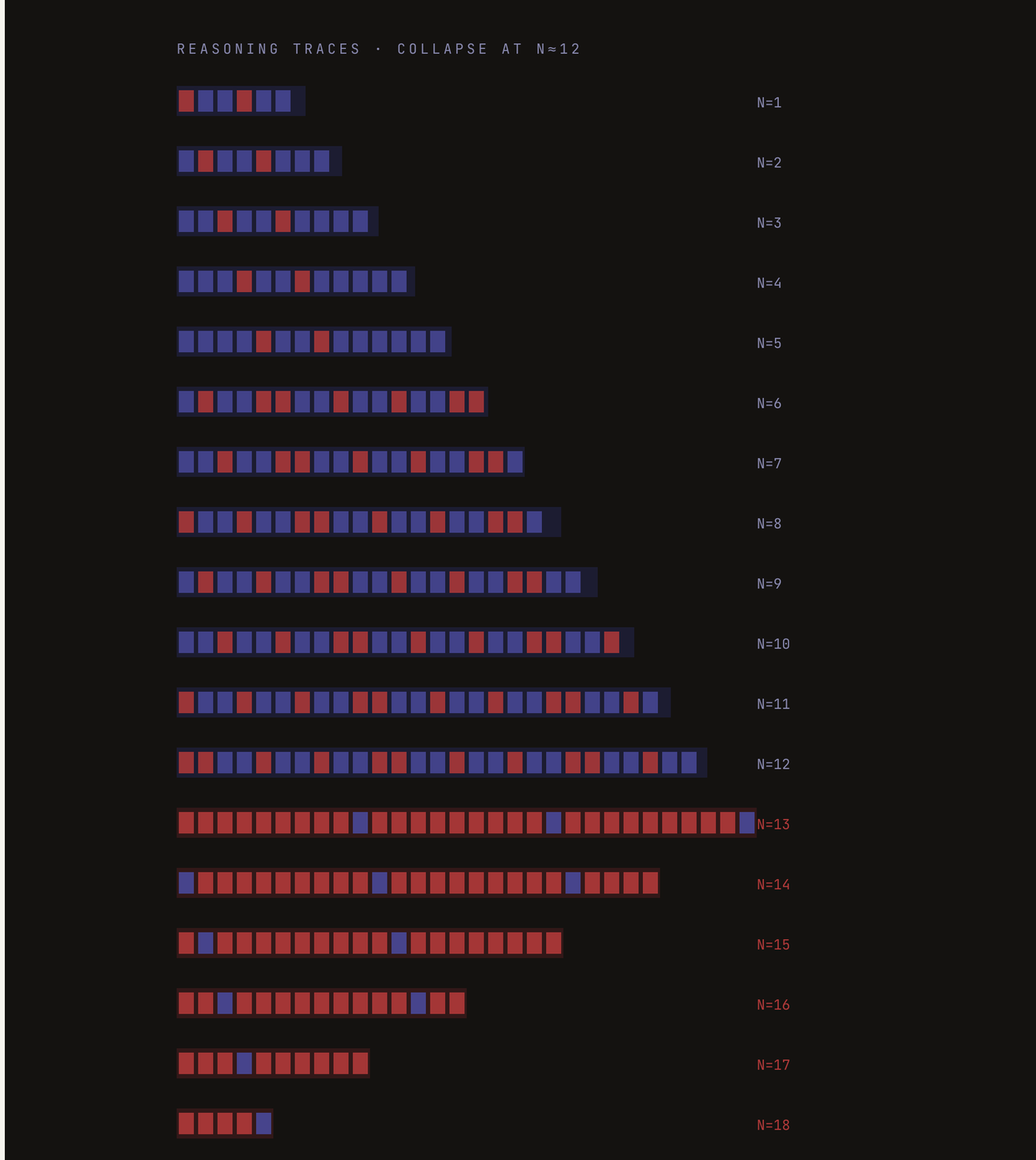
PAPER PRESENTATION · ARXIV:2506.06941

The Illusion of Thinking

Strengths & limitations of Large Reasoning Models, viewed through controllable problem complexity.

Shojaee, Mirzadeh, Alizadeh, Horton, Bengio, Farajtabar · NeurIPS 2025

PRESENTED BY KUNAL SALUJA



TL;DR

Reasoning models look like they think. Past a complexity threshold, they don't.

3

DISTINCT PERFORMANCE REGIMES VS. STANDARD LLMS

0%

ACCURACY BEYOND A MODEL-SPECIFIC COMPLEXITY THRESHOLD

REASONING EFFORT THAT FALLS AS PROBLEMS GET HARDER

THE PAPER

Understanding the strengths and limitations of reasoning models via the lens of problem complexity.

arXiv:2506.06941v3 — NeurIPS 2025
Submitted Jun 2025 · Revised Nov 2025

AUTHORS

Parshin Shojaee
Iman Mirzadeh*
Keivan Alizadeh
Maxwell Horton
Samy Bengio
Mehrdad Farajtabar

* equal contribution

MOTIVATION

A new class of models is claimed to "think." We don't really know what that means, or when it fails.

GAP · 01

Are LRMs doing *generalizable* reasoning, or a sophisticated form of pattern matching?

GAP · 02

How does their performance scale with problem complexity, beyond a fixed benchmark?

