HardcoreLogic

Challenging Large Reasoning Models with Long-tail Logic Puzzle Games

Jingcong Liang[1*] Shijun Wan[1*] Xuehai Wu[1*] Yitong Li[2]
Qianglong Chen[2] Duyu Tang[2] Siyuan Wang[3] Zhongyu Wei[1,4]

[1]Fudan University [2]Huawei Technologies Ltd.
[3]University of Southern California [4]Shanghai Innovation Institute

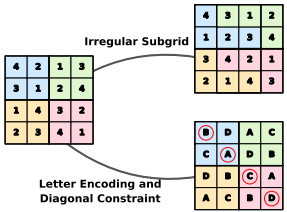
[*]Equal Contribution

ICLR 2026

Motivation: Do LRMs Truly Reason on Logic Puzzles?

Large Reasoning Models (LRMs) achieve impressive scores on logic puzzle benchmarks (e.g., ZebraLogic). **But are they genuinely reasoning?**

- Existing puzzle benchmarks are dominated by **canonical** forms (e.g., standard 9×9 Sudoku)
- Models may **overfit** to memorized formats and solution patterns
- Failure modes on **non-canonical variants**:
  1. Fail to understand new rules → faulty reasoning
  2. Apply rigid strategies → mismatched solutions



Two long-tail Sudoku variants

Key Question:

Can LRMs flexibly apply rules to **long-tail** puzzle variants?

HardcoreLogic: Benchmark Design

5,000+ puzzles across **10 logic games**, systematically transformed along **3 dimensions**:

Increased Complexity (IC)

IC1 **Search Space Expansion** — larger grids, more empty cells

IC2 **Constraint Strengthening** — partial & entangled hints

Uncommon Elements (UE)

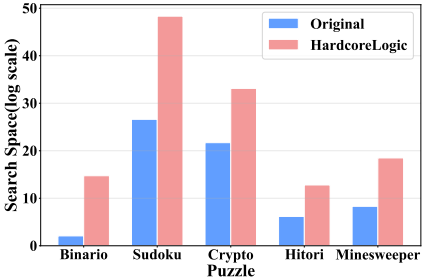
UE1 **Form Variation** — letter encoding, irregular subgrids

UE2 **Rule Variation** — diagonal constraints, multi-hop

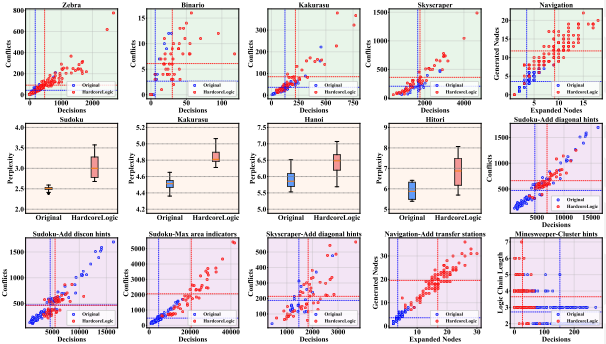
Unsolvable Puzzles (UP)

Deliberately designed with no valid solution to test whether models can detect logical contradictions

Complexity Analysis: HardcoreLogic is Harder

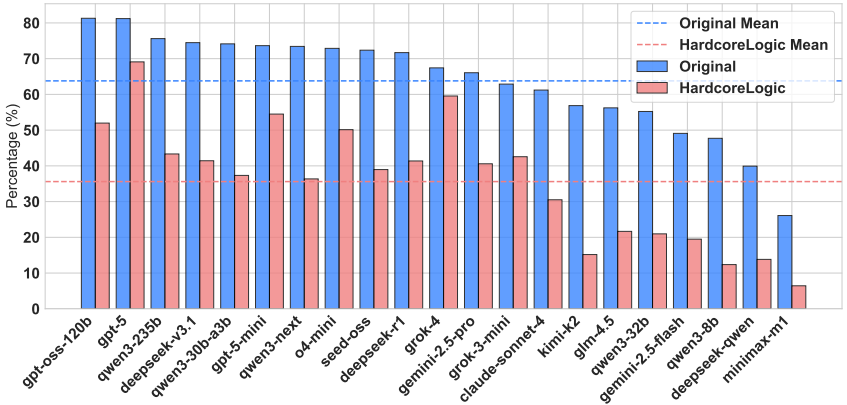


Search space comparison on IC1: HardcoreLogic vs. Original (existing benchmarks)



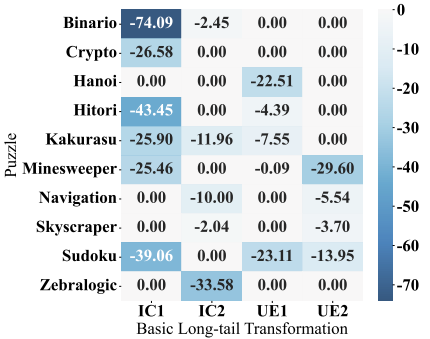
Average complexity comparison on IC2 (green), UE1 (orange) and UE2 (purple)

Overall Results: Significant Performance Degradation



All 21 LRMs show substantial accuracy drops from Original to HardcoreLogic.
Average accuracy: **~64%** (Original) → **~35%** (HardcoreLogic) — nearly **halved**.

Per-Transformation Analysis



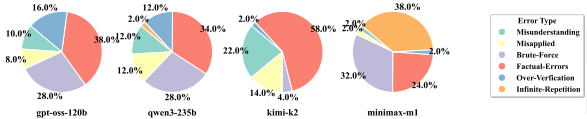
Average accuracy change (%) per puzzle and transformation type across all open-source models

Key Findings:

- **IC1 (Search Space Expansion)** causes the largest drops —up to **–74%** (Binario)
- **UE (Uncommon Elements)** also hurt performance significantly, even when puzzle difficulty is *not* increased
  - **UE1:** Sudoku — irregular grid patterns
  - **UE2:** Minesweeper — special mine rule

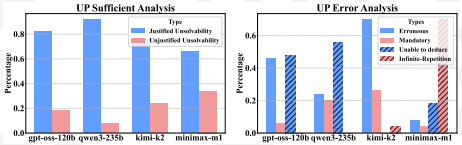
Error Analysis: Solvable & Unsolvable Puzzles

Solvable Puzzle Errors (6 categories):



- **Factual errors** dominate
- Stronger models show more **brute-force** attempts

Unsolvable Puzzle Analysis:



- **(In)sufficiency:** Weak models output “unsolvable” when they *fail to find* answers
- **Error mode:** Weaker models *force out* solutions; stronger ones get lost in *reasoning depth*

Thank you

Paper

Dataset

Code

Thank you!