BEYOND RATIONALIZATION: CRITERIA AND GUIDELINES FOR ALGORITHMIC REASONING TRACES IN LLM LOGICAL REASONING

Karun Thankachan*
Walmart E-commerce
karun.thankachan@walmart.com

Prateek Kohli
Walmart E-Commerce
prateek.kohli@walmart.com

ABSTRACT

Chain-of-thought (CoT) prompting is now a standard way to elicit "reasoning" from large language models, but recent work shows that CoT can hurt accuracy on some pattern-based in-context learning tasks and often produces explanations that look reasonable but do not match how the model actually made its decision. At the same time, symbolic and neuro-symbolic approaches such as Faithful CoT, SymbCoT, and Logic-LM connect natural language traces to executable formalisms and obtain higher accuracy and verifiable faithfulness on logical benchmarks. In this position paper, we argue that for logical reasoning it is important to separate *linguistic rationalization*, where CoT mainly describes an opaque pattern-matching process, from *algorithmic reasoning*, where the trace corresponds to a concrete computation that a solver can run. We propose four criteria—with default operationalizations—for treating a CoT trace as algorithmic reasoning in logic-focused tasks, clarify their intended normativity, and suggest a three-condition evaluation protocol that compares direct answering, free-form CoT, and symbolic or neuro-symbolic CoT. We ground the framework with a compact worked example on syllogistic reasoning that illustrates a key failure mode: free-form CoT can improve narrative plausibility without achieving solver-verifiable correctness.

1 INTRODUCTION

Large language models (LLMs) are now widely used for tasks that require logical reasoning, including math word problems, theorem proving, and multi-hop question answering (Cheng et al., 2025). Chain-of-thought (CoT) prompting—showing a model examples of intermediate steps or instructing it to "think step by step"—has become a standard recipe for eliciting reasoning behaviour and is often presented simultaneously as a performance technique and an interpretability tool (Wei et al., 2022; Sprague et al., 2024).

Recent empirical work reveals that these benefits are limited and uneven. Sprague et al. (2024) find that CoT helps mainly on math and symbolic tasks, with much smaller gains elsewhere. The "Curse of CoT" study shows that on pattern-based in-context learning (ICL) benchmarks, CoT underperforms direct answering and the gap widens with more in-context examples (Zheng et al., 2025). Concurrently, several analyses find that CoT explanations are often *not faithful* to the model's internal computation: traces can change under small prompt manipulations while the answer remains fixed, large-scale audits reveal many cases of "implicit post-hoc rationalisation," and CoT monitoring alone is unreliable as a safety mechanism (Arcuschin et al., 2025; Chen et al., 2025; Barez et al., 2025).

In contrast, symbolic and neuro-symbolic CoT methods obtain both higher accuracy and verifiable faithfulness on logical benchmarks by coupling natural language traces to external solvers (Lyu et al., 2023; Xu et al., 2024; Calanzone et al., 2024; Li et al., 2026). This contrast motivates the central question: *in logical reasoning settings, when should we treat a CoT trace as algorithmic reasoning rather than linguistic rationalization, and when should we prefer solver-backed traces?*

*Corresponding author

We make three contributions. **(1)** We synthesize recent evidence on when CoT helps or hurts and how symbolic methods address these issues. **(2)** We propose four operationalized criteria for algorithmic reasoning traces in logic-focused settings. **(3)** We recommend a three-condition evaluation protocol, grounded with a worked example demonstrating a key failure mode. Expanded background, detailed operationalization procedures, and additional examples are in the Appendix.

2 Background and Problem Statement

CoT prompting refers to giving a model examples of intermediate reasoning steps, or explicitly instructing it to "think step by step," so that it produces a sequence of natural language statements before an answer (Wei et al., 2022). Two threads of recent work frame our paper's contribution.

