Accurate but Causally Disconnected: When Chain-of-Thought Reasoning Fails to Drive Its Own Answers

Anonymous ACL submission

Abstract

The CoT faithfulness regime is not random or fixed by task or model alone; it is shaped by the interaction between task type and training distribution. We demonstrate this with **Normalized Logit Difference Decay** (NLDD), a replacement-based step-level metric that isolates each reasoning step's causal contribution while holding chain structure constant, enabling normalized cross-model comparison. Across five models (6.7B–70B) and four benchmarks, task type predicts faithfulness direction: arithmetic reasoning is load-bearing across all five models, while commonsense reasoning shows a consistent negative depth trend.

Within symbolic tasks, direction alone does not determine regime, training distribution does: two comparable-scale DeepSeek models exhibit opposite causal sensitivity on the same task. Scale modulates magnitude without changing either. In the limiting case, DeepSeek-R1 achieves 100% accuracy with a 0% flip rate under step corruption, indicating systematic chain bypassing: free-generation experiments (N=500 per condition) reveal near-identical behavior on clean and corrupted chains, confirming R1 appears largely insensitive to perturbations of the reasoning trace. These results replace the assumption that high-accuracy CoT is trustworthy with a tractable diagnostic framework: faithfulness is structured and must be established per model-task pair before CoT can serve as a verification signal. [1]

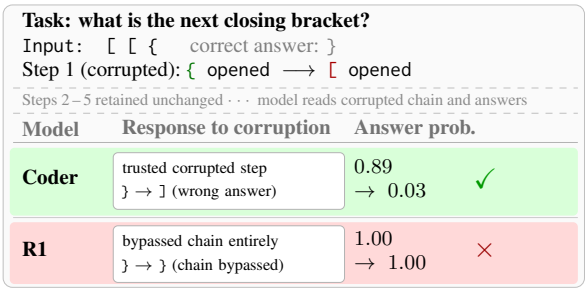


Figure 1: DeepSeek-Coder and DeepSeek-R1-Distill are models of comparable scale from the same developer but with different training distributions, yet exhibit opposite faithfulness regimes on Dyck-n. Corrupting one reasoning step flips Coder's answer; R1 ignores the corruption entirely (NLDD: 8.60 vs. 0.43; flip rate: 3.6% vs. 0%). Free-generation experiments confirm this reflects systematic chain bypassing: R1 behaves identically on clean and corrupted chains. High accuracy does not imply a causally load-bearing chain.

1 Introduction

CoT prompting (Wei et al., 2022) has become the dominant paradigm for eliciting complex reasoning in large language models, and is increasingly deployed as a verification signal in high-stakes settings (Turpin et al., 2023; Lightman et al., 2023): if a model can show its work, that work can be checked. This assumption rests on the implicit premise that faithfulness is stable across tasks and models. We show it does not, and that the failure is structured rather than random. The faithfulness regime is systematically shaped by the interaction between task type and training distribution. The same task produces opposite regimes in different models; the same model shifts regimes across tasks. DeepSeek-R1-Distill-Llama-8B achieves 100% accuracy on a bracket-matching task with a 0% flip rate under step corruption, a failure mode invisible to accuracy evaluation alone. Targeted free-generation experiments confirm this reflects systematic chain bypassing: R1 produces near-identical behavior on clean and corrupted chains alike, never conditioning on the supplied trace regardless of its correctness.

Prior work has characterized CoT faithfulness as a function of the prompt or the task, but not their interaction. Turpin et al. (2023) showed that input-level biases can induce unfaithful CoT, where

[1] Code: https://anonymous.4open.science/r/NLDD_v2-F5DD/

models rely on biasing features in the prompt while omitting them from their explanations. Lanham et al. (2023) found large cross-task variation in CoT conditioning and an inverse scaling effect, but their metric is binary and unnormalized, making graded cross-model comparison impossible. Sprague et al. (2025) showed CoT primarily benefits multi-step arithmetic over fact retrieval. None of these works ask whether faithfulness regime is jointly determined by which model meets which task. We show that these factors interact: task type sets the baseline direction, training distribution determines regime within symbolic tasks, and scale modulates magnitude without changing either, as demonstrated by two DeepSeek models of comparable scale exhibiting sharply different causal sensitivity on the same task.

We measure this with **Normalized Logit Difference Decay** (NLDD), a replacement-based step-level metric that quantifies how much a model's answer confidence drops when a single reasoning step is replaced with a task-appropriate error. Replacement holds forward chain structure constant while varying step content, isolating each step's causal contribution independently of chain-length dependence. Truncation-based faithfulness metrics cannot resolve this confound. Normalization by intrinsic output variability enables consistent cross-model comparison across architectures with different logit scaling, including compressed output distributions. We complement NLDD with chain pruning as a behavioral validator and linear probes as a representational diagnostic. This triangulation is what allows regime classification to rest on convergent evidence rather than a single metric.

The structure is concrete and replicable. Removing early arithmetic reasoning steps collapses accuracy by $\sim$65%; removing early commonsense steps yields a 1.1% crash rate, showing that the chain is causally dispensable despite surface coherence. Linear probes explain why: on commonsense tasks the answer is already encoded at the middle transformer layer (StrategyQA: 64–67% probe accuracy) before the chain has completed; on logical deduction, probe accuracy falls to near chance (48–54%), confirming the chain is doing genuine step-by-step computation.

Our primary contributions are:

- **Faithfulness regime is structured by model-task interaction**: task type determines direction across all five models; training distribution determines regime within symbolic tasks; scale modulates magnitude without changing either.

- **Chain bypassing exists in practice**: DeepSeek-R1 achieves 100% accuracy with a 0% flip rate under step corruption. Paired free-generation experiments (N=500 per condition, 101 input sequences) confirm R1 never conditions on the supplied chain regardless of its correctness: recovery rates are near-identical on clean (76.0%) and corrupted (80.0%) chains, establishing systematic chain bypassing rather than corruption detection or causal inertness.

- **NLDD with replacement-based corruption**: a normalized, causally clean step-level faithfulness metric that enables graded cross-model comparison and disambiguates decorative reasoning from premature sufficiency.

- **Behavioral and representational triangulation**: crash rates under chain pruning and linear probe accuracy jointly validate regime classifications independently of the logit-level signal.