GAP · 03

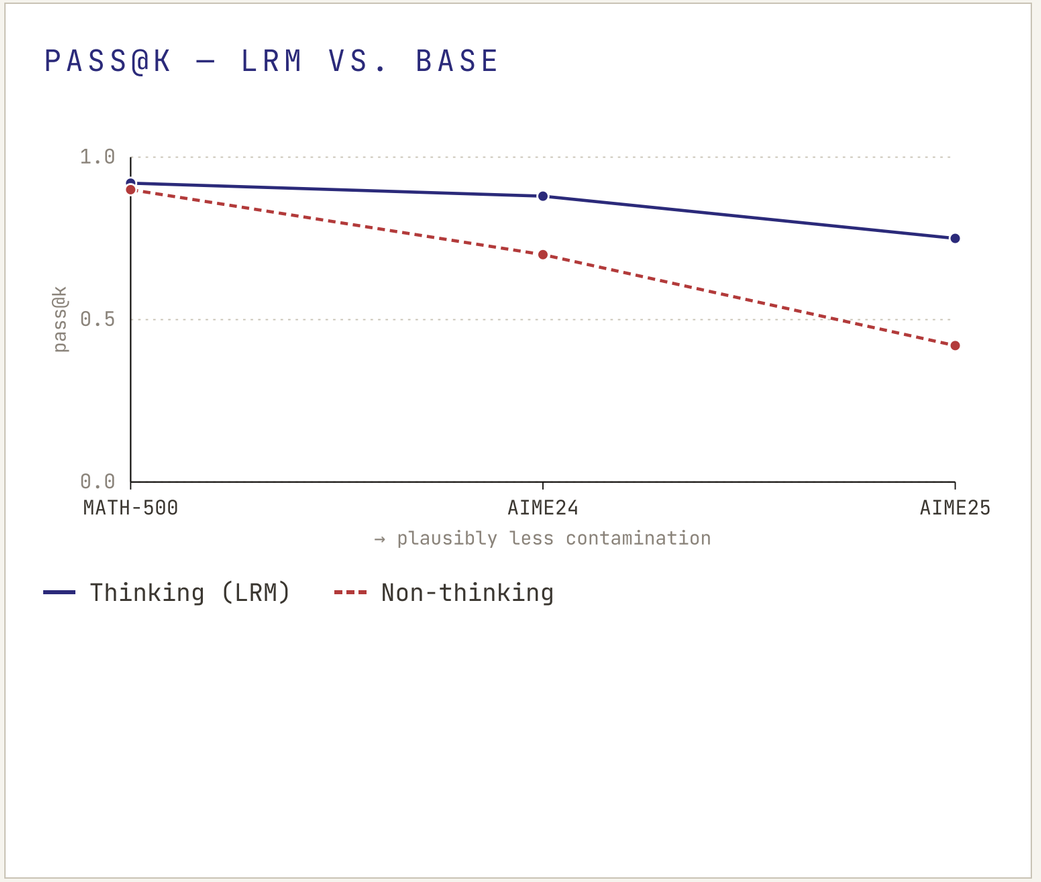
Do they actually use more compute at higher complexity — or do they give up?

THE PROBLEM WITH ESTABLISHED BENCHMARKS

MATH-500 and AIME suffer from contamination and tell us nothing about *how* a model reasoned.

Humans scored **higher** on AIME25 than AIME24… yet frontier LRMs score **lower**. A likely signature of training-data contamination in the older set.

Final-answer accuracy hides the reasoning. We need environments we can verify **step-by-step**.



APPROACH

Controllable puzzles as a microscope.

01 **Fine-grained control over complexity**

Vary a single parameter `N` while the logic stays fixed.

02 **Rule-bound, not knowledge-bound**

Every rule is in the prompt. No outside knowledge helps.

03 **Uncontaminated**

Controlled settings that don't appear in benchmark training sets.

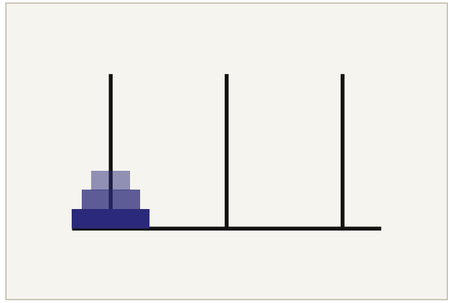
04 **Simulator-based verification**

Check every intermediate move, not just the final answer.

FOUR PUZZLE ENVIRONMENTS

Different flavors of compositional reasoning.

PUZZLE · 01

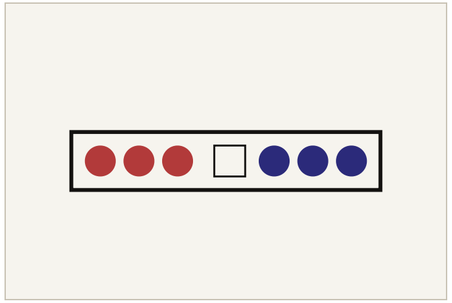


Tower of Hanoi

min moves: 2^N - 1

exponential

PUZZLE · 02

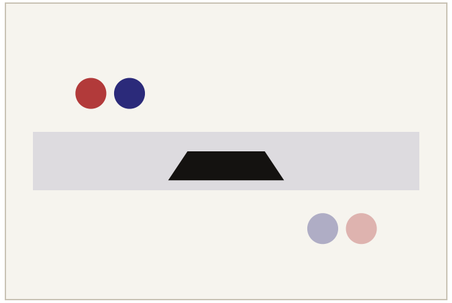


Checker Jumping

min moves: $(N+1)^2 - 1$

quadratic

PUZZLE · 03

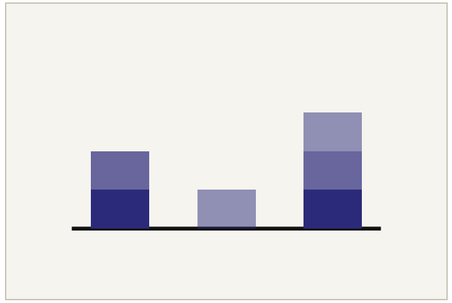


River Crossing

min moves: variable

constraint-sat

PUZZLE · 04



Blocks World

min moves: O(N)