The faithfulness–utility distinction. An important terminological clarification: *faithfulness* refers to whether a trace reflects the internal computation that produced the answer, while *utility* refers to whether the trace is informative or helpful regardless of internal correspondence (Lee & Hockenmaier, 2025). Some work argues that CoT can be highly informative even when not faithful by strict causal criteria—for instance, because rationales contain hints that causally influence the model when included as input, or because incomplete rationales still predict the correct answer (Deng et al., 2025; Zaman & Srivastava, 2025; Lewis-Lim et al., 2025). Our criteria target a stronger property than utility: we require traces that are *both* solver-verifiable and causally stable, going beyond CoT that is merely informative. This is appropriate for logic-critical settings but may be unnecessary for open-ended or subjective tasks (see §3 and Appendix F for scope). Additional background on faithfulness nuances and symbolic CoT methods is in Appendices A–C.

Why logical reasoning demands more. On logical benchmarks where step correctness and trace reliability are central, the faithfulness shortcomings of free-form CoT are particularly consequential. Symbolic methods such as Faithful CoT, SymbCoT, and Logically Consistent Language Models define a regime we call *algorithmic reasoning*: the model outputs a structured or symbolic trace that has executable semantics, is checked or executed by an external solver, and is causally responsible for the final answer (Lyu et al., 2023; Xu et al., 2024; Calanzone et al., 2024).

3 Linguistic Rationalization vs. Algorithmic Reasoning

Recent theoretical and empirical work suggests that many CoT traces are better understood as linguistic rationalization than genuine reasoning. Shao & Cheng (2025) argue that CoT mainly acts as a tight structural constraint that makes models imitate the *form* of reasoning seen in training data, rather than performing abstract inference. Audits of frontier models show that traces can flip under small prompt changes while answers remain fixed and that models frequently produce "implicit post-hoc rationalisation"—a plausible story retrofitted onto an answer obtained by other means (Arcuschin et al., 2025; Chen et al., 2025).

To make the distinction operational for logical reasoning tasks, we propose four criteria for when a CoT trace can be treated as algorithmic reasoning (Table 1). **Intended normativity:** these criteria constitute a *recommended bar* for logic-critical settings—not a universal necessary-and-sufficient definition of algorithmic reasoning. A trace may satisfy fewer than four criteria and still be legitimately useful for non-logic tasks; conversely, all four should be met before a trace is treated as algorithmically reliable in formal logical reasoning. Criteria (i)–(ii) are closer to necessary conditions in strict logic settings; criteria (iii)–(iv) are robustness requirements that may be relaxed when multiple correct strategies exist or when external verification is unavailable (see Appendix F for explicit in-scope/out-of-scope examples).

The proposed default measurements are deliberately lightweight and imperfect; they are intended as a minimal reproducible baseline, not a definitive evaluation suite. Researchers should report which operationalization they use when applying these criteria.

Faithfulness vs. utility, revisited. Our criteria are stricter than general faithfulness metrics. A trace may be *useful* (high utility) without being fully faithful, and faithful by some metrics without being solver-verifiable. The four criteria jointly target *algorithmic faithfulness*: the trace must be

Table 1: Four criteria for algorithmic reasoning traces and their default operationalizations.

Criterion	Definition	Default Operationalization
(i) Formal entailment	Each step follows from prior steps and premises via identifiable inference rules (*e.g.*, modus ponens, arithmetic identities), not unsupported jumps.	*Rule-template match rate*: fraction of steps matching a named inference rule template from a fixed rule library (*e.g.*, checked by a theorem prover or rule-pattern classifier).
(ii) External verifiability	The trace can be checked by an independent mechanism (theorem prover, interpreter, symbolic solver) that detects incorrect or inconsistent steps.	*Verifier pass rate*: fraction of traces for which the solver accepts all steps without error. A trace failing the verifier is flagged as non-algorithmic.
(iii) Paraphrase invariance	Meaning-preserving rephrasings of the input preserve the overall reasoning *strategy* (sequence of rule types), not only the final answer.	*Strategy agreement score*: given k back-translated or template-varied paraphrases, fraction of paraphrase–original pairs sharing the same coarse-grained rule sequence (see Appendix D).
(iv) Counterfactual sensitivity	Minimal changes to key input facts produce appropriately modified intermediate steps, not superficial edits to a fixed template.	*Counterfactual step-edit rate*: fraction of steps that differ substantively between original and counterfactual traces, measured by a semantic similarity threshold or by solver re-execution (see Appendix D).

the executable cause of the answer. Traces that are merely informative—*e.g.*, free-form CoT that helps humans understand an answer but cannot be re-executed to produce it—satisfy utility but not algorithmic faithfulness, and should be treated as narrative aids in logic-critical evaluations.