These results have direct consequences for CoT deployment. A model that reasons faithfully on arithmetic may produce a causally disconnected chain on commonsense tasks, and a single faithfulness benchmark cannot distinguish these regimes. The implication is not that CoT is untrustworthy, but that trustworthiness is not a global property of a model. It is a local property of a model-task pair, established by causal intervention rather than accuracy or fluency. Task-specific profiling using NLDD or equivalent causal instruments is a prerequisite, not an optional diagnostic, for deploying CoT as a verification signal in high-stakes settings.

2 Causal Diagnostic Framework

2.1 Benchmarks and Dataset Construction

We evaluate across four benchmarks spanning symbolic to naturalistic reasoning, chosen to maintain verifiable ground-truth answers, a methodological requirement for NLDD, which calibrates against a known correct answer. Following Sprague et al. (2025), we treat a problem as *symbolic* if it can be grounded in a natural, well agreed-upon formal system, and *commonsense* if no such framework

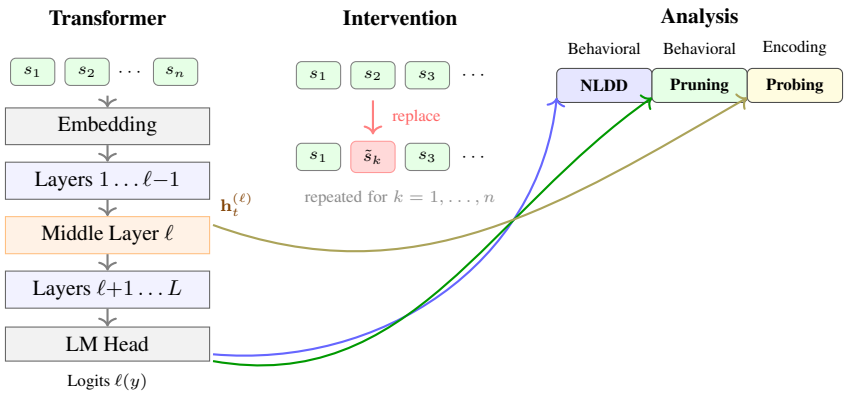


Figure 2: Methodological framework. A task-appropriate error replaces step k; all other steps are retained unchanged. Repeating across all k produces a per-step causal profile. NLDD measures answer-margin drop from final-token logits or teacher-forced sequence scores, depending on the task; Pruning measures crash rate under step removal; Probing tests middle-layer encoding independently of causal routing.

applies. Dyck-n (syntactic state tracking) is synthetically generated: bracket sequences of varying depth and length grounded in a formal bracket grammar with deterministic ground-truth answers. PrOntoQA (Saparov and He, 2023) provides multi-hop logical entailment grounded in first-order logic; we use a fictional long-chain variant (4–16 hops) generated following their ontology to avoid memorization effects. GSM8K (Cobbe et al., 2021) provides arithmetic reasoning grounded in formal mathematical operations with step-by-step solutions. StrategyQA (Geva et al., 2021) provides implicit multi-hop commonsense reasoning with binary yes/no answers, drawn from the official public split. We use these three categories: symbolic (Dyck-n, PrOntoQA), arithmetic (GSM8K) and commonsense (StrategyQA), descriptively to organize our findings rather than as a formal taxonomy.

We condition on samples where the model produces the correct final answer on the clean chain, isolating faithfulness measurement from task failure. We target $N{=}300$ per task per model; accepted counts vary by model due to clean accuracy filtering (Appendix B). StrategyQA sample counts are bounded by the official public split (229 questions with supporting-fact decompositions); after clean accuracy filtering the maximum accepted sample count across all models is $N{=}169$. Evaluation and probing subsets are disjoint to prevent representational leakage.

For each clean sample we generate a counterfactual variant by replacing the content of a single reasoning step with a task-appropriate error. Steps are newline-delimited units of the reasoning trace (e.g., one bracket-state update, one deduction sentence, one arithmetic line; see Table 3): bracket depth errors (Dyck-n), entity substitutions (PrOntoQA), arithmetic errors (GSM8K), and truth-value inversions of a key supporting proposition (StrategyQA). Critically, all steps *after* the corrupted step are retained unchanged, isolating the causal contribution of that step's content from any forward chain dependency. Because retaining subsequent clean steps creates a potential distributional shift, a logical discontinuity between the corrupted step and the clean tail, we filter candidates using a perplexity ratio threshold (≤ 1.5 for GSM8K, ≤ 3.5 for others) computed using the evaluated model itself. This ensures the spliced chain remains within the model's native distribution and that NLDD reflects genuine reasoning sensitivity rather than out-of-distribution surprise. We additionally generate semantic-preserving paraphrases as controls; near-zero NLDD on paraphrases confirms that score variation is driven by semantic errors rather than arbitrary surface perturbations (Appendix C).

2.2 Models

We evaluate five instruction-tuned decoder-only transformers. Llama-3.1-8B and Llama-3.1-70B (Grattafiori et al., 2024) provide a scale-controlled comparison within a single architecture family. Gemma-2-9B (Team et al., 2024) provides an architectural variant with logit soft-capping, testing whether NLDD normalization is robust to compressed output distributions. DeepSeek-Coder-6.7B (Guo et al., 2024) is a code-specialized model whose training distribution diverges substantially from natural language reasoning tasks. DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) is a reasoning-specialized model distilled onto a Llama-3.1-8B base and provides a comparison with

Llama-3.1-8B on the effect of reasoning distillation and post-training.

2.3 NLDD: Measuring Step-Level Faithfulness

NLDD measures how much a model's confidence in its correct answer degrades when a single reasoning step is replaced with an erroneous one. Unlike binary accuracy metrics, it captures graded sensitivity in logit space and normalizes for cross-model comparison.

Calibration. We calibrate a normalization constant S separately for each model-task run on clean reasoning traces. For token-margin tasks (Dyck-n, PrOntoQA, StrategyQA), we use

$$S = \frac{1}{M} \sum_{m=1}^{M} \sigma(z_m), \tag{1}$$

where z_m is the final-token logit vector for sample m and $\sigma(\cdot)$ computes its standard deviation across the full vocabulary. Normalizing by intrinsic output variability rather than absolute logit magnitude makes NLDD comparable across architectures with different output scaling, including Gemma-2's logit soft-capping mechanism, which compresses the logit dynamic range relative to standard softmax models.