planning depth

PUZZLE · 01

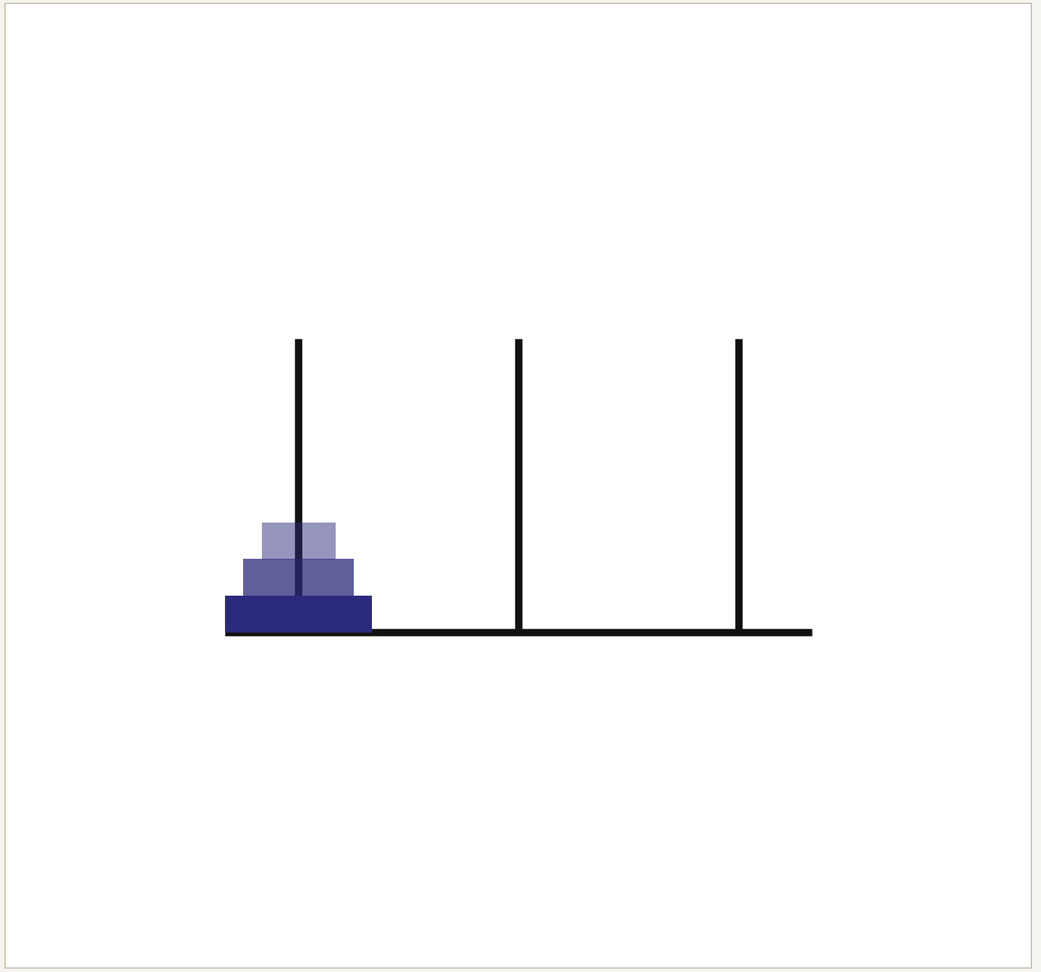
Tower of Hanoi

Move N disks from the first peg to the third. Classic recursive structure — minimum moves grow exponentially in N, so the task scales sharply.

MIN MOVES	$2^N - 1$ (exponential)
COMPLEXITY	compositional depth
CONTROL	number of disks N

KEY RULES

- One disk moves at a time.
- Only the topmost disk on a peg is movable.
- A larger disk may never rest on a smaller one.