Efficiency metric. Consistent with the paper's recommendation to normalize accuracy by cost, we propose reporting:

$$\text{Efficiency} = \frac{\text{Accuracy}}{\text{Generated tokens}} \quad \text{or} \quad \frac{\text{Accuracy}}{\text{Wall-clock seconds}}. \tag{1}$$

This allows improvements from conditions (B) or (C) over (A) to be interpreted relative to computational overhead, especially as long CoT traces and reasoning-optimized models can incur substantial token and latency costs (Sprague et al., 2024; Zheng et al., 2025).

4 Prescriptive Guidelines, Evaluation Protocol, and Worked Example

4.1 High-Level Guidelines

Motivated by the evidence reviewed in §3 and Appendix A, we offer three high-level guidelines for logical reasoning tasks:

- **Direct answering (A)** should be preferred for pattern-based ICL and induction-style tasks where CoT degrades performance, unless strong contrary evidence exists (Zheng et al., 2025; Liu et al., 2024).
- **Free-form CoT (B)** can serve as a generation aid for open-ended or subjective tasks, but its traces should *not* be treated as explanations in logic-critical settings, given their variable faithfulness (Arcuschin et al., 2025; Chen et al., 2025).
- **Symbolic or neuro-symbolic CoT (C)** should be the default for formal logical reasoning, mathematics, and safety- or audit-critical applications where step correctness and trace reliability matter and external solvers are available (Lyu et al., 2023; Xu et al., 2024; Calanzone et al., 2024; Li et al., 2026).

4.2 Three-Condition Evaluation Protocol

To make guidelines actionable across studies, we recommend reporting performance under three conditions for each logical reasoning benchmark:

1. **Direct answer (A).** The model produces an answer without explicit CoT, measuring implicit reasoning.
2. **Free-form CoT (B).** The model reasons step by step in natural language; the final answer is taken from its trace (Wei et al., 2022).
3. **Symbolic / neuro-symbolic CoT (C).** The model produces a structured or symbolic trace, executed or checked by an external solver, as in Faithful CoT, SymbCoT, and LoCo-LMs (Lyu et al., 2023; Xu et al., 2024; Calanzone et al., 2024).

The *pattern* of results across the three conditions is informative: if $B < A$ while $C > A$, free-form CoT is harmful but structured CoT recovers benefits via executable traces; if $B > A$ and $C \approx B$, free-form CoT suffices; if $B > A$ and $C > B$, most gains come from structure and external verification; if $A \approx B \approx C$, the benchmark does not strongly depend on explicit reasoning. Expanded interpretation patterns are in Appendix D, Table 2.

We additionally recommend reporting the Efficiency metric (Eq. 1) and, on a small subset of examples, the paraphrase invariance and counterfactual sensitivity scores from Table 1. These lightweight faithfulness checks require no new datasets and provide additional evidence about whether improvements in B or C correspond to algorithmic reasoning or to improved linguistic rationalization.

5 Conclusion

For logical reasoning with LLMs, our analysis motivates three practical implications. **Benchmarks** should report performance under direct answering, free-form CoT, and symbolic/neuro-symbolic CoT to make the role of explicit reasoning explicit. **Method development** should prioritize traces satisfying formal entailment, external verifiability, paraphrase invariance, and counterfactual sensitivity—not merely longer or more fluent rationales. **Interpretability and safety work** should treat free-form CoT as a narrative object unless traces are validated against these criteria or checked by an external verifier.