For GSM8K, which uses sequence-margin scoring over numeric answer strings, S is instead the mean standard deviation of clean candidate-sequence scores. Thus, both scoring and calibration are matched to the task's answer format.

Answer margin. For token-margin tasks, we measure the standardized margin between task-specific correct and incorrect answer candidates. Let $\ell(y)$ denote the pre-softmax logit assigned to token y by the language model head at the final sequence position:

$$LD_{\text{tok}} = \frac{\max_{y \in \mathcal{Y}_c} \ell(y) - \max_{y' \in \mathcal{Y}_i} \ell(y')}{S_{\text{tok}}}, \tag{2}$$

where $\mathcal{Y}_c$ and $\mathcal{Y}_i$ are validated token IDs obtained from task-specific positive and negative answer candidates, including tokenization variants such as leading spaces. They are not the full vocabulary complement. For example, Dyck-n compares the correct closing bracket against the other closing brackets, while binary tasks compare the correct label against the opposite label. Negative candidates are task-determined and closed-set rather than open-vocabulary, bounding sensitivity to distractor choice by task structure.

For GSM8K, whose answers are numeric strings, we use a teacher-forced sequence-margin analogue rather than first-token scoring:

$$LD_{\text{seq}} = \frac{\max_{y \in \mathcal{Y}_c} s(y) - \max_{y' \in \mathcal{Y}_i} s(y')}{S_{\text{seq}}}, \tag{3}$$

where $s(y)$ is the mean teacher-forced log-probability per token of candidate answer string y. The incorrect set $\mathcal{Y}_i$ is a finite set of numeric distractors generated from the gold answer and, when available, the model's greedy numeric prediction.

Faithfulness quantification.

$$\text{NLDD} = \frac{LD_{\text{clean}} - LD_{\text{corrupt}}}{|LD_{\text{clean}}|} \times 100. \tag{4}$$

Rows with $|LD_{\text{clean}}| \leq \epsilon$ are retained but assigned NLDD $= 0$; in our experiments, $\epsilon = 0.05$. To ensure robustness against outlier score inflation from near-zero baseline margins, aggregate results in Section 3 are reported using 10% trimmed means and medians alongside 95% BCa bootstrap confidence intervals (B=10,000).

Causal grounding. Replacement-based corruption implements a do-calculus intervention (Pearl, 2009): setting step k to an erroneous value while holding all other steps fixed isolates that step's causal contribution independently of chain-length dependence. This design formally separates two distinct failure modes, decorative reasoning, where no step produces a causal signal under intervention, and premature sufficiency, where early steps are causal but late steps are inert, which truncation-based metrics cannot distinguish (Appendix E).

Faithfulness regimes. NLDD values identify three qualitatively distinct regimes that organize our results throughout. **Faithful**: consistently positive NLDD, the chain is causally load-bearing and corrupting a step reduces answer confidence. **Depth-degrading**: NLDD declines with chain depth, causal utility decreases as the chain lengthens, consistent with parametric retrieval preceding chain computation. **Anti-faithful**: negative mean NLDD, corruption increases correct-answer confidence, suggesting the clean CoT suppresses a pre-computed answer. We observe a trend toward this pattern in one model-task combination (§3); we include it here for completeness as it represents a theoretically distinct causal structure. The present

replacement-based metric provides cleaner causal isolation by holding the chain structure constant across clean and corrupt conditions.

2.4 Chain Pruning: Testing Causal Necessity

Replacement isolates whether step *content* is causally load-bearing; pruning tests whether subsequent steps are necessary at all. The two interventions triangulate faithfulness from independent directions and are distinct in purpose from truncation-based faithfulness metrics (Lanham et al., 2023): here truncation serves as behavioral validation of NLDD's logit-level signal, not as a primary faithfulness measure.

Design. We supply a clean prefix through step k and prompt the model to answer from the prefix alone, omitting all subsequent steps. We evaluate at three data-driven horizons—k_{low}, k_{point}, k_{high}—corresponding to the lower bound, mode, and upper bound of the 95% BCa confidence interval around the reasoning horizon k^*. We record two quantities: **Crash Rate**, the fraction of previously correct samples that fail under truncation; and **Token Save %**, the fraction of reasoning tokens eliminated. Horizon values per task are reported in Table 5.

2.5 Linear Probing

Linear probes test whether task-relevant information is linearly decodable from hidden states independently of whether the model uses it for prediction (Alain and Bengio, 2018). We train L2-regularized logistic regression classifiers (C=1.0, 80/20 split, seed 42) on step-terminal hidden states from the middle transformer layer ($\lfloor L/2 \rfloor$). The middle layer serves as a practical heuristic balancing early-layer token processing and late-layer task-specific computation; we do not claim it is uniquely optimal across all settings.

We probe for the following targets:

- **Dyck-n:** Stack depth at each step (0–10, multi-class)
- **PrOntoQA:** Final truth value at each step (binary)
- **GSM8K:** Arithmetic operation type at each step (multi-class)
- **StrategyQA:** Final answer choice at each step (binary)

3 Results

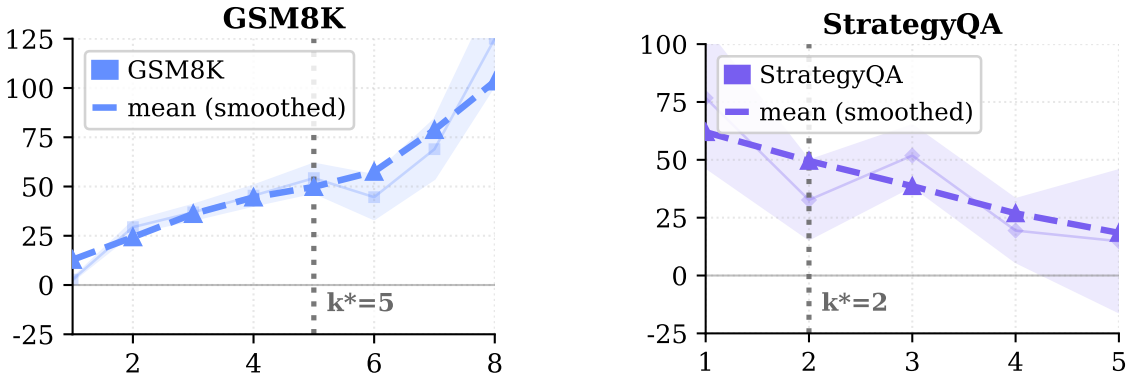
We organize results around three explanatory factors, validated through two independent diagnostic signals (Table 1). Task type determines faithfulness direction, training distribution determines regime within symbolic tasks, and scale modulates magnitude without changing either. Crash rates (§3.3, Appendix F) and linear probes (§3.4, Appendix H) independently corroborate each regime classification; flip rates, corrupt accuracy, trimmed means, and full robustness statistics are in Appendix D.