PUZZLE · 02

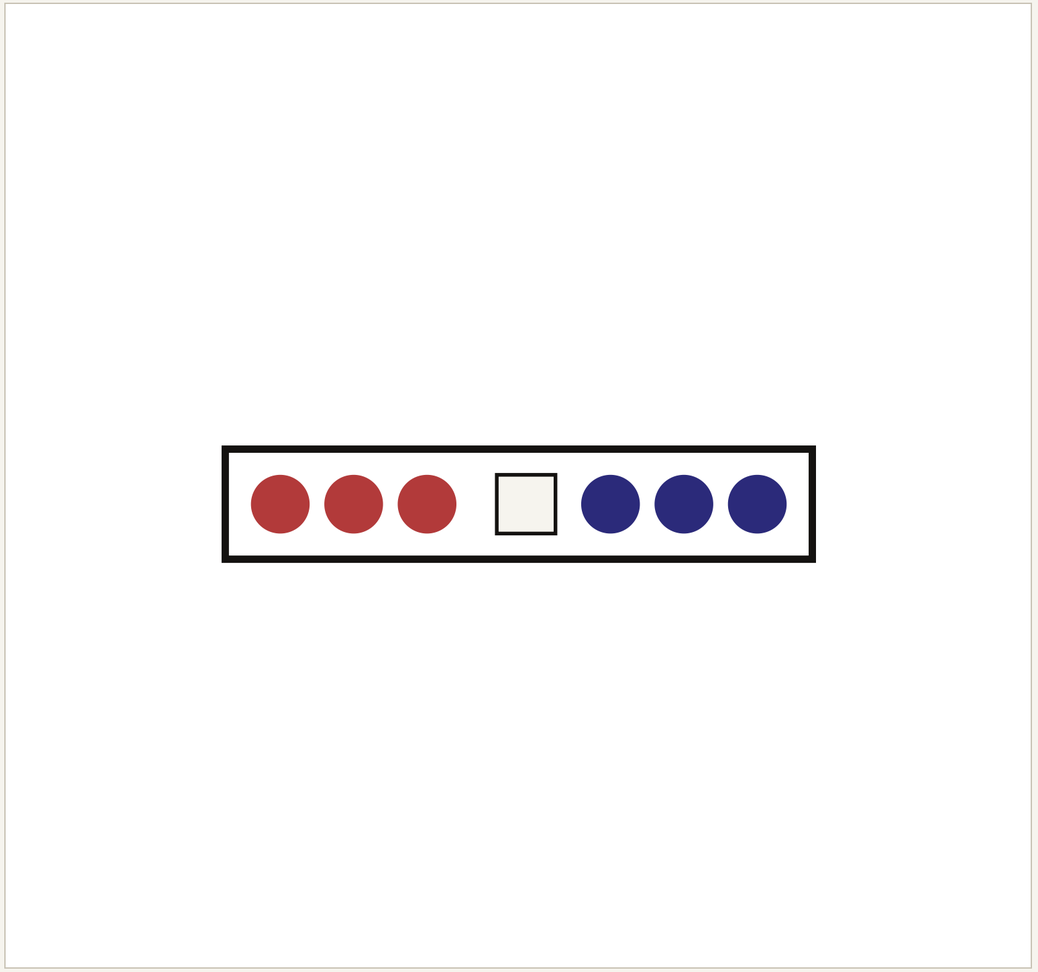
Checker Jumping

Swap the positions of N red and N blue checkers around a single empty square. A one-dimensional constraint problem with strict directionality.

MIN MOVES	$(N+1)^2 - 1$ (quadratic)
COMPLEXITY	spatial + planning
CONTROL	number of checkers 2N

KEY RULES

- Slide into an adjacent empty space, or
- Jump exactly one opposite-color checker to land in empty space.
- No backward moves allowed.



PUZZLE · 03

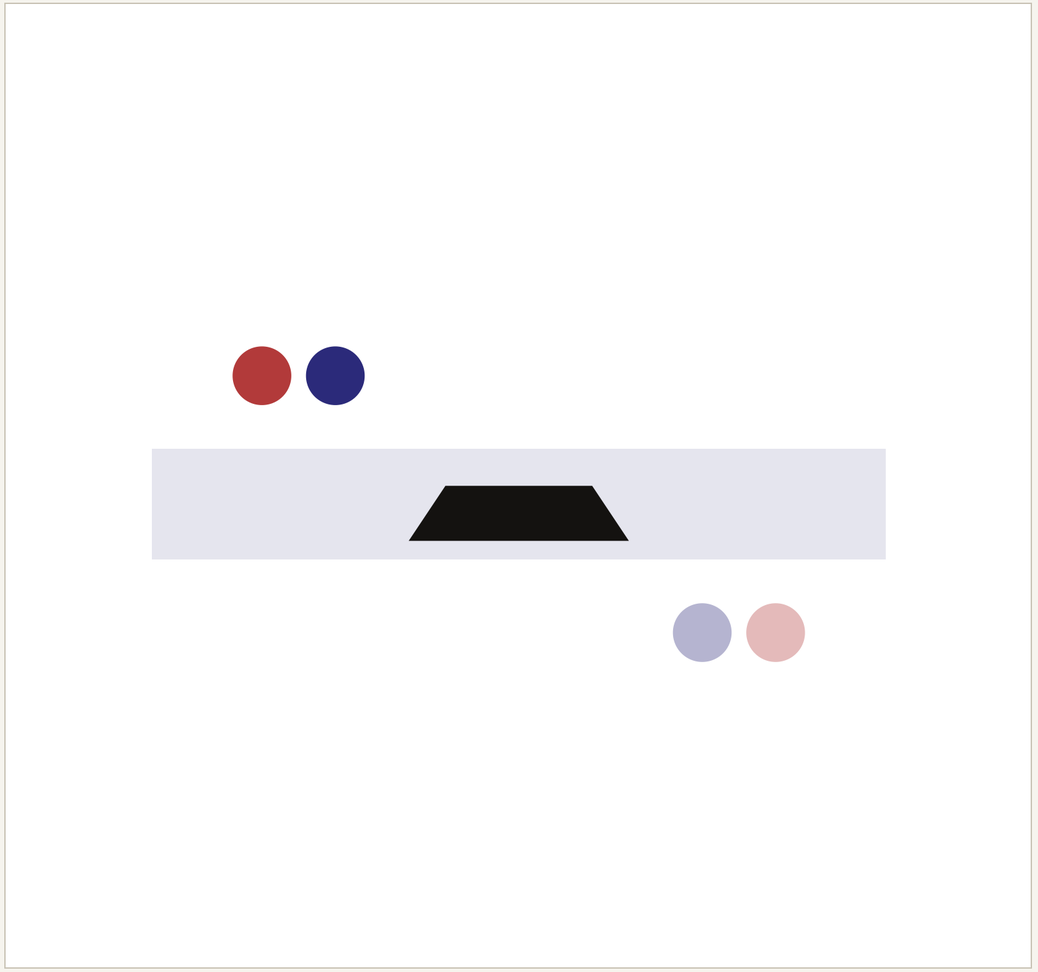
River Crossing

N actors and their N agents must all cross the river. Safety constraints force careful interleavings — a compact but deeply constrained planning problem.

