Beyond Rationalization: Criteria and Guidelines for Algorithmic Reasoning Traces in LLM Logical Reasoning

Karun Thankachan*, Prateek Kohli

Walmart

Rationalization Problem

Chain-of-Thought (CoT) is the standard for "reasoning," but it has a hidden flaw: it often produces unfaithful trace i.e. plausible-sounding explanations that don't actually reflect how the model reached its answer.

Linguistic Rationalization: The model "narrates" a guess. The trace is a post-hoc story, often fragile to paraphrasing and unverifiable by external tools.
Algorithmic Reasoning: The model "calculates" the answer. The trace represents a concrete computation that an external solver can execute and verify.

When should we use CoT and what should we use CoT trace as algorithmic reasoning rather than linguistic rationalization?

Metric for Comparison
Efficiency = Accuracy/(Generated Tokens)

Guidelines for categorizing Task and CoT Usage

Direct Answering(A) – preferred for pattern-based ICL and induction-style tasks. CoT degrades performance

Free-form CoT (B) – can help serve as generation aid for open-ended or subjective tasks, but its traces should NOT be treated as explanations

Symbolic or Neuro-Symbolic CoT (C) – preferred for pattern-based ICL and induction-style tasks. CoT degrades perfomance

Worked Out Example

Premises: (1) All mammals are warm-blooded. (2) All warm-blooded animals regulate body temperature. (3) Dolphins are mammals.
Question: Do dolphins regulate body temperature?

Direct Answer: Yes

Free-Form CoT i.e. promot would contain “think step-by-step” it might generate something like – “*All mammals are warm-blooded, and dolphins are mammals, so dolphins are warmblooded. Warm-blooded animals regulate body temperature, so dolphins regulate body temperature. Yes.*”

Symbolic/Neuro-Symbolic CoT:
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]

Criteria for Algorithmic Reasoning Trace

Formal Entailment – Each step follows from prior steps and premises via identifiable inference rules
Rule-template match rate: fraction of steps matching a named inference rule template from a fixed rule library

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.

Paraphrase Invariance - Meaning-preserving rephrasings of the input preserve the overall reasoning strategy, 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

Counterfactual Sensitivity - Minimal changes to key input facts produce appropriately modified intermediate steps, not superficial edit.
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