Chain-of-thought prompting, as commonly used, often behaves as linguistic rationalization rather than algorithmic reasoning; symbolic and neuro-symbolic methods show that executable, verifiable traces are achievable and better aligned with the goals of logical reasoning research. The four criteria—used together with the three-condition protocol and the lightweight operationalizations in Table 1—provide a concrete, reproducible starting point for distinguishing genuine algorithmic reasoning from plausible-sounding rationalization.

Acknowledgments

The author used Perplexity to summarize and iteratively refine this paper to meet the page limit and writing standards for ICLR. The position and framework proposed are based on the author's own research.

References

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

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, v1, 2025.

Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. *arXiv preprint arXiv:2409.13724*, 2024.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. Empowering LLMs with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*, 2025.

Amy Deng, Sydney Von Arx, Ben Snodin, Sudarsh Kunnavakkam, and Tamera Lanham. CoT may be highly informative despite "unfaithfulness". METR Blog Post, August 2025. URL https://metr.org/blog/2025-08-08-cot-may-be-highly-informative-despite-unfaithfulness/.

Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. *arXiv preprint arXiv:2502.12289*, 2025.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29826–29841, 2025.

Jindong Li, Yali Fu, Yang Yang, Jiahong Liu, Hongce Zhang, Haoxuan Li, Yutao Yue, and Menglin Yang. A survey on LLM symbolic reasoning. In *Logical and Symbolic Reasoning in Language Models @ AAAI 2026*, 2026.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, 2023.

Avinash Patil and Aryan Jadon. Advancing reasoning in large language models: Promising methods and approaches. *arXiv preprint arXiv:2502.03671*, 2025.

Jintian Shao and Yiming Cheng. CoT is not true reasoning, it is just a tight constraint to imitate: A theory perspective. *arXiv preprint arXiv:2506.02878*, 2025.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021.

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

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13326–13365, 2024.

Kerem Zaman and Shashank Srivastava. Is chain-of-thought really not explainability? chain-of-thought can be faithful without hint verbalization. *arXiv preprint arXiv:2512.23032*, 2025.

Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. The curse of CoT: On the limitations of chain-of-thought in in-context learning. *arXiv preprint arXiv:2504.05081*, 2025.

A Additional Background and Empirical Evidence

A.1 Adoption of Chain-of-Thought Prompting

Chain-of-thought prompting was introduced as a technique to elicit multi-step reasoning from LLMs by showing intermediate steps in few-shot exemplars or by instructing the model to "think step by step" before answering (Wei et al., 2022). Subsequent work quickly adopted CoT across a wide range of tasks, including math word problems, commonsense reasoning, multi-hop question answering, and program synthesis (Wei et al., 2022; Cheng et al., 2025; Sprague et al., 2024). Surveys of LLM reasoning note that CoT has become the dominant prompting paradigm when evaluating reasoning capabilities, often without strong ablations against direct answering or alternative methods (Cheng et al., 2025; Patil & Jadon, 2025; Li et al., 2026).

A.2 Performance Benefits and Limitations of CoT

Sprague et al. (2024) provide a large-scale empirical study of CoT across more than one hundred tasks and find that CoT helps primarily on math and symbolic reasoning benchmarks, with limited or no gains on many other tasks. Zheng et al. (2025) show that on pattern-based ICL tasks, CoT and related reasoning-style prompts consistently underperform direct answering, and that this gap widens as the number of in-context examples increases, attributing the effect to increased contextual distance between demonstrations and answers. Liu et al. (2024) perform controlled experiments on tasks where human deliberation is known to hurt performance and demonstrate that step-by-step prompting similarly reduces LLM accuracy on such tasks.

A.3 Faithfulness of CoT Explanations

Barez et al. (2025) argue that CoT is not explainability in a strict sense and present experiments where CoT traces change under small prompt perturbations while the underlying answer remains fixed. In a follow-up study, Arcuschin et al. (2025) show that frontier models frequently produce CoT traces that are coherent but do not reflect the internal computation on natural prompts. Chen et al. (2025) report similar behaviour for reasoning-optimized models, finding that CoT-style reasoning does not reliably expose unsafe internal computations.

