The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

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

Apple

Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think". Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counterintuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low-complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.

1 Introduction

Large Language Models (LLMs) have recently evolved to include specialized variants explicitly designed for reasoning tasks—Large Reasoning Models (LRMs) such as OpenAI's o1/o3 [1, 2], DeepSeek-R1 [3], Claude 3.7 Sonnet Thinking [4], and Gemini Thinking [5]. These models are new artifacts, characterized by their "*thinking*" mechanisms such as long Chain-of-Thought (CoT) with self-reflection, and have demonstrated promising results across various reasoning benchmarks. Their

*Equal contribution.

†Work done during an internship at Apple.

{p_shojaee, imirzadeh, kalizadehvahid, mchorton, bengio, farajtabar}@apple.com

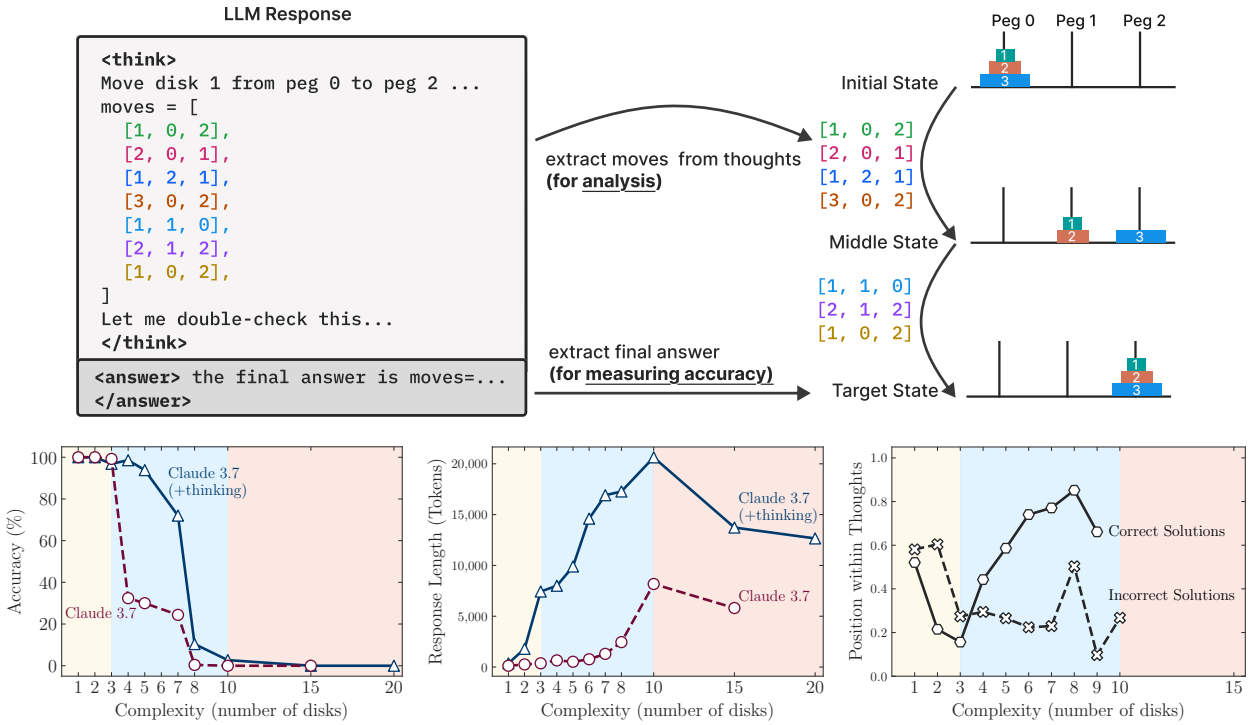

Figure 1: **Top**: Our setup enables verification of both final answers and intermediate reasoning traces, allowing detailed analysis of model thinking behavior. **Bottom left & middle**: At low complexity, non-thinking models are more accurate and token-efficient. As complexity increases, reasoning models outperform but require more tokens—until both collapse beyond a critical threshold, with shorter traces. **Bottom right**: For correctly solved cases, Claude 3.7 Thinking tends to find answers early at low complexity and later at higher complexity. In failed cases, it often fixates on an early wrong answer, wasting the remaining token budget. Both cases reveal inefficiencies in the reasoning process.

emergence suggests a potential paradigm shift in how LLM systems approach complex reasoning and problem-solving tasks, with some researchers proposing them as significant steps toward more general artificial intelligence capabilities.

