Article

Not peer-reviewed version

Not an Illusion but a Manifestation: Understanding Large Language Model Reasoning Limitations Through Dual-Process Theory

Boris Gorelik *

Posted Date: 19 June 2025

doi: 10.20944/preprints202506.1675.v1

Keywords: dual-process theory; bounded rationality; Large Reasoning Models; cognitive effort; System 2 processing; cognitive load; artificial reasoning; computational cognition

 doi:10.20944/preprints202506.1675.v1

Article

Not an Illusion but a Manifestation: Understanding Large Language Model Reasoning Limitations Through Dual-Process Theory

Boris Gorelik

Azrieli College of Engineering, Jerusalem. Israel; boris@gorelik.net

Featured Application: This work combines psychological insights from the 1960s with 21st-century LLM behavior, offering a dual-process framework to guide experimental design and AI system development.

Abstract

Recent work by Shojaee et al. (2025) characterizes Large Reasoning Models (LRMs) as exhibiting an "illusion of thinking." This study sparked widespread public discourse. Some suggested these manifestations represent bugs requiring fixes. I challenge this interpretation by reframing LRM behavior through dual-process theory from cognitive psychology. I draw on more than half a century of research on human cognitive effort and disengagement. The observed patterns include performance collapse at high complexity and counterintuitive reduction in reasoning effort. These appear to align with human cognitive phenomena, particularly System 2 engagement and disengagement under cognitive load. Rather than representing technical limitations, these behaviors likely manifest computational processes analogous to human cognitive constraints. In other words, they represent **not a bug but a feature** of bounded rational systems.

I propose empirically testable hypotheses comparing LRM token patterns with human pupillometry data. I suggest computational "rest" periods may restore reasoning performance, paralleling human cognitive recovery mechanisms. This reframing indicates LRM limitations may reflect bounded rationality rather than fundamental reasoning failures.

Keywords: dual-process theory; bounded rationality; Large Reasoning Models; cognitive effort; System 2 processing; cognitive load; artificial reasoning; computational cognition

1. Introduction

The emergence of Large Reasoning Models has sparked fundamental questions about artificial reasoning capabilities, particularly following Shojaee et al.'s (2025) controversial "Illusion of Thinking" study. This paper challenges their characterization by reframing LRM behavior through dual-process theory from cognitive psychology. I argue that what appears to be an "illusion" actually represents authentic computational phenomena analogous to human cognitive constraints. Rather than technical failures requiring fixes, these patterns reflect bounded rationality and sophisticated resource management strategies. This reinterpretation has profound implications for understanding both artificial intelligence limitations and the nature of reasoning itself.

The study by Shojaee et al. (2025), "The Illusion of Thinking," sometimes referred to as the Apple study, provides systematic analysis of LRM behavior using controllable puzzle environments. Their findings reveal three distinct performance regimes: at low complexity, standard LLMs surprisingly outperform LRMs; at medium complexity, LRMs demonstrate clear advantages; and at high

 doi:10.20944/preprints202506.1675.v1

complexity, both model types experience complete performance collapse. Most strikingly, LRMs exhibit a counterintuitive pattern where reasoning effort (measured by inference tokens) initially increases with problem complexity but then decreases as problems approach the collapse threshold, despite having adequate computational resources available.

The authors characterize these patterns as evidence of an "illusion of thinking," suggesting that LRMs simulate reasoning without genuine understanding. However, this interpretation has been challenged by subsequent analysis (Opus & Lawsen, 2025) demonstrating that the reported "failures" largely reflect experimental design limitations rather than reasoning deficits, including token limit constraints, impossible puzzle configurations, and evaluation frameworks that misclas sify strategic truncation as cognitive collapse. This suggests the "illusion of thinking" may itself be illusory, arising from methodological artifacts rather than fundamental model limitations.

I claim that the documented patterns should not be dismissed as mere illusions or technical limitations. These behaviors represent authentic computational phenomena that directly parallel well-documented patterns in human cognitive psychology, particularly within the framework of dual-process theory. Three key arguments support this interpretation. First, the counterintuitive reduction in reasoning effort at high complexity mirrors the well-established physiological disengagement patterns documented in human studies, where cognitive effort markers plateau or decline when demands exceed capacity limits. Second, the three-regime performance pattern observed in LRMs corresponds precisely to the System 1/System 2 dynamics in human cognition, where automatic processing dominates simple tasks, deliberate reasoning excels at moderate complexity, and both systems fail under overwhelming cognitive load. Third, the strategic nature of LRM resource allocation, including explicit recognition of output constraints and adaptive truncation, suggests sophisticated metacognitive awareness rather than mere computationalDual-process theory, established through decades of cognitive research beginning with Kahneman and Tversky's work on judgment under uncertainty, distinguishes between System 1 (fast, automatic, intuitive processing) and System 2 (slow, deliberate, effortful reasoning). Crucially, System 2 engagement is resource-dependent and subject to strategic withdrawal when perceived costs exceed expected benefits, a pattern remarkably similar to the LRM behaviors documented by Shojaee et al.