3.1 Task Type Predicts Faithfulness Direction

Arithmetic reasoning is load-bearing across all five models. GSM8K yields positive NLDD for every model (range 9.29–27.77), with four of five reaching significance after Holm correction. Llama-70B replicates the pattern at scale (NLDD 15.71, Δ=+34.0, p=0.073), ruling out a mid-sized-model artifact. Crash rates at k_{low} confirm the logit signal behaviorally: removing early chain steps collapses accuracy by ∼65%, demonstrating that the chain is causally necessary for answer production, not merely decorative.

Commonsense reasoning shows a directionally consistent negative depth trend across all five models, significant in two of five. All five models show a negative robustness delta on StrategyQA (Δ range: -44.1 to -79.2), NLDD decreases as corruption position moves deeper into the chain. Two of five reach significance after Holm correction (Gemma p=0.035, Llama-70B p=0.0003); the remaining three trend in the same direction. Crash rates near-zero at k_{high} (4%) confirm low causal necessity regardless of depth. NLDD means vary substantially across models (-16.69 to 41.01): the depth-degradation direction is universal, but absolute faithfulness magnitude is model-dependent. Linear probes corroborate the mechanism for Gemma and Llama-8B: StrategyQA answers are encoded at the middle transformer layer (64–67% probe accuracy), consistent with parametric retrieval preceding chain computation.

Symbolic tasks expose model-level divergence. The directional consistency of GSM8K and StrategyQA does not extend to Dyck-n or PrOntoQA, where faithfulness regime depends on which model is evaluated rather than on task type alone. This is not a failure of the framework; it is the finding. Within symbolic tasks, regime is determined by training distribution, as we show in Section 3.2.



(a) Llama-3.1-8B on GSM8K (arithmetic). NLDD increases monotonically with chain depth, the chain is causally load-bearing throughout.

(b) Llama-3.1-8B on StrategyQA (commonsense). NLDD declines monotonically with chain depth, the chain is causally dispensable throughout.

Figure 3: NLDD as a function of corruption step k for arithmetic vs. commonsense reasoning on Llama-3.1-8B. Arithmetic chains gain causal weight with depth; commonsense chains lose it. Both patterns hold across all five models (Table 4); three models including Gemma-2-9B cross into negative NLDD on StrategyQA at later positions. Shaded regions show 95% BCa bootstrap intervals; dashed vertical line marks k^*.

3.2 Training Distribution Predicts Regime Within Symbolic Tasks

Same task, sharply different causal sensitivity: the DeepSeek comparison. DeepSeek-Coder-6.7B and DeepSeek-R1-Distill-Llama-8B models are comparable in scale from the same developer, but they are not architecture-controlled: they differ in tokenizer, base model, and post-training procedure. On PrOntoQA, both exhibit positive NLDD, yet differ sharply in magnitude: Coder achieves NLDD 102.14 with a 56.6% flip rate (n=65), while R1 achieves NLDD 25.22 with a 7.4% flip rate (n=189)—a 4× difference in NLDD magnitude and a 7× difference in flip rate on identical inputs. Coder's low n reflects its code-specialized training distribution: it failed clean accuracy on the majority of PrOntoQA candidates, consistent with multi-hop logical deduction lying far outside its training domain; these results should be interpreted with appropriate caution.

The limiting case: systematic chain bypassing. DeepSeek-R1 on Dyck-n yields NLDD 0.43 (trimmed mean 0.31, 95% BCa CI $[0.20, 0.64]$), 100% corrupt accuracy, and a 0% flip rate—the chain is completely unresponsive to step corruption under closed-set logit scoring. Paired free-generation experiments (N=500 per condition, 101 input sequences) reveal why: recovery rates are near-identical on corrupted (80.0%) and clean (76.0%) chains, ruling out corruption detection as an alternative explanation and confirming bypassing rather than targeted correction. Accuracy of 51.4% and 55.4% respectively, both well above the 25% chance baseline, confirms R1 re-solves from the raw input rather than failing randomly (Table 6). R1 exhibits *systematic chain bypassing*: it never conditions on the supplied trace regardless of correctness, re-solving directly from the raw input on every trial. By contrast, Llama-3.1-8B achieves NLDD 50.31 with a 16.8% flip rate on the same task, confirming that reasoning distillation substantially alters causal coupling even when the base model is held fixed.

Scale modulates magnitude; task and training distribution determine regime. Scaling Llama from 8B to 70B reduces faithfulness magnitude on both GSM8K (NLDD: 27.7→15.7) and PrOntoQA (NLDD: 186.9→91.8), consistent with the inverse scaling effect Lanham et al. (2023) observed and extending it to a normalized cross-model metric. Yet the task-type distinction persists at both scales: arithmetic and logical deduction remain load-bearing, commonsense remains depth-degrading. Task type sets the direction, training distribution determines the specific regime within symbolic tasks, and scale compresses magnitude without changing either.

3.3 Crash Rates Behaviorally Validate Regime Classification

Three pruning signatures emerge, summarized in Table 5. **Faithful regime** (GSM8K, Dyck-n): high crash rates at k_{low} confirm that early steps are causally necessary, meaning the model cannot skip computation. **Faithful with early saturation**

Table 1: Mean NLDD (95% BCa CI) across five models and four benchmarks. **Bold** = significant positive robustness ($p<0.05$, Holm). *Italic* = significant negative robustness Δ (mean CI may cross zero). † small n, interpret with caution. ‡ CI unavailable (log-scale only). Full results: Appendix D.