Despite these claims and performance advancements, the fundamental benefits and limitations of LRMs remain insufficiently understood. Critical questions still persist: Are these models capable of generalizable reasoning, or are they leveraging different forms of pattern matching [6]? How does their performance scale with increasing problem complexity? How do they compare to their non-thinking standard LLM counterparts when provided with the same inference token compute? Most importantly, what are the inherent limitations of current reasoning approaches, and what improvements might be necessary to advance toward more robust reasoning capabilities?

We believe the lack of systematic analyses investigating these questions is due to limitations in current evaluation paradigms. Existing evaluations predominantly focus on established mathematical and coding benchmarks, which, while valuable, often suffer from data contamination issues and do not allow for controlled experimental conditions across different settings and complexities. Moreover, these evaluations do not provide insights into the structure and quality of reasoning traces. To understand the reasoning behavior of these models more rigorously, we need environments that enable controlled experimentation.

In this study, we probe the reasoning mechanisms of frontier LRMs through the lens of problem

complexity. Rather than standard benchmarks (e.g., math problems), we adopt controllable puzzle environments that let us vary complexity systematically—by adjusting puzzle elements while preserving the core logic—and inspect both solutions and internal reasoning (Fig. 1, top). These puzzles: (1) offer fine-grained control over complexity; (2) avoid contamination common in established benchmarks; (3) require only the explicitly provided rules, emphasizing algorithmic reasoning; and (4) support rigorous, simulator-based evaluation, enabling precise solution checks and detailed failure analyses. Our empirical investigation reveals several key findings about current Language Reasoning Models (LRMs): First, despite their sophisticated self-reflection mechanisms learned through reinforcement learning, these models fail to develop generalizable problem-solving capabilities for planning tasks, with performance collapsing to zero beyond a certain complexity threshold. Second, our comparison between LRMs and standard LLMs under equivalent inference compute reveals three distinct reasoning regimes (Fig. 1, bottom). For simpler, low-compositional problems, standard LLMs demonstrate greater efficiency and accuracy. As problem complexity moderately increases, thinking models gain an advantage. However, when problems reach high complexity with longer compositional depth, both model types experience complete performance collapse (Fig. 1, bottom left). Notably, near this collapse point, LRMs begin reducing their reasoning effort (measured by inference-time tokens) as problem complexity increases, despite operating well below generation length limits (Fig. 1, bottom middle). This suggests a fundamental inference time scaling limitation in LRMs' reasoning capabilities relative to problem complexity. Finally, our analysis of intermediate reasoning traces or thoughts reveals complexity-dependent patterns: In simpler problems, reasoning models often identify correct solutions early but inefficiently continue exploring incorrect alternatives—an "overthinking" phenomenon. At moderate complexity, correct solutions emerge only after extensive exploration of incorrect paths. Beyond a certain complexity threshold, models completely fail to find correct solutions (Fig. 1, bottom right). This indicates LRMs possess limited self-correction capabilities that, while valuable, reveal fundamental inefficiencies and clear scaling limitations.

These findings highlight both the strengths and limitations of existing LRMs, raising questions about the nature of reasoning in these systems with important implications for their design and deployment. Our key contributions are:

- We question the current evaluation paradigm of LRMs on established math benchmarks and design a controlled experimental testbed by leveraging algorithmic puzzle environments that enable controllable experimentation with respect to problem complexity.
- We show that state-of-the-art LRMs (e.g., o3-mini, DeepSeek-R1, Claude-3.7-Sonnet-Thinking) still fail to develop generalizable problem-solving capabilities, with accuracy ultimately collapsing to zero beyond certain complexities across different environments.
- We find that there exists a scaling limit in the LRMs' reasoning effort with respect to problem complexity, evidenced by the counterintuitive decreasing trend in the thinking tokens after a complexity point.
- We question the current evaluation paradigm based on final accuracy and extend our evaluation to intermediate solutions of thinking traces with the help of deterministic puzzle simulators. Our analysis reveals that as problem complexity increases, correct solutions systematically emerge at later positions in thinking compared to incorrect ones, providing quantitative insights into the self-correction mechanisms within LRMs.
- We uncover surprising limitations in LRMs' ability to perform exact computation, including their failure to benefit from explicit algorithms and their inconsistent reasoning across puzzle types.