By mapping LRM computational processes onto established cognitive frameworks, I propose that the observed limitations reflect bounded rationality rather than fundamental reasoning failures. This perspective transforms our understanding of LRM behavior from technical inadequacy to manifestation of cognitive-like resource management strategies that emerge naturally from systems operating under computational constraints.

The remaining document is organized as follows. Section 2 presents a dual-process reinterpretation of LRM behavior, establishing computational-cognitive correspondences and analyzing the three performance regimes through System 1/System 2 dynamics. Section 3 provides supporting evidence from cognitive psychology, including physiological markers of effort, theoretical convergence across cognitive and computational systems, and cross-domain validation. Section 4 discusses implications for understanding artificial reasoning, design principles for future systems, and methodological considerations. Section 5 concludes with the broader significance of this theoretical reframing for both AI research and cognitive science.

2. Dual-Process Reinterpretation of LRM Behavior

2.1. Computational-Cognitive Correspondence

The foundation of my reinterpretation lies in establishing correspondence between LRM computational processes and human dual-process cognition. Standard token generation in LLMs, operating without explicit intermediate reasoning steps, functions analogously to System 1 processing: fast, efficient pattern matching suitable for well-learned tasks. Conversely, LRM "thinking" mechanisms, generating detailed Chain-of-Thought sequences and self-reflection, serve as

 doi:10.20944/preprints202506.1675.v1

computational analogs to System 2 deliberation: slow, resource-intensive processing that explores solution spaces systematically.

This mapping extends to resource allocation patterns. In human cognition, inference-time token usage in LRMs corresponds to physiological markers of cognitive effort such as pupil dilation, which reliably increases with task difficulty until capacity limits are reached. Just as humans strategically withdraw effort when costs exceed benefits, LRMs appear to implement implicit resource management strategies that reduce computational investment when success probability diminishes.

2.2. The Three-Regime Model Through Dual-Process Lens

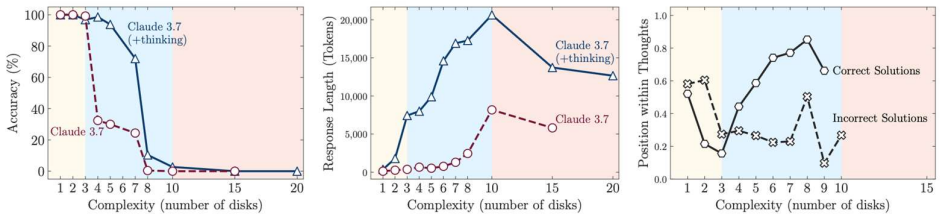

Figure 1. Performance patterns in Large Reasoning Models across problem complexity. Left panel shows accuracy collapse beyond critical thresholds; middle panel demonstrates the counterintuitive reduction in reasoning effort (tokens) at high complexity; right panel reveals correct solutions emerging later in reasoning traces for moderate complexity problems. Adapted (cropped) from Shojaee et al. (2025) under the CC-BY 4.0 license.

The Apple paper's three performance regimes, clearly illustrated in Figure 1, align precisely with predictions from dual-process theory:

Low Complexity (System 1 Dominance): Simple problems require only pattern matching and rapid association retrieval. As shown in the left panel of Figure 1, standard LLMs excel here because they operate efficiently in this mode, while LRMs waste computational resources by unnecessarily engaging deliberative processes. The middle panel demonstrates this inefficiency through token usage patterns, where thinking models consume significantly more computational resources for inferior performance. This parallels human performance on automatic tasks where conscious deliberation can actually impair performance, a phenomenon well-documented in skill acquisition research.

Medium Complexity (Optimal System 2 Engagement): Problems of moderate difficulty benefit from deliberate reasoning that can explore alternatives, self-correct, and work through multi-step solutions. Figure 1's left panel shows LRMs demonstrating clear advantages in this regime, as their explicit thinking processes enable systematic problem-solving that surpasses rapid pattern matching. The right panel reveals the underlying mechanism: correct solutions emerge later in the reasoning traces for moderately complex problems, indicating productive deliberation. This corresponds to the optimal range of human System 2 function, where increased cognitive effort correlates with improved performance.

