The Illusion of Thinking: What Reasoning Models Get Right (and Where They Break)

Reasoning models look like a big step forward: they generate a long chain of intermediate steps, then land on an answer. On math and coding benchmarks, that often works. But a question keeps bothering me: are these models actually getting better at reasoning, or are they just performing well on the kinds of problems we already know how to measure?

A NeurIPS 2025 paper called *The Illusion of Thinking* tackles that question using controllable puzzle environments. The key trick is that the puzzles let researchers dial up compositional complexity while keeping the underlying logic consistent, so you can study not only final accuracy, but also what happens inside the "thinking trace" as problems get harder.

Why problem complexity matters

Most evaluations emphasize "did the model get the right answer?" on well-known benchmark distributions. The paper argues that this is incomplete (and sometimes misleading), because it doesn't reveal how reasoning behavior changes when you systematically push difficulty beyond the familiar range.

By controlling complexity directly, the authors can observe when a model's apparent reasoning ability is robust and when it starts to behave more like pattern-matching under stress.

Key findings that stood out

1. Accuracy can collapse once problems cross a certain complexity threshold, rather than degrading gradually.
2. The scaling behavior is counter-intuitive: "reasoning effort" (often measured via how much the model writes/uses its thinking tokens) increases with complexity up to a point, then drops as tasks get even harder—even when token budget is still available.
3. Under matched inference compute, the paper describes three regimes:
    - Low complexity: standard LLMs can outperform LRMs, suggesting extra "thinking" can add noise when tasks are easy.
    - Medium complexity: LRMs tend to do better, where structured intermediate reasoning actually helps.
    - High complexity: both approaches can fail badly, highlighting a fundamental limitation rather than a tuning issue.
4. Exact computation remains a weak spot: the traces often look heuristic and inconsistent, which is a red flag for tasks that require algorithmic, deterministic steps.

What this changes for evaluation

The takeaway isn't "reasoning models are useless." It's that we should be more careful about what we infer from benchmark wins. If accuracy collapses beyond a complexity boundary, then "more tokens" or "more compute at inference" isn't a universal fix—and it becomes important to test models in settings where difficulty is controlled, not just sampled.

It also reinforces something that's easy to forget: a convincing chain-of-thought can be a UI artifact, not proof of stable internal computation. If the trace quality degrades or becomes

inconsistent as complexity rises, the model may be narrating a path rather than executing one.

Why I'm paying attention

This line of work feels important because it forces a sharper definition of "reasoning." If a model can only reason inside a comfort zone, then the real problem becomes: how do we build systems that fail predictably, expose uncertainty, and reliably handle tasks that demand exactness?