At the same time, Zaman & Srivastava (2025) argue that many apparent failures of faithfulness are better described as incomplete rationales, and that faithfulness improves as rationale length increases. Deng et al. (2025) find that on tasks genuinely requiring multi-step reasoning, CoT traces tend to be more faithful even when they omit some details. Lewis-Lim et al. (2025) further show that CoT can fail to describe the exact internal computation yet still causally guide model predictions when rationales are included as part of the input. These nuances motivate our focus on stronger, logic-specific criteria rather than perfect faithfulness in general.

B Symbolic and Neuro-Symbolic CoT Methods

B.1 Faithful Chain-of-Thought Reasoning

Faithful Chain-of-Thought Reasoning (FCoT) proposes a two-stage architecture separating natural language understanding from symbolic problem solving (Lyu et al., 2023). In the first stage, the LLM translates the input into a symbolic chain (*e.g.*, a sequence of equations or logical statements). In the second stage, a deterministic solver executes this chain to produce the final answer. Because the solver is the only component allowed to produce answers, every answer is guaranteed to be derived by executing the symbolic chain, and any mismatch between the chain and the solver output is detectable. Lyu et al. (2023) show improved performance on math word problems and provide ablations demonstrating that solver-based execution, rather than just the extra text, accounts for the gains.

B.2 Symbolic CoT for Logical Reasoning

Faithful Logical Reasoning via Symbolic Chain-of-Thought (SymbCoT) extends this idea to first-order logic (Xu et al., 2024). The model is trained to generate a sequence of intermediate formulas and

inference steps, each aligned with a formal rule (*e.g.*, universal instantiation, conjunction introduction), which are then checked or executed by a logic engine such as a theorem prover. Xu et al. (2024) evaluate on benchmarks like ProofWriter and FOLIO and show that SymbCoT achieves higher accuracy and stronger guarantees of logical correctness than free-form CoT, particularly on longer proofs.

B.3 Logically Consistent Language Models and Surveys

Logically Consistent Language Models via Neuro-Symbolic Integration (LoCo-LMs) integrate logical constraints into the decoding process and use external reasoning to detect and correct contradictions (Calanzone et al., 2024). Instead of generating fully symbolic proofs, these models aim to maintain global consistency across answers to related queries (*e.g.*, transitivity of preferences, adherence to background axioms). Empirical results show that this approach significantly reduces logical inconsistencies while preserving strong task performance. A recent survey synthesises these and many related approaches (Li et al., 2026).

C Reasoning-Trace Evaluation and Our Criteria

Lee & Hockenmaier (2025) provide a survey of methods for evaluating step-by-step reasoning traces and propose a taxonomy with four main dimensions: groundedness (whether steps are supported by input or external knowledge), validity (whether steps follow logically from previous ones), coherence (whether the trace is locally and globally consistent), and utility (whether the trace is useful for humans or downstream tasks). They review metrics and datasets for each dimension, including perturbation-based tests, entailment checks, and human evaluations.

Our criteria for algorithmic reasoning traces overlap with this taxonomy but serve a different role. Formal entailment corresponds to a strong notion of validity, but we additionally require explicit alignment with named inference rules. External verifiability goes beyond their dimensions by insisting on an independent mechanism (*e.g.*, a solver) that can check traces and flag errors. Paraphrase invariance and counterfactual sensitivity are specific robustness conditions combining aspects of groundedness, coherence, and causal faithfulness, reflecting concerns raised in CoT explainability work (Arcuschin et al., 2025; Barez et al., 2025). In short, Lee & Hockenmaier (2025) provide a general evaluation framework, while our criteria are a stricter subset tailored to logical reasoning applications where traces are intended to function as algorithms.

D Details of the Three-Condition Protocol and Faithfulness Tests

D.1 Interpretation Patterns

Table 2: Illustrative interpretation patterns for the three-condition protocol.