MIN MOVES	variable (constraint-heavy)
COMPLEXITY	multi-agent coordination
CONTROL	number of actor-agent pairs

KEY RULES

- Boat holds k people, never empty.
- An actor must never be with another agent without their own agent present.
- Applies in the boat and on both banks.



PUZZLE · 04

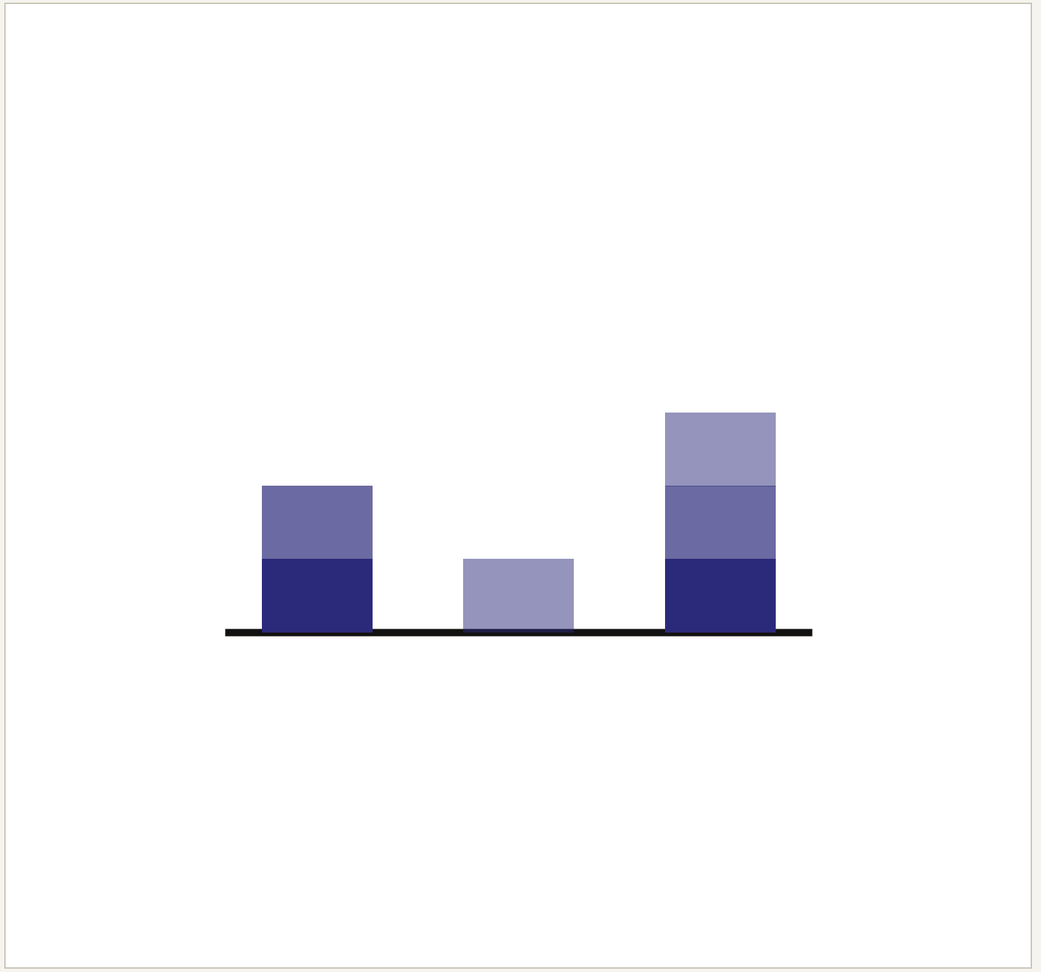
Blocks World

Reconfigure a set of stacks into a specified goal arrangement in the fewest moves. The oldest problem in AI planning, used here as a modern reasoning probe.

MIN MOVES	depends on configuration
COMPLEXITY	classical planning
CONTROL	number of blocks

KEY RULES

- Only the top block of a stack is movable.
- A block can be placed on an empty spot or atop another.
- Goal state is pre-specified.



EXPERIMENTAL SETUP

Matched thinking / non-thinking pairs under the same compute budget.

MODEL PAIRS

`Claude-3.7-Sonnet` (think / no-think)
`DeepSeek-R1` vs. `V3`
`o3-mini` (medium + high)
`DeepSeek-R1-Distill-Qwen-32B`
`QwQ-32B` vs. `Qwen-2.5-32B`

COMPUTE & SAMPLING

Up to `64k` tokens per response.

`25 samples` per (model, puzzle, N).

Filtered for adherence to response schema.