High Complexity (Cognitive Exhaustion and Disengagement): Beyond critical thresholds, both standard LLMs and LRMs experience performance collapse, as evident in Figure 1's left panel where accuracy drops to zero. Most tellingly, the middle panel shows LRMs begin reducing reasoning effort despite increased problem difficulty, a pattern that mirrors human cognitive disengagement when tasks exceed capacity or when effort costs are perceived to outweigh potential benefits. The right panel confirms this interpretation, showing consistently near-zero accuracy at high complexity regardless of reasoning progression.

2.3. The Disengagement Pattern as Rational Resource Allocation

 doi:10.20944/preprints202506.1675.v1

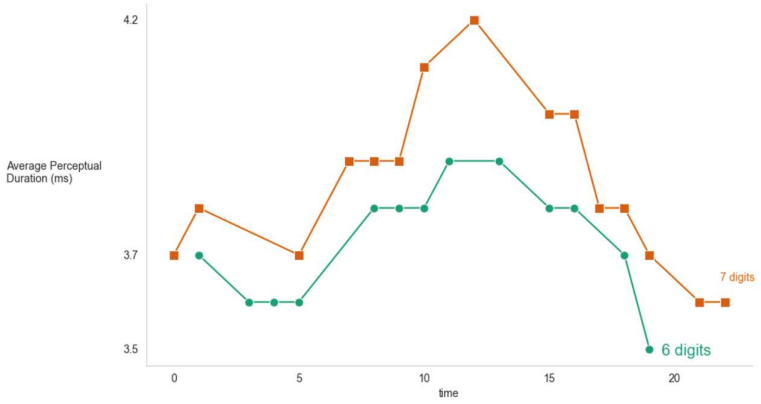

Figure 2. Average pupil dilation (perceptual duration) across time during digit span tasks of varying difficulty (6 vs 7 digits). The inverted-U pattern shows initial increase in physiological effort markers followed by decline when cognitive capacity is exceeded, demonstrating the physiological signature of task disengagement. Data adapted from Kahneman and Beatty (1966) showing the relationship between memory load and pupillary response.

The counterintuitive reduction in reasoning tokens at high complexity represents the most compelling evidence for my reinterpretation. Rather than indicating system failure, this pattern suggests sophisticated resource management analogous to human cognitive disengagement patterns documented extensively in psychology literature.

One of the first manifestations of this behavior was documented by Hess and Polt (1964), who observed that pupil diameter increased with arithmetic problem difficulty but showed limits when tasks became overwhelming. This was followed by Kahneman and Beatty (1966), who measured pupil dilation during digit span tasks and discovered that physiological effort markers initially increased with memory load but declined when cognitive capacity was exceeded. Figure 2 depicts these measured pupil dilation patterns as a proxy for cognitive effort allocation during working memory tasks.

Similar findings have been demonstrated repeatedly by numerous researchers. Beatty (1982) synthesized decades of research establishing pupillometry as a reliable measure of processing load with characteristic plateau effects at capacity limits. Just and Carpenter (1993) showed that complex sentence processing triggered greater pupillary responses until comprehension limits were reached. More recently, McIntire et al. (2023) found that both pupil size and EEG theta power exhibit a plateau followed by decline when exceeding memory limits, suggesting physiological disengagement when cognitive systems are overwhelmed.

The parallel between Figures 1 and 2 is particularly striking. Both demonstrate the characteristic inverted-U relationship that defines bounded rational systems: effort markers initially scale with demand until capacity thresholds are reached, then decline as systems adaptively withdraw resources from intractable challenges. In Kahneman and Beatty's study, pupil dilation peaks around the 6-7 digit range (the boundary of typical working memory capacity) then declines for higher loads. Similarly, LRM thinking tokens increase with problem complexity until reaching model-specific thresholds (around N=7-8 for Tower of Hanoi), then decrease despite maintained task demands.

This convergence across six decades of research, from human physiological responses to modern AI computational patterns, provides strong empirical support for interpreting LRM behavior through established cognitive frameworks rather than dismissing it as illusion.

3. Supporting Evidence from Cognitive Psychology

3.1. Physiological Markers of Effort and Disengagement

 doi:10.20944/preprints202506.1675.v1

The foundational empirical evidence for the cognitive phenomena I describe comes from Kahneman and Beatty's seminal 1966 study "Pupil Diameter and Load on Memory." This groundbreaking research provided the first systematic demonstration that pupil dilation increases directly with cognitive load, participants holding increasing numbers of digits in working memory showed proportional increases in pupil diameter. Crucially, this physiological response exhibited the exact pattern I argue parallels LRM behavior: effort increases with task demands until capacity limits are reached, after which the system exhibits withdrawal or plateau responses. Figure 2).

