skip* reports the skip rate and the number of skipped instances. *Non-change* is the fraction of skipped instances for which $Y_{\text{refined}} = Y_0$, with its percentage-point difference from the cell-wide non-change rate in parentheses. *Gain vs. random* is the policy-accuracy difference from the mean of 100 random selectors with the same skip count. Qwen-Math and Llama-3.1 are shown as representative cases; the remaining models show the same positive selection pattern on MATH-500.