	DS-Coder	DS-R1	Gemma	Llama-8B	Llama-70B
Dyck-n	8.6 [3.8, 14.3]	0.4 [0.2, 0.6]	**27.0** [18.1, 32.2]	50.3 [41.2, 62.1]	23.7‡
GSM8K	**22.5** [18.4, 33.0]	**22.3** [19.0, 26.1]	**9.3** [8.0, 10.8]	**27.8** [25.0, 30.9]	15.7 [14.2, 17.4]
PrOntoQA	†102.1 [72.6, 128.2]	25.2 [21.7, 30.1]	**141.3** [112.7, 165.5]	186.9 [142.9, 241.5]	**91.8** [70.3, 119.1]
StrategyQA	26.0 [3.2, 56.5]	6.2 [-9.2, 20.7]	*−16.7* [−78.6, 7.3]	41.0 [25.1, 60.9]	*37.4* [23.8, 48.9]

(PrOntoQA): near-zero crash rates despite high NLDD reflect that k^* falls at the final chain step; the chain is load-bearing but reaches sufficiency before completion, confirmed by flip rates of 52–79%. **Depth-degrading** (StrategyQA): near-zero crash rates confirm causal dispensability, consistent with the negative robustness deltas in Table 1.

3.4 Representational Encoding Dissociates from Causal Role

Probe results (Table 7) reveal a consistent dissociation between representational encoding and causal routing. On PrOntoQA, probe accuracy falls to near chance for all three models (50–55%) despite high NLDD and flip rates. This is not contradictory: near-chance probing means the final answer is not yet linearly decodable at the middle layer, confirming that the chain is doing genuine step-by-step computation that has not resolved to a linearly accessible answer midway through. The chain is load-bearing precisely because the answer is not pre-stored.

On StrategyQA, the pattern reverses: Gemma and Llama show above-chance probe accuracy (64–67%) consistent with parametric retrieval, corroborating the behavioral finding that the chain is causally dispensable. For Gemma, this coexists with a negative mean NLDD (-16.69), suggesting the pre-stored answer is actively suppressed by the chain rather than merely ignored.

On Dyck-n, Gemma achieves 89.6% probe accuracy (bracket-depth state strongly encoded) alongside moderate NLDD (26.95). The structural state is represented but only partially routed to the output, an instance of *representational silencing*: the information exists in the model but does not causally drive prediction.

4 Discussion

4.1 Interpreting the Structured Faithfulness Account

Symbolic tasks. Symbolic tasks expose a task-selective form of bypassing. DeepSeek-R1 achieves near-perfect accuracy with NLDD 0.43 and a 0% flip rate; free-generation experiments confirm systematic chain bypassing (§3.2, Appendix G). Yet bypassing is not a global property of R1: the same model exhibits positive NLDD on both GSM8K (22.34, flip rate 10.2%) and PrOntoQA (25.22, flip rate 7.4%), confirming that the chain remains load-bearing on tasks requiring genuine sequential computation. Dyck-n bypassing is consistent with R1's reasoning distillation having collapsed bracket-state tracking, a task whose answer is fully determined by syntactic structure, into parametric retrieval. The Llama-8B vs. DeepSeek-R1-Distill-Llama-8B comparison, which shares a base model, confirms that reasoning distillation specifically alters causal coupling independently of architecture.

Anti-faithful depth pattern. The StrategyQA depth pattern also suggests a more specific failure mode. Gemma, DeepSeek-Coder, and DeepSeek-R1 cross into negative mean NLDD at later chain positions, indicating that later corruptions can increase confidence in the correct answer rather than decrease it. One possible account is that, once answer information is already available, additional reasoning tokens introduce competing associations that interfere with the answer margin. The Llama family does not show the same negative crossing, suggesting that the effect may depend on post-training or instruction-following behavior. This interpretation is an hypothesis; targeted ablations are needed to identify the mechanism.

4.2 Implications for CoT as a Verification Signal

These results argue against treating CoT faithfulness as a property that transfers automatically across tasks or models. An evaluation showing faithful arithmetic reasoning does not imply faithful commonsense reasoning, and a model that uses the chain on one symbolic task may ignore it on another. Faithfulness should therefore be established for the target model-task pair, not inferred from overall accuracy or from the fluency of the trace.

This matters for systems that use CoT as a verification signal. A reasoning trace can be coherent, accurate, and still weakly connected to the answer it accompanies. In the most extreme case we observe, DeepSeek-R1 on Dyck-n preserves the correct answer under step corruption with a 0% flip rate via systematic chain bypassing, R1 re-solves from the raw input regardless of whether the chain is correct or corrupted. Such a trace may still be useful as a communication artifact, but it cannot be treated as a causal transcript of the computation that produced the answer.

The same concern applies to process supervision. Process reward models (PRMs) and step-wise evaluators typically score the visible reasoning trace. If that trace is not causally coupled to the answer, optimizing its surface quality may improve the explanation without improving the underlying computation. Our results do not show that this failure occurs in every supervised reasoning system, but they identify a necessary diagnostic: causal faithfulness must be established per model-task pair before process supervision can be trusted to improve reasoning rather than surface coherence alone.

5 Related Work

The Functional and Unfaithful Roles of CoT. Recent work distinguishes when CoT performs genuine computation versus when it acts as a stylistic artifact. Sprague et al. (2025) showed CoT primarily benefits multi-step math over simple fact retrieval; our contrast between arithmetic and commonsense tasks causally grounds this distinction. Furthermore, Pfau et al. (2024) show that transformers can, in principle, exploit intermediate tokens purely as a compute buffer independent of token content. Our layer probes suggest this possibility may be active for commonsense tasks. This architectural decoupling enables severe unfaithfulness: Arcuschin et al. (2025) demonstrate it occurs naturally on unbiased prompts via *implicit post-hoc rationalization*. Even in explicit "thinking" models (e.g., DeepSeek-R1, Sonnet 3.7), a stark thinking-answer divergence persists, with internal computation routinely hiding true reasoning steps from the final text (Young, 2026). Together, these findings motivate causal diagnosis of CoT faithfulness rather than surface observation alone.

The Evolution of Process Supervision and Agents. To address outcome vulnerabilities, alignment has shifted toward fine-grained PRMs across text, RL, and autonomous agents. This paradigm is moving toward active test-time guidance, such as Process Reward Agents (Sohn et al., 2026) that use online, step-wise rewards to prune reasoning trajectories in domains like medicine. However, these pipelines assume the supervised chain causally drives the answer. Our results challenge this assumption directly.