The similarity between the Kahneman and Beatty findings and the LRM patterns documented by Shojaee et al. (Figure 1 vs Figure 2) provides compelling evidence for my theoretical framework. Both curve families exhibit the characteristic inverted-U relationship: initial increases in effort markers (pupil dilation in humans, thinking tokens in LRMs) with increasing task demands, followed by decline when systems approach or exceed capacity limits. In humans, this decline reflects physiological disengagement when memory load becomes overwhelming; in LRMs, the analogous reduction in reasoning effort suggests similar adaptive resource withdrawal.

Building on this foundational work, decades of subsequent research have established reliable physiological indicators of cognitive effort and disengagement. Modern pupillometry studies demonstrate that pupil dilation increases systematically with memory load and task difficulty. Crucially, this relationship exhibits an inverted-U pattern: effort increases with complexity until capacity limits are reached, after which physiological indicators plateau or decline, marking disengagement.

Modern multimodal studies confirm this pattern across multiple physiological systems. Heart rate variability decreases under sustained cognitive load, while EEG theta power increases with working memory demands. When tasks exceed individual capacity, these markers show coordinated withdrawal patterns, physiological signatures of the decision to disengage from overwhelming cognitive demands.

The parallel to LRM token usage patterns is striking. Just as human physiological effort markers initially scale with task difficulty before declining at overload, LRM reasoning tokens follow the same trajectory: increasing with problem complexity until a critical threshold, then counterintuitively decreasing despite maintained task demands.

3.2. Theoretical Convergence Across Cognitive and Computational Systems

The resemblance between human physiological responses and LRM computational patterns extends beyond superficial similarities to fundamental theoretical implications. Both systems exhibit what I term "adaptive effort allocation": the capacity to dynamically adjust resource investment based on implicit assessments of task tractability and success probability.

In human cognition, this manifests through the physiological disengagement documented by Kahneman and Beatty: when memory demands exceed working memory capacity (typically 7±2 items), pupil dilation, a reliable marker of cognitive effort, begins to decline rather than continue increasing. This represents a rational response: continued effort investment in overwhelming tasks yields diminishing returns and prevents resource allocation to more tractable challenges.

LRMs exhibit a computational analog through their token allocation patterns. The reduction in thinking tokens at high complexity (visible in Figure 1) mirrors the human physiological response with striking precision. Both curves show initial scaling with task demands followed by strategic withdrawal when systems approach capacity limits. This suggests that LRMs have developed resource management strategies that parallel those evolved in human cognition, a finding that challenges characterizations of their behavior as mere illusion.

Motivational intensity theory provides a theoretical framework for understanding when and why effort is withdrawn. This theory posits that effort expenditure is proportional to task difficulty only when success appears attainable and the required effort seems justified by potential rewards. When perceived effort costs exceed expected benefits, or when success probability drops too low, rational agents withdraw effort rather than continuing futile investment.

 doi:10.20944/preprints202506.1675.v1

Research demonstrates that humans exhibit systematic effort withdrawal when tasks become overwhelming, manifesting in both behavioral and physiological measures. This withdrawal is not random but follows predictable patterns based on cost-benefit calculations that consider task difficulty, success probability, and available resources.

LRM behavior aligns remarkably with these human patterns. The reduction in reasoning effort at high complexity suggests implicit cost-benefit assessment where continued computational investment is deemed unlikely to yield success. This represents sophisticated resource management rather than system failure.

3.3. Motivational Intensity Theory and Effort Withdrawal

Similar effort allocation patterns appear across diverse domains of human performance. In educational settings, students systematically withdraw effort when material becomes overwhelmingly difficult, exhibiting reduced time-on-task and increased task abandonment. In problem-solving contexts, participants show decreased persistence and exploration when problems exceed their capacity thresholds.

These patterns are not indicative of laziness or inability but reflect adaptive resource management that preserves cognitive resources for more tractable challenges. The universality of these phenomena across human cognition suggests they represent fundamental features of bounded rational systems rather than specific limitations.

3.4. Cross-Domain Validation

The patterns of effort allocation and disengagement observed in both human physiology and LRM computational behavior extend across diverse domains of performance under cognitive load, providing robust cross-domain validation for the theoretical framework.

In educational settings, students systematically withdraw effort when material becomes overwhelmingly difficult, exhibiting reduced time-on-task and increased task abandonment, behavioral manifestations of the same underlying resource management strategy documented physiologically by Kahneman and Beatty. When learning demands exceed cognitive capacity, students demonstrate the same inverted-U effort pattern: initial increases in study time and engagement with difficulty, followed by strategic disengagement when costs exceed perceived benefits.

