Structural reasoning — or surface-level sensitivity?

Evaluating Robustness of Reasoning Models on Parameterized Logical Problems

Naïm Es-Sebbani · Esteban Marquer · Yakoub Salhi · Zied Bouraoui

CRIL – CNRS & Université d'Artois, Lens · GREYC, Université de Caen Basse Normandie

ICML 2026 · Seoul, South Korea · 06–11 July 2026

01

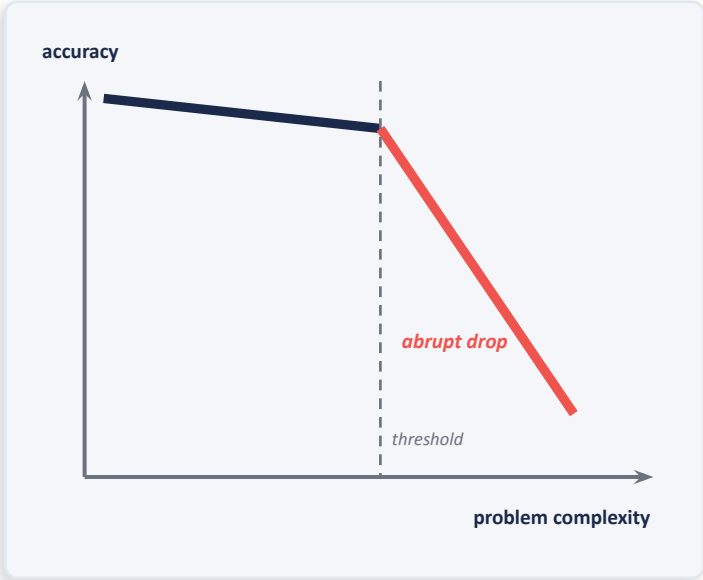
Reasoning LLMs ace combinatorial tasks — until they don’t.

On planning, puzzles and SAT, **LRMs often beat standard LLMs.**

But accuracy **drops sharply** as instances get harder (Hazra et al., 2025; Shojaee et al., 2025).

One interpretation: some LRMs imitate **solver-like steps** — propagation, backtracking, self-correction (Hazra et al., 2025).

→ If it is real reasoning: accuracy stays **stable** under semantics-preserving shifts, dropping **only** with structural difficulty.



Illustrative — after Shojaee et al., 2025; Hazra et al., 2025

01

Existing benchmarks can’t decide

SAT-in-natural-language

Mostly vary the **surface wording** of fixed instances.

Richardson & Sabharwal, 2022 · Pan et al., 2025 · Wei et al., 2025

Proof / trace-based

Grade the **explicit derivations** the model emits.

Saha et al., 2020 · Tafjord et al., 2021 · Creswell et al., 2023

Known risk: high accuracy from ***shortcut learning*** rather than rule-based generalization (Zhang et al., 2023).

The common gap: none isolates **genuine structure tracking** from surface difficulty.

→ *We need difficulty tuned along interpretable structural axes, while satisfiability is preserved.*

Our contribution

A parameterised 2-SAT diagnostic that pulls the two readings apart.

1

Parameterised families

Structured 2-CNF families + semantics-preserving perturbations: clause reordering, filler clauses, variable renaming.

2

Two metrics

Score both the SAT/UNSAT decision and the validity of the produced assignment — they often disagree.

3

Systematic study

Ten reasoning models (14B ⟶ 120B) probed across all generators, exposing generator-specific brittleness.

02 Why 2-SAT?

Polynomial-time, yet structurally rich.

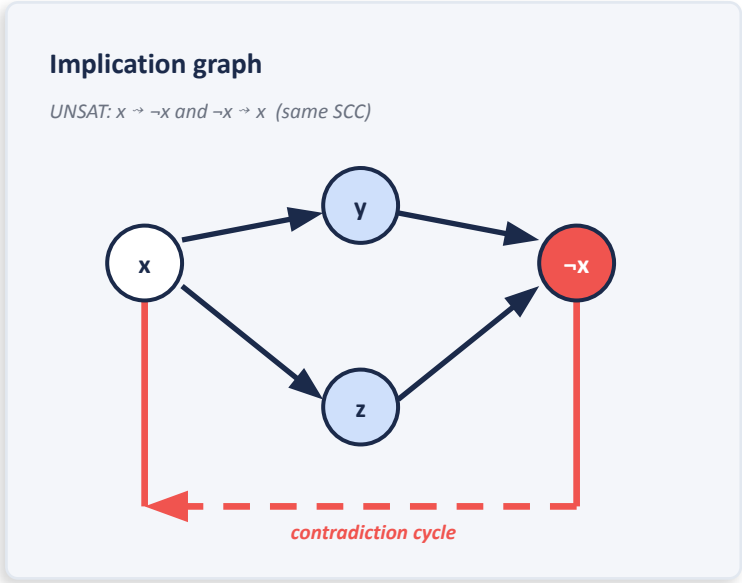
2-SAT is solvable in linear time — failure can't be blamed on raw computational hardness. It reveals a reasoning gap, not a search gap.

Exact graph characterisation.

A clause $(\ell_1 \vee \ell_2)$ reads as $\neg\ell_1 \Rightarrow \ell_2$ and $\neg\ell_2 \Rightarrow \ell_1$. UNSAT $\Leftrightarrow$ some x and ¬x lie in the same SCC of the implication graph.