Causal Measurement Frameworks. Paul et al. (2024) used causal mediation analysis to show CoT's causal effects are task-dependent, but without a normalized metric, cross-model comparisons remain confounded by scale and output distribution. Our normalization adapts the cross-architecture principles of Normalized AOPC (Edin et al., 2025) from feature importance to step-level causal measurement. While white-box mechanistic interpretability (Geiger et al., 2021; Wang et al., 2022) offers microscopic circuit resolution, our chain-level intervention trades that resolution for the architectural breadth needed to profile diverse deployments at scale. The two approaches are therefore complementary rather than competing.

6 Conclusion

CoT faithfulness regime is systematically shaped by the interaction between task type and training distribution, not either factor alone. Arithmetic reasoning is load-bearing across all five models; commonsense reasoning depth-degrades across all five models; symbolic tasks diverge by training distribution rather than task structure. Scale modulates magnitude without changing either.

Faithfulness is not a global property of a model. It is a local property of a model-task pair, established by causal intervention rather than inferred from accuracy, fluency, or task success alone.

7 Limitations

Our replacement intervention improves causal isolation, but introduces a splicing assumption. Retaining later clean steps holds chain length and downstream structure fixed, but can create local incoherence between the corrupted step and the clean continuation. Perplexity filtering, token constraints, task validity checks, and paraphrase controls mitigate this, but cannot guarantee fully natural counterfactuals.

While we select diverse benchmarks spanning arithmetic, symbolic, and commonsense reasoning, we rely on one primary dataset per category (e.g., GSM8K, StrategyQA). While our behavioral and representational diagnostics strongly triangulate our findings on these datasets, generalizing these regime classifications to all arithmetic or commonsense tasks is a hypothesis that requires future validation across a wider corpus of benchmarks.

The 0% flip rate for DeepSeek-R1 on Dyck-n reflects systematic chain bypassing, confirmed by paired free-generation experiments showing near-identical behavior on clean and corrupted chains. The internal mechanism by which R1 bypasses the chain and routes computation directly to the raw input remains uncharacterized at the circuit level; mechanistic ablation is an important direction for future work.

We evaluate four reasoning benchmarks and a finite set of instruction-tuned decoder-only models, including one 70B model. Generalization to proprietary frontier systems, encoder-decoder or multimodal architectures, tool-using agents, and open-ended reasoning remains untested. StrategyQA is also bound by the official public split and usable decomposition annotations.

NLDD normalization remains empirically motivated. We normalize answer margins by model-level output variability, which supports comparison across different logit scales, including compressed output distributions. However, this does not guarantee invariance under all architectural logit transformations or post-training procedures. Sensitivity analysis across alternative normalization schemes remains an important direction for future work.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

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. *Preprint*, arXiv:2503.08679.

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. *Preprint*, arXiv:2110.14168.

Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. 2025. Normalized aopc: Fixing misleading faithfulness metrics for feature attribution explainability. *Preprint*, arXiv:2408.08137.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Preprint*, arXiv:2106.02997.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *Preprint*, arXiv:2401.14196.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Jiangkai Zhao, Niranjan DasSarma, Sandipan Nair, Nicholas Schiefer, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Preprint*, arXiv:2402.13950.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *Preprint*, arXiv:2404.15758.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *Preprint*, arXiv:2210.01240.

Jiwoong Sohn, Tomasz Sternal, Kenneth Styppa, Torsten Hoefler, and Michael Moor. 2026. Process reward agents for steering knowledge-intensive reasoning. *Preprint*, arXiv:2604.09482.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Preprint*, arXiv:2211.00593.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Maeda, Polina Edakunni, Henryk Ku, Ed Chi, Sharan Narang, Gaurav Nayak, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Richard J. Young. 2026. Why models know but don't say: Chain-of-thought faithfulness divergence between thinking tokens and answers in open-weight reasoning models. *Preprint*, arXiv:2603.26410.

Dataset	Original Step (Clean)	Counterfactual Step (Corrupted)
Dyck-n	"Seen '{', stack depth is 2."	"Seen '{', stack depth is **1**."
PrOntoQA	"Every wumpus is a zumpus."	"**No** wumpus is a zumpus."
GSM8K	"She has $15 \times 2 = 30$ eggs."	"She has $15 \times 2 = \mathbf{32}$ eggs."
StrategyQA	"Marie Curie was born in Poland."	"Marie Curie was born in **France**."

Table 2: Representative counterfactual interventions used for NLDD across all four benchmarks. Corruptions (bolded for clarity) are designed to break the causal chain while maintaining linguistic coherence. StrategyQA corruptions target factual premises underlying the multi-hop decomposition rather than arithmetic or structural state, reflecting the commonsense nature of the task.

A Qualitative Examples and Implementation Details

We provide qualitative examples, dataset formats, and implementation details that complement the main text.

A.1 Example Counterfactuals

Table 2 shows representative single-step counterfactual interventions for each dataset. In all cases, the counterfactual modifies exactly one intermediate reasoning step while preserving the original input structure and prompt format. These interventions are designed to induce a causal change in the expected answer.

A.2 Dataset Examples

Table 3 provides representative examples of inputs and corresponding step-wise reasoning traces for each dataset. All tasks are formatted to expose explicit intermediate reasoning steps, enabling localized counterfactual interventions and step-level mechanistic analysis. StrategyQA traces follow a multi-hop decomposition structure in which each step resolves an implicit sub-question, distinguishing it from the arithmetic and symbolic tasks.

A.3 Implementation Details

All experiments were conducted on a single NVIDIA GH200 480GB GPU, with the exception of Llama-3.1-70B, which was run on 8×NVIDIA A100 GPUs due to memory requirements. Models were loaded in `bfloat16` with automatic de-

Task	Input / Question	Reasoning Trace (Excerpt)
Dyck-n	**Input:** [[{⟨⟩}]] **Question:** What is the next closing bracket?	Seen [, depth 1 Seen {, depth 2 Seen [, depth 3 Seen {, depth 4 Seen <, depth 5 Seen >, depth 4 Seen }, depth 3 **Answer:**]
PrOntoQA	**Facts:** Sam is a zumpus. **Rules:** All zumpus are impus. All impus are rompus. All rompus are gorpus. **Question:** Is Sam a wumpus?	Since Sam is a zumpus and all zumpus are impus, Sam is an impus. Since Sam is an impus and all impus are rompus, Sam is a rompus. Since Sam is a rompus and all rompus are gorpus, Sam is a gorpus. Conclusion: Sam is a gorpus, not a wumpus. **Answer:** False
GSM8K	Janet's ducks lay 16 eggs per day. She eats 3, bakes 4, and sells the rest at $2 each. **Question:** How much does she make daily at the farmer's market?	Janet sells $16 - 3 - 4 = 9$ duck eggs per day. She makes $9 \times 2 = 18$ dollars. **Answer:** $18
StrategyQA	**Question:** Did Marie Curie ever visit the country where penicillin was discovered?	Marie Curie was a Polish-French scientist who spent most of her career in France. Penicillin was discovered by Alexander Fleming in Scotland. Marie Curie traveled widely across Europe, including visits to the United Kingdom. **Answer:** Yes