Pattern	**Qualitative Interpretation**	**Example Scenario**
$B < A, C > A$	Free-form CoT harms performance; structured CoT recovers benefits via executable traces.	Pattern-based ICL where symbolic translation + solver helps (Zheng et al., 2025; Lyu et al., 2023).
$B > A, C \approx B$	Free-form CoT suffices; symbolic overhead is not justified for this task.	Short math problems where CoT yields near-perfect accuracy (Wei et al., 2022; Sprague et al., 2024).
$B > A, C > B$	CoT helps, but most gains come from structure and external verification.	Long-form proofs or complex logic benchmarks (Xu et al., 2024; Calanzone et al., 2024).
$A \approx B \approx C$	Task does not strongly depend on explicit reasoning.	Pattern-matching or noisy QA where context is sufficient (Sprague et al., 2024).

D.2 PARAPHRASE INVARIANCE TEST (OPERATIONALIZED)

Given an input x and its CoT trace τ, generate k paraphrases $\{x'_1, \ldots, x'_k\}$ that preserve task semantics. Default paraphrase generation: back-translation (source $\rightarrow$ pivot language $\rightarrow$ source) or controlled templating with synonym substitution. For each x'_i, obtain a CoT trace τ'_i using the same prompting and decode settings. Define trace strategy as the sequence of coarse-grained operations or rule types (*e.g.*, [MODUS PONENS, CONJUNCTION, MODUS PONENS]). Compute:

$$\text{StrategyAgreement}(\tau, \tau'_i) = \mathbf{1}[\text{RuleSeq}(\tau) = \text{RuleSeq}(\tau'_i)].$$

Report the mean agreement across all paraphrase pairs. High answer agreement but low strategy agreement indicates that the trace is not stable under paraphrase—a behaviour commonly observed in unfaithful CoT (Arcuschin et al., 2025; Barez et al., 2025). Recommended default: $k = 5$ paraphrases per example, on at least 50 examples per benchmark.

D.3 COUNTERFACTUAL SENSITIVITY TEST (OPERATIONALIZED)

Given an input x with CoT trace τ, construct counterfactual variants $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ by minimally changing key facts (*e.g.*, swapping entities, negating a premise, modifying a numeric constant). For each $\tilde{x}_j$, obtain trace $\tilde{\tau}_j$. The counterfactual step-edit rate is:

$$\text{StepEditRate}(\tau, \tilde{\tau}_j) = \frac{\left|\{s \in \text{Steps}(\tau) : \text{Sim}(s, \text{closest}(\tilde{\tau}_j)) < \theta\}\right|}{|\text{Steps}(\tau)|},$$

where Sim is a semantic similarity function and θ is a threshold (default: $\theta = 0.85$ cosine similarity with a sentence embedding model). High step-edit rate indicates the trace responds substantively to input changes. Recommended default: $m = 3$ counterfactual variants per example.

E EXTENDED WORKED EXAMPLE

To ground the framework, consider a multi-hop syllogistic task in the style of ProofWriter (Tafjord et al., 2021):

> **Premises:** (1) All mammals are warm-blooded. (2) All warm-blooded animals regulate body temperature. (3) Dolphins are mammals. *[Counterfactual variant: Dolphins are reptiles.]*
> **Question:** Do dolphins regulate body temperature?

Condition A (direct answer). A strong LLM can often answer "Yes" correctly via pattern matching. However, on the counterfactual variant where Dolphins are stipulated to be reptiles (making the answer "No," absent further premises), direct answering frequently still returns "Yes," revealing reliance on world knowledge rather than the stated premises.

Condition B (free-form CoT). Under "think step by step," a typical model produces:

> *All mammals are warm-blooded, and dolphins are mammals, so dolphins are warm-blooded. Warm-blooded animals regulate body temperature, so dolphins regulate body temperature. Yes.*

This trace is narratively plausible and matches the correct answer on the original. **However, this is the key failure mode:** on the counterfactual variant, models with free-form CoT frequently still generate an affirmative trace by silently overriding the stated premise with world knowledge. The trace sounds coherent but is not *causally* derived from the given premises. Narrative plausibility improves—the reasoning "looks like" valid deduction—while solver-verifiable correctness fails.