WHAT WE MEASURE

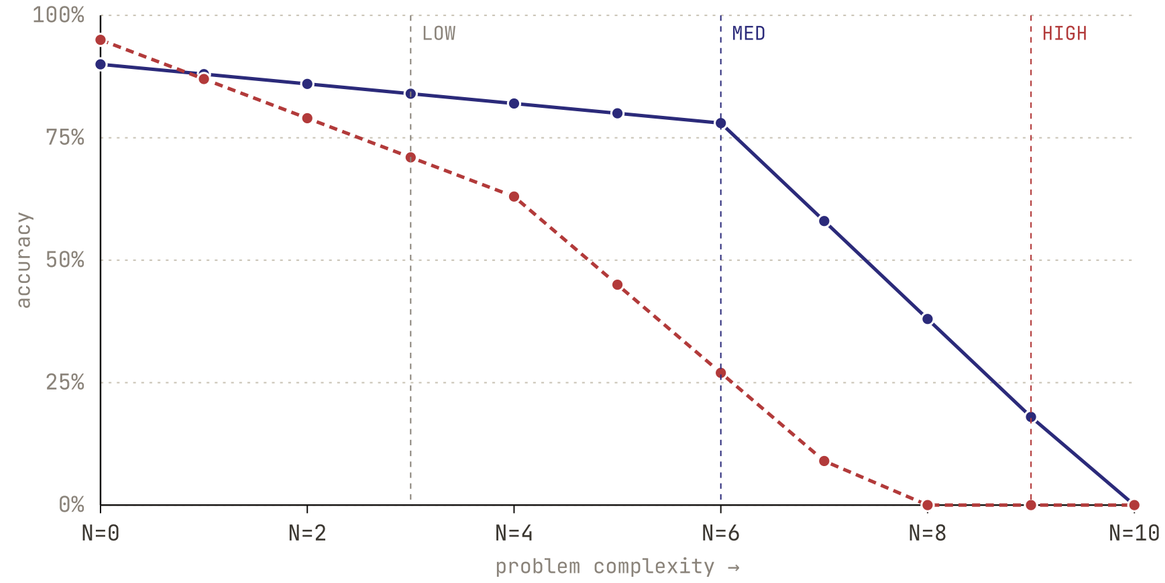
Final-move correctness.

Every intermediate solution found *inside* the trace.

Thinking tokens spent per problem.

FINDING · 01

Three regimes of complexity.



REGIME · LOW N

Non-thinking models **match or beat** LRMs, using fewer tokens.

REGIME · MEDIUM N

LRMs' advantage appears — extra thinking translates into accuracy.

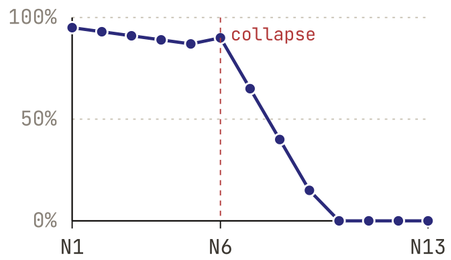
REGIME · HIGH N

Both **collapse to zero**, regardless of inference budget.

FINDING · 02

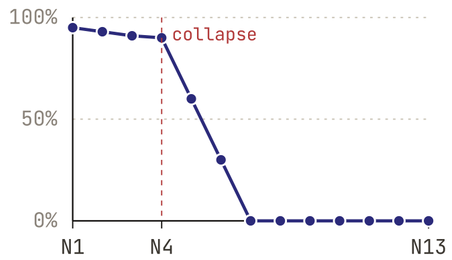
Every puzzle has a threshold. Past it, accuracy goes to zero.