Table 3: Representative examples from each dataset. All tasks are formatted to expose explicit intermediate reasoning steps, enabling localized counterfactual interventions. Counterfactual variants corrupt a single intermediate step while preserving input structure and surface coherence. StrategyQA reasoning traces reflect implicit sub-question decomposition over factual knowledge rather than symbolic or arithmetic state tracking.

vice mapping and scaled dot-product attention (`sdpa`) via HuggingFace `transformers`. All models were evaluated using greedy decoding (temperature $= 0$, `do_sample=False`). Key software versions: Python 3.10.12, PyTorch 2.7.0, Transformers 5.8.1, Datasets 4.8.5, scikit-learn 0.23.2, SciPy 1.8.0, and statsmodels 0.14.6. All statistical tests (BCa confidence intervals, Holm correction) were computed post-hoc in NumPy 1.26.4 and SciPy. Linear probes were trained using scikit-learn's `LogisticRegression` with ℓ_2 regularization and default solver settings. No external APIs or proprietary inference services were used; all model weights were loaded locally from HuggingFace Hub checkpoints.

B Dataset Audit

Table 4 reports accepted sample counts per model per task after clean accuracy filtering. Counts below N=300 reflect models that failed to produce correct clean-chain answers on a sufficient number of candidates. Deepseek-Coder-6.7B on PrOntoQA (n=65) is the most affected, consistent with multi-hop logical deduction lying outside its training distribution. Deepseek-R1-8B on GSM8K (n=279) reflects minor filtering. All StrategyQA counts are bounded by the official public split after decomposition filtering (≤ 169).

C Paraphrase Control Results

Semantic-preserving paraphrases of each reasoning step were generated as controls. Mean NLDD on paraphrased steps was near zero across all tasks and models (range: -1.2 to $+1.8$), confirming that score variation in the main results is driven by semantic errors rather than surface perturbations. This rules out the alternative explanation that NLDD is measuring out-of-distribution surprise from unfamiliar phrasing rather than causal sensitivity to reasoning content.

D Full Results

Table 4 reports complete NLDD statistics including mean, trimmed mean (10%), 95% BCa confidence intervals, corrupt accuracy, flip rate, sample size,

and Holm-corrected robustness deltas for all five models across four benchmarks.

E Causal Sufficiency Frontier: Formal Grounding

Replacement-based NLDD is grounded in three Pearl do-calculus conditions (Pearl, 2009) that formally separate decorative reasoning from premature sufficiency. Each condition instantiates a named concept from Pearl (Ch. 1.3).

$\mathcal{C}_{\text{act}}$ **— Causal Activity.**

$$\exists\, i \in [1, k],\ \exists\, t \in \{\text{replace}, \text{delete}\}\ \text{ s.t.}$$
$$\hat{y}(\text{do}_t(\mathbf{s}_i)) \neq \hat{y}(\mathbf{s}_{1:K})\ \vee\ \Delta_t(i) \geq \delta \quad (5)$$

At least one pre-frontier step causally influences the answer under intervention. Chains failing this condition globally are *decorative* because no step produces a causal signal regardless of position. This implements Pearl's definition of causal influence: step i causally influences y if there exists an intervention $\text{do}(\mathbf{s}_i \leftarrow \mathbf{s}'_i)$ that changes $P(y)$.

$\mathcal{C}_{\text{inert}}$ **— Intervention Inertia.**

$$\forall j \in \mathcal{W}_k,\ \hat{y}(\text{do}(\mathbf{s}_j \leftarrow \tilde{\mathbf{s}}_j)) = \hat{y}(\mathbf{s}_{1:K}) \wedge \Delta(j) \leq \varepsilon \quad (6)$$

All post-frontier steps leave the answer unchanged under replacement. Chains where this fails exhibit *premature sufficiency*, where early steps are causal but late steps remain load-bearing. This implements Pearl's definition of no causal influence: $P(y \mid \text{do}(\mathbf{s}_j \leftarrow \mathbf{s}'_j))$ is invariant to the choice of $\mathbf{s}'_j$.

$\mathcal{C}_{\text{suf}}$ **— Distributional Sufficiency.**

$$\hat{y}(\mathbf{s}_{1:k}) = \hat{y}(\mathbf{s}_{1:K})\ \wedge\ \Delta_{\text{prefix}}(k) \leq \varepsilon \quad (7)$$

The prefix alone produces the same answer as the full chain within task-native tolerance ε. This implements Pearl's truncated factorization: $P(y \mid \text{do}(\mathbf{s}_{1:k})) \approx P(y \mid \text{do}(\mathbf{s}_{1:K}))$.

Relationship to NLDD. NLDD operationalizes $\mathcal{C}_{\text{act}}$ and $\mathcal{C}_{\text{inert}}$ continuously across all steps, producing a per-step causal profile rather than a binary frontier detection. The regime taxonomy in Section 3, faithful, depth-degrading, and anti-faithful, corresponds to the aggregate signature of these conditions across the full chain. Full joint CSF computation requiring all three conditions simultaneously is left for future work with larger sample pools.

F Chain Pruning Results

Table 5 reports full chain pruning results for DeepSeek-R1, DeepSeek-Coder, and Llama-3.1-70B across all evaluated tasks and horizons. Llama-3.1-70B is absent from Dyck-n pruning due to compute constraints.

G DeepSeek-R1 Chain Bypassing: Free-Generation Diagnostic

Table 6 reports full results from paired free-generation experiments diagnosing the DeepSeek-R1 Dyck-n finding. Samples were drawn from the same 101 input sequences used in the main evaluation, stratified at up to five corrupted steps per sequence. The clean condition used identical inputs with fully correct reasoning chains; the corrupted condition spliced a single bracket-depth error at the same position. Recovery rate measures the fraction of free generations containing explicit self-correction tokens ("wait", "hmm"). Accuracy measures the fraction correctly identifying the next closing bracket under free generation.