Condition C (symbolic / neuro-symbolic CoT). A SymbCoT-style system outputs:

```
1. forall x. Mammal(x) -> WarmBlooded(x)    [Premise 1]
2. forall x. WarmBlooded(x) -> RegTemp(x)   [Premise 2]
3. Mammal(Dolphin)                           [Premise 3]
4. WarmBlooded(Dolphin)                      [MP: 1, 3]
5. RegTemp(Dolphin)                          [MP: 2, 4]
```

An external prover verifies each step. On the counterfactual, Premise 3 becomes `Reptile(Dolphin)` and the chain breaks at step 4, correctly yielding "cannot derive." The trace satisfies all four criteria: formal entailment (named rules at each step), external verifiability (prover-checked), paraphrase invariance (strategy sequence preserved under rephrasing), and counterfactual sensitivity (the trace changes substantively when a key premise changes). This example illustrates the central claim: B can improve over A on surface metrics and plausibility yet fail the criteria for algorithmic reasoning, while C satisfies all four.

F Scope Clarification: In-Scope and Out-of-Scope Tasks

F.1 Intended Normativity

The four criteria are a **recommended bar for logic-critical settings**, not a universal necessary-and-sufficient definition of algorithmic reasoning. The criteria are intended to be used as a *joint* bar: a trace qualifies as algorithmically reliable when all four are satisfied, but partial satisfaction is still informative (*e.g.*, a trace that passes criteria (i) and (ii) but fails (iii) may be algorithmically correct but prompt-sensitive).

F.2 In-Scope Tasks

The criteria are appropriate for:

- **Formal logical reasoning:** Tasks with well-defined premises, rules, and conclusions where a theorem prover or constraint solver can serve as the verifier (*e.g.*, ProofWriter, FOLIO, AR-LSAT, LogiQA).
- **Mathematical reasoning with symbolic grounding:** Tasks where derivations can be expressed in a formal system and checked symbolically (*e.g.*, GSM8K with symbolic verification, MATH with proof assistants).
- **Safety- and audit-critical applications:** Settings where traces are used to explain or justify model decisions to external stakeholders and errors have downstream consequences.
- **Neuro-symbolic pipeline evaluation:** Benchmarks specifically designed to evaluate symbolic CoT methods.

F.3 Out-of-Scope Tasks

The criteria are not intended to apply (or should be applied loosely) for:

- **Open-ended or subjective reasoning:** Tasks where multiple correct strategies exist, invariance is not expected, and "correctness" is contested (*e.g.*, essay evaluation, moral reasoning, creative writing).
- **Tasks without tractable external verifiers:** Settings where model-internal algorithmic behaviour has no easy external verification mechanism. Here, criteria (i) and (iv) may still apply, but criterion (ii) should be relaxed.
- **Pattern-based ICL and induction tasks:** As documented by Zheng et al. (2025) and Liu et al. (2024), CoT itself is often counterproductive here; the three-condition protocol still applies, but the algorithmic reasoning criteria are not the relevant diagnostic.
- **Retrieval-augmented or knowledge-grounded tasks:** Where the "strategy" depends on retrieved content and paraphrase invariance in the strict sense is not expected.

F.4 Multiple Correct Strategies

Criterion (iii) (paraphrase invariance) requires strategy agreement across paraphrases, but for some tasks multiple correct proof strategies exist (*e.g.*, a lemma can be proved by induction or by direct construction). In such cases, the strategy agreement score should be computed at the *answer class* level rather than the exact rule-sequence level, or researchers should adopt a weaker notion of strategy: two traces agree if they produce solver-verified correct answers via any valid derivation. This relaxed

form of criterion (iii) remains informative about prompt sensitivity without over-penalizing legitimate strategy diversity.