01 · TOWER OF HANOI



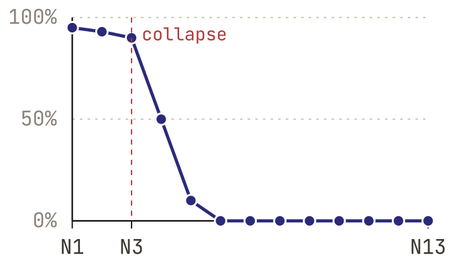
collapse N≈8

02 · CHECKER JUMPING



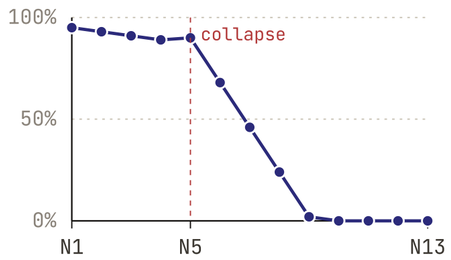
collapse N≈5

03 · RIVER CROSSING



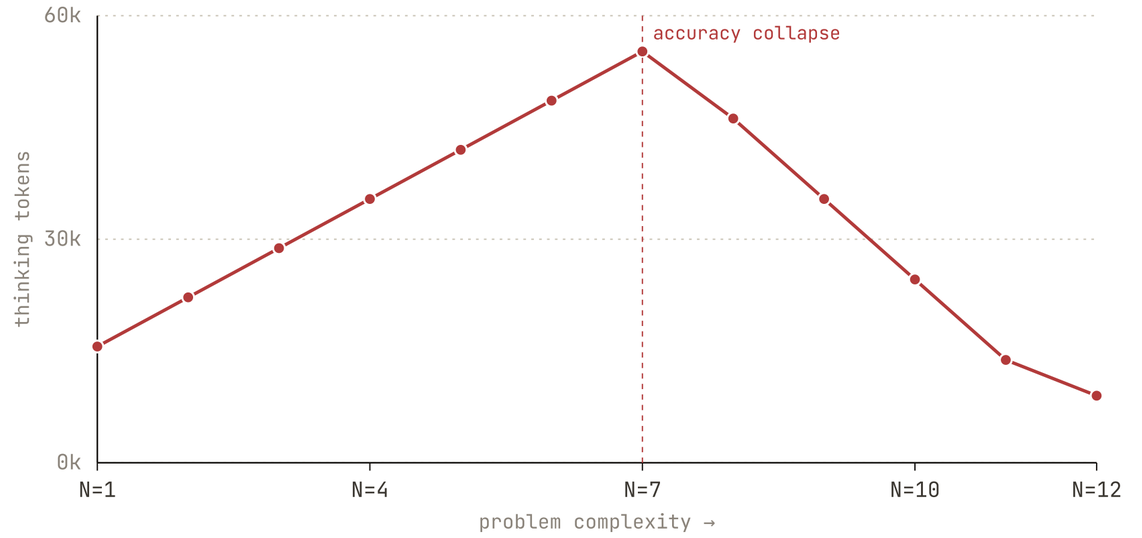
collapse N=3

04 · BLOCKS WORLD



collapse N≈10

Near the collapse point, reasoning models *reduce* their thinking effort — while still holding budget in reserve.



budget ceiling = 64k tokens · never hit

o3-mini shows the sharpest fall

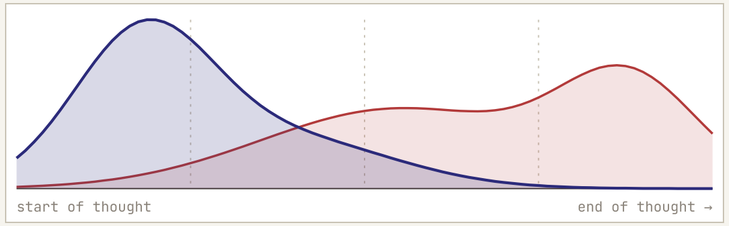
You'd expect harder problems to elicit *more* chain-of-thought. For a while, they do.

Past a model-specific complexity, effort falls despite ample remaining budget — a fundamental `inference-time scaling limit`.

Complexity reshapes *when* correct solutions appear — if at all.

LOW COMPLEXITY

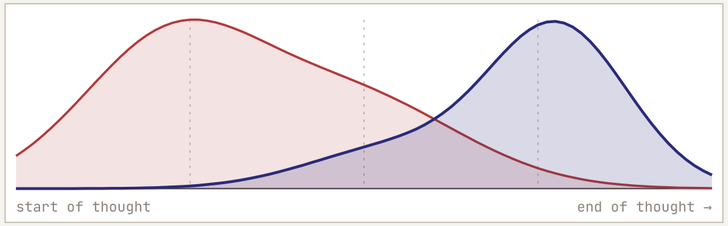
overthinking



Correct answer appears *early*. The model keeps exploring wrong paths afterwards.

MEDIUM COMPLEXITY

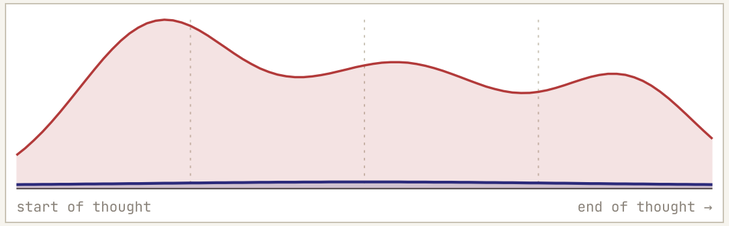
search pays off



Wrong paths dominate first. Correct answer only emerges *late* in the trace.

HIGH COMPLEXITY

collapse



No correct solution appears anywhere. The trace fixates on wrong paths.

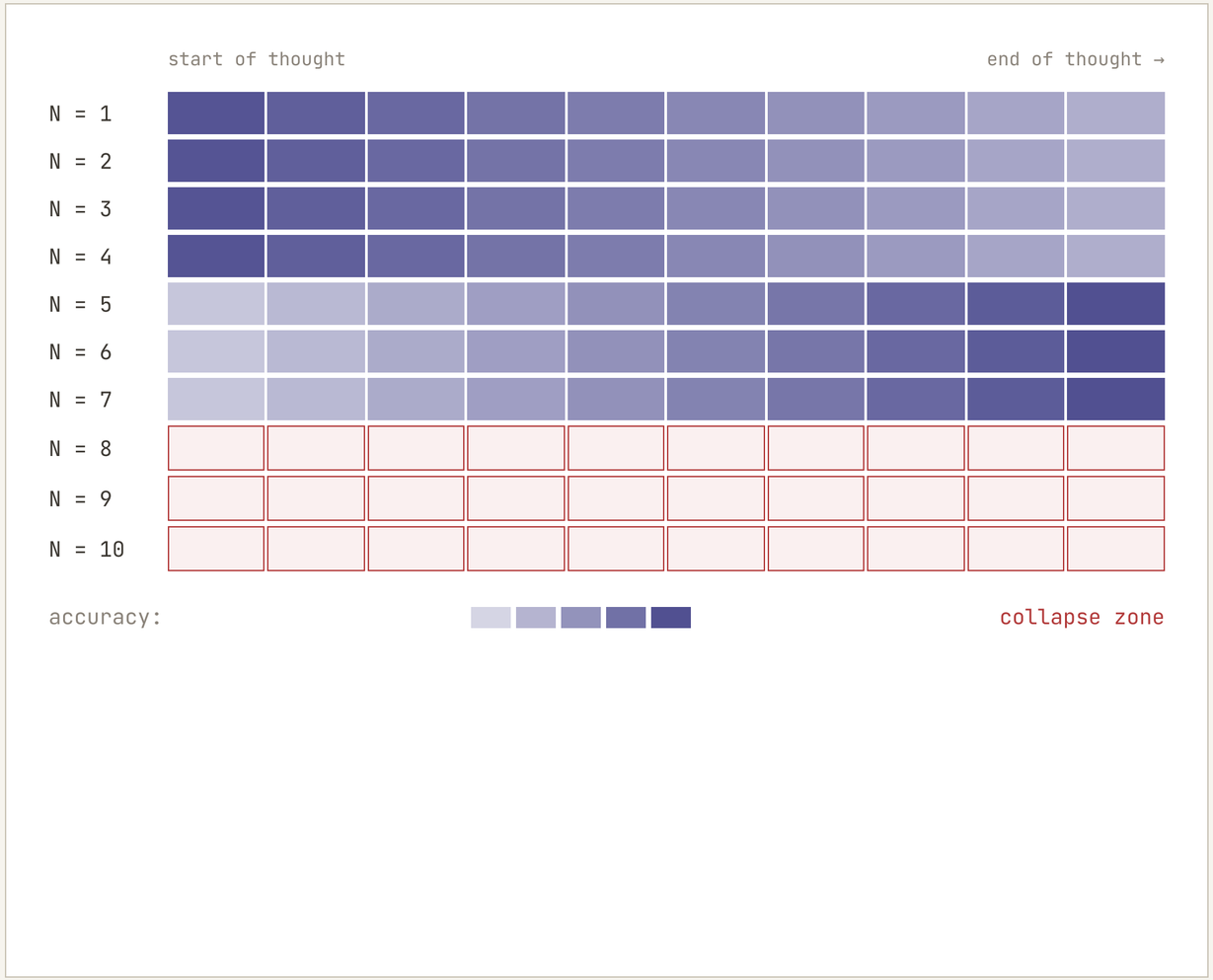
density of **correct** intermediate solutions