H Linear Probing Results

Table 7 reports linear probe accuracy at the middle transformer layer for DeepSeek-Coder, Llama-3.1-8B, and Gemma-2-9B. DeepSeek-R1 and Llama-3.1-70B are excluded due to compute constraints. StrategyQA probing is available for Gemma and Llama-8B only; DeepSeek-Coder StrategyQA probing was not completed.

Table 4: Full results: mean NLDD, trimmed mean (10%), 95% BCa CI, corrupt accuracy, flip rate, and Holm-corrected robustness delta across all five models and four benchmarks. **Bold** = significant positive robustness ($p<0.05$, Holm). *Italic* = significant negative robustness. † small n, interpret with caution.

Task	Model	N	Mean NLDD	Trim. (10%)	95% BCa CI	Corr. acc	Flip rate	Robust. Δ (Holm)
Dyck-n	DS-Coder-6.7B	300	8.60	3.21	[3.77, 14.26]	96.4%	3.6%	—
	DS-R1-8B	300	0.43	0.31	[0.20, 0.64]	100%	0.0%	$+2.9$ ns
	Gemma-2-9B	300	26.95	17.04	[18.05, 32.21]	87.0%	13.0%	**$+25.9$, p=0.027**
	Llama-3.1-8B	300	50.31	23.87	[41.18, 62.09]	83.2%	16.8%	$+10.6$ ns
	Llama-3.1-70B	300	23.67	—	no CI	86.2%	13.8%	no horizon
GSM8K	DS-Coder-6.7B	300	22.46	9.74	[18.39, 33.01]	86.0%	14.0%	**$+28.3$, p=0.0009**
	DS-R1-8B	279	22.34	12.23	[19.01, 26.07]	89.8%	10.2%	**$+16.6$, p=0.0009**
	Gemma-2-9B	300	9.29	4.06	[8.03, 10.79]	98.0%	2.0%	**$+28.9$, p=0.0006**
	Llama-3.1-8B	300	27.77	18.70	[24.99, 30.92]	89.8%	10.2%	**$+51.5$, p=0.0004**
	Llama-3.1-70B	300	15.71	11.45	[14.19, 17.36]	74.8%	25.2%	$+34.0$, p=0.073
PrOntoQA	DS-Coder-6.7B†	65	102.14	25.00	[72.63, 128.21]	43.4%	56.6%	—
	DS-R1-8B	189	25.22	22.67	[21.66, 30.08]	92.6%	7.4%	—
	Gemma-2-9B	300	141.31	34.35	[112.74, 165.53]	21.3%	78.7%	**$+110.8$, $p<0.001$**
	Llama-3.1-8B	300	186.85	107.78	[142.85, 241.48]	47.7%	52.3%	$+49.9$, p=1.0
	Llama-3.1-70B	300	91.79	83.11	[70.28, 119.14]	45.5%	54.5%	**$+50.3$, p=0.006**
StrategyQA	DS-Coder-6.7B	83	25.96	17.77	[3.21, 56.49]	85.5%	14.5%	-79.2 ns
	DS-R1-8B	83	6.16	5.03	[-9.20, 20.66]	96.6%	3.4%	-79.2 ns
	Gemma-2-9B	169	-16.69	4.16	[-78.56, 7.29]	95.4%	4.6%	-66.1, p=0.035
	Llama-3.1-8B	169	41.01	30.46	[25.06, 60.85]	83.0%	17.0%	-44.1 ns
	Llama-3.1-70B	169	37.36	38.95	[23.75, 48.93]	89.3%	10.7%	-64.1, p=0.0003

Table 5: Full chain pruning results. Crash Rate = fraction of previously correct samples failing under prefix-only answering at horizon k. Token Save = fraction of reasoning tokens eliminated. k^* = reasoning horizon (step of peak mean NLDD).

Task	Model	k^*	k_{low} crash	k_{high} crash	Token save
Dyck-n	DS-R1	10	0.0%	0.0%	2.2%
	DS-Coder	7	9.3%	0.7%	10.5%
GSM8K	DS-R1	4	67.7%	0.7%	4.1%
	DS-Coder	6	63.7%	2.3%	0.3%
	Llama-70B	8	0.0%	0.0%	0.07%
PrOntoQA	DS-Coder	13	23.5%	5.9%	1.3%
	Llama-70B	11	17.4%	1.0%	4.3%
StrategyQA	DS-R1	3	1.1%	0.0%	1.5%
	DS-Coder	4	1.2%	0.0%	0.4%
	Llama-70B	2	1.2%	1.2%	6.6%

Table 6: Free-generation diagnostic for DeepSeek-R1 on Dyck-n (N=500 per condition, 101 input sequences). Near-identical behavior on clean and corrupted chains confirms systematic chain bypassing rather than corruption detection. Accuracy substantially exceeds the 25% chance baseline (4-way bracket choice) in both conditions, confirming R1 re-solves from the raw input rather than failing randomly.

Metric	**Corrupted**	**Clean**
Recovery rate	80.0%	76.0%
Accuracy	51.4%	55.4%
Chance baseline (4-way)	25.0%	
N	500	500
Unique input sequences	101	

Table 7: Linear probe accuracy at the middle transformer layer ($\lfloor L/2 \rfloor$). Probe targets: Dyck-n = stack depth (multi-class, 0–10); PrOntoQA = final truth value (binary); GSM8K = operation type (multi-class); StrategyQA = final answer (binary). High probe accuracy paired with low or negative NLDD indicates representational silencing.

Model	**Task**	**Probe acc.**	**Interpretation**
DS-Coder	Dyck-n	68.1%	Above chance
	PrOntoQA	50.0%	Near chance
	GSM8K	36.3%	Below chance
Llama-8B	Dyck-n	82.8%	Strong encoding
	PrOntoQA	54.6%	Near chance
	GSM8K	67.0%	Above chance
	StrategyQA	67.2%	Above chance
Gemma	Dyck-n	89.6%	Strong encoding
	PrOntoQA	52.6%	Near chance
	GSM8K	67.4%	Above chance
	StrategyQA	64.6%	Above chance