The simplicity is the test.

Easy for a SAT solver — that's the point. It isolates structural reasoning from search hardness; moving to 3-SAT would conflate the two.



03

Five parameterised generators

Each axis is a surgical probe for one reasoning competency.

ImplicationCycle	UNSAT	**Long-range implication tracking**	*Close a contradiction cycle $\ell_1 \to \ldots \to \neg\ell_1$.*
EquivalenceCore	SAT	**Equivalence-class propagation**	*Collapse bound variables from a small free set.*
Backbone	SAT	**Forced-value inference**	*Planted backbone + a monotone remainder.*
MonoBridge	SAT	**Sensitivity to one bridge clause**	*A single bridge couples two monotone regions.*
Symmetry / Redundancy	SAT or UNSAT	**Reuse under renaming**	*Two isomorphic copies: $\Psi \wedge \rho(\Psi)$.*

03 One clause → many surfaces

2-CNF clause

$(\neg a \vee b)$

Template verbalisers

deterministic · invertible · 6 surfaces

logic Either a is false or b is true.
team a is on Blue or b is on Red.
letter a doesn't pass the letter, or b does.
room room a is dark, or room b is lit.
social a is absent, or b attends.
door door a is closed, or door b is open.

LLM augmentation

3 themes · 2-step validator

An LLM rewrites the two literals as a short narrative paragraph in a chosen theme **(spy · heist · detective)**. A two-step validator then re-extracts both entities and their polarity from the generated text — mismatches trigger regeneration.

Pipeline & evaluation

Sat.

Decision accuracy

Is the SAT/UNSAT call correct?

Wit.

Witness validity (SAT only)

On a SAT call, does the model’s assignment actually satisfy every clause? Checked deterministically with pysat.

→ Two separate metrics. They often disagree

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	·	—	·	—	·	—	·	—
	20	·	—	·	—	·	—	·	—
	50	·	—	·	—	·	—	·	—
	100	·	—	·	—	·	—	·	—
EquivalenceCore · SAT									
	10	·	·	·	·	·	·	·	·
	15	·	·	·	·	·	·	·	·
	20	·	·	·	·	·	·	·	·
	50	·	·	·	·	·	·	·	·

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	0.0	—	·	—	·	—	·	—
	20	18.3	—	·	—	·	—	·	—
	50	11.7	—	·	—	·	—	·	—
	100	20.0	—	·	—	·	—	·	—
EquivalenceCore · SAT									
	10	100.0	·	·	·	·	·	·	·
	15	98.3	·	·	·	·	·	·	·
	20	96.7	·	·	·	·	·	·	·
	50	93.3	·	·	·	·	·	·	·

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	0.0	—	·	—	·	—	·	—
	20	18.3	—	·	—	·	—	·	—
	50	11.7	—	·	—	·	—	·	—
	100	20.0	—	·	—	·	—	·	—
EquivalenceCore · SAT									
	10	100.0	1.7	·	·	·	·	·	·
	15	98.3	0.0	·	·	·	·	·	·
	20	96.7	0.0	·	·	·	·	·	·
	50	93.3	0.0	·	·	·	·	·	·

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	0.0	—	·	—	100.0	—	·	—
	20	18.3	—	·	—	96.7	—	·	—
	50	11.7	—	·	—	81.7	—	·	—
	100	20.0	—	·	—	76.7	—	·	—
EquivalenceCore · SAT									
	10	100.0	1.7	·	·	100.0	91.7	·	·
	15	98.3	0.0	·	·	100.0	53.3	·	·
	20	96.7	0.0	·	·	100.0	63.3	·	·
	50	93.3	0.0	·	·	98.3	8.3	·	·

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	0.0	—	·	—	100.0	—	·	—
	20	18.3	—	·	—	96.7	—	·	—
	50	11.7	—	·	—	81.7	—	·	—
	100	20.0	—	·	—	76.7	—	·	—
EquivalenceCore · SAT									
	10	100.0	1.7	100.0	86.7	100.0	91.7	100.0	83.3
	15	98.3	0.0	96.7	45.0	100.0	53.3	100.0	46.7
	20	96.7	0.0	86.7	43.3	100.0	63.3	100.0	50.0
	50	93.3	0.0	83.3	1.7	98.3	8.3	100.0	8.3

04 Results

		Llama-3.3 70B		QwQ 32B		Qwen3-Next 80B		GPT-OSS 120B	
Generator	|C|	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.	Sat.	Wit.
ImplicationCycle · UNSAT									
	5	0.0	—	98.3	—	100.0	—	91.7	—
	20	18.3	—	65.0	—	96.7	—	93.3	—
	50	11.7	—	40.0	—	81.7	—	56.7	—
	100	20.0	—	45.0	—	76.7	—	40.0	—
EquivalenceCore · SAT									
	10	100.0	1.7	100.0	86.7	100.0	91.7	100.0	83.3
	15	98.3	0.0	96.7	45.0	100.0	53.3	100.0	46.7
	20	96.7	0.0	86.7	43.3	100.0	63.3	100.0	50.0
	50	93.3	0.0	83.3	1.7	98.3	8.3	100.0	8.3

Structural reasoning
— or surface-level sensitivity?

Thank you.

Questions are very welcome.

Naïm Es-Sebbani · essebbani@cril.fr