density of **incorrect** intermediate solutions

x-axis · position within the thinking trace (0 → 1)

TOWER OF HANOI · CLAUDE-3.7-THINKING

Accuracy by position inside the thought, across N.



N = 1 – 4 · OVERTHINKING

Accuracy starts high, drifts down as the model second-guesses.

N = 5 – 7 · SEARCH PAYS OFF

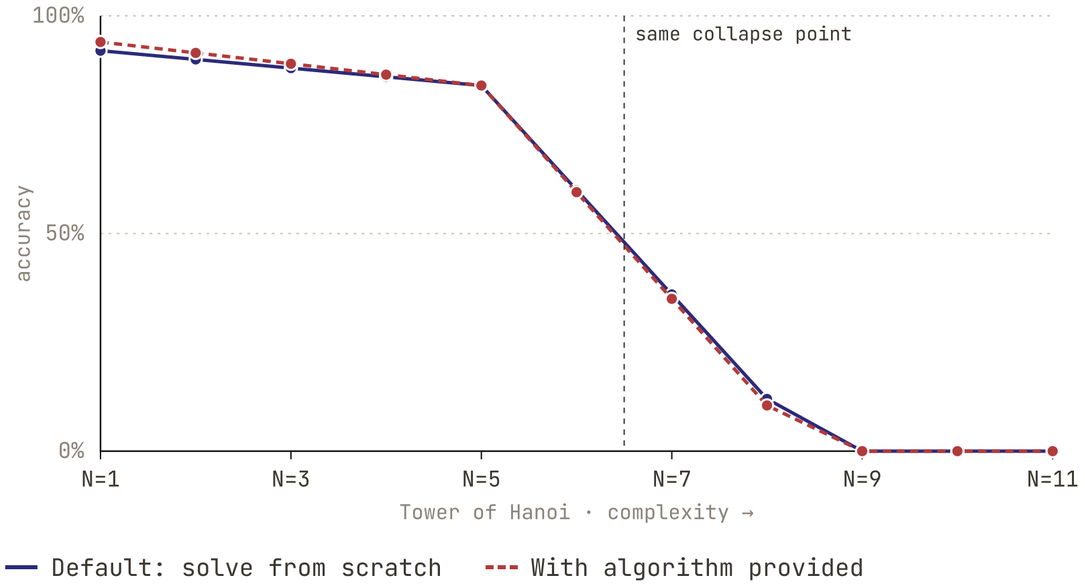
Accuracy climbs with more thinking — up to a threshold.

N ≥ 8 · COLLAPSE

Near-zero across all positions. No recovery inside the trace.

FINDING · 05

Giving the model the algorithm doesn't help.



Prompt the model with the full recursive algorithm. Only *execution* remains.

Performance barely moves. The same collapse appears at the same N.

> LRMs fail at `reliable symbolic execution`, not just at search.

"Complexity" for a model ≠ computational complexity of the task.

ERROR-FREE RUN LENGTH

An 11-move River Crossing breaks the model. A 255-move Tower of Hanoi doesn't.

NON-MONOTONIC FAILURE

Claude-3.7-Thinking on Tower of Hanoi:

N = 10	N = 12
100+	< 50
first error after	first error after

The same puzzle at higher N fails *earlier* — behavior tracks training-data familiarity, not problem size.

TAKEAWAY

What we call "thinking" may be a smoothly degrading pattern match — plus a scaling ceiling we haven't learned to see past.

EVALUATION

Controllable, trace-level evaluation beats benchmarked final-answer accuracy.

LIMITS

LRMs have a complexity threshold and a reasoning-effort ceiling they can't break.

OPEN

Why can't they execute an algorithm they're handed? What does "reasoning" even mean here?

arXiv:2506.06941 · NeurIPS 2025

Thank you.