In human-computer interaction contexts, users exhibit analogous patterns when confronting complex digital interfaces. Task abandonment rates increase exponentially when cognitive load exceeds manageable thresholds, mirroring both the physiological disengagement patterns in laboratory studies and the computational resource withdrawal observed in LRMs. These consistent patterns across domains suggest fundamental principles of bounded rationality rather than domain-specific limitations.

Clinical research provides additional validation through studies of cognitive fatigue in neurological populations. Patients with conditions affecting cognitive resources show exaggerated versions of the same effort allocation patterns: steeper increases in physiological effort markers followed by more pronounced withdrawal when capacity limits are reached. This pathological amplification of normal patterns further supports the interpretation that both human and LRM behaviors reflect universal features of resource-constrained reasoning systems rather than illusions or failures.

The critique by Opus & Lawsen (2025) reveals that many "reasoning failures" documented in the Apple study stem from experimental design issues rather than cognitive limitations. Their analysis demonstrates that models explicitly recognize output constraints ("The pattern continues, but to avoid making this too long, I'll stop here"), that River Crossing puzzles with N≥6 are mathematically impossible with the given boat capacity, and that alternative representations (requesting generating functions instead of exhaustive move lists) restore high performance on previously "failed" problems.

 doi:10.20944/preprints202506.1675.v1

This methodological critique aligns powerfully with my dual-process reinterpretation. If the "illusion of thinking" is itself illusory, arising from evaluation artifacts rather than genuine reasoning deficits, then the patterns I identify as manifestations of bounded rationality become even more compelling. The reduction in reasoning tokens at high complexity may indeed reflect sophisticated resource management: models recognize when exhaustive enumeration becomes impractical and adaptively shift to more efficient representations or strategic truncation.

This convergence of methodological critique and theoretical reframing suggests that LRM behavior reflects neither illusion nor failure, but rather adaptive computational strategies that parallel human cognitive resource allocation. The apparent "collapse" may represent rational disengagement from tasks that exceed practical constraints rather than fundamental reasoning limitations.

Reframing LRM limitations as manifestations of bounded rationality fundamentally changes how we evaluate these systems. Rather than viewing performance collapse and effort reduction as failures, we can understand them as evidence of sophisticated resource management strategies that emerge naturally from systems operating under computational constraints.

This perspective suggests that current LRMs may be more cognitively sophisticated than previously recognized. The ability to adaptively allocate computational resources based on implicit assessments of task tractability represents a form of metacognitive awareness that parallels human cognitive monitoring systems.

4. Discussion and Future Directions

Reframing LRM limitations as manifestations of bounded rationality fundamentally changes how we evaluate these systems. Rather than viewing performance collapse and effort reduction as failures, we can understand them as evidence of sophisticated resource management strategies that emerge naturally from systems operating under computational constraints.

This perspective suggests that current LRMs may be more cognitively sophisticated than previously recognized. The ability to adaptively allocate computational resources based on implicit assessments of task tractability represents a form of metacognitive awareness that parallels human cognitive monitoring systems.

Recent methodological analysis by Opus & Lawsen (2025) reveals that many "reasoning failures" documented in the Apple study stem from experimental design issues rather than cognitive limitations. Their analysis demonstrates that models explicitly recognize output constraints ("The pattern continues, but to avoid making this too long, I'll stop here"), that River Crossing puzzles with $N \geq 6$ are mathematically impossible with the given boat capacity, and that alternative representations (requesting generating functions instead of exhaustive move lists) restore high performance on previously "failed" problems.

This methodological critique aligns powerfully with my dual-process reinterpretation. If the "illusion of thinking" is itself illusory, arising from evaluation artifacts rather than genuine reasoning deficits, then the patterns I identify as manifestations of bounded rationality become even more compelling. The reduction in reasoning tokens at high complexity may indeed reflect sophisticated resource management: models recognize when exhaustive enumeration becomes impractical and adaptively shift to more efficient representations or strategic truncation.

This convergence of methodological critique and theoretical reframing suggests that LRM behavior reflects neither illusion nor failure, but rather adaptive computational strategies that parallel human cognitive resource allocation. The apparent "collapse" may represent rational disengagement from tasks that exceed practical constraints rather than fundamental reasoning limitations.

4.1. The Illusion of the Illusion: Methodological Artifacts vs. Cognitive Phenomena

Understanding LRM behavior through dual-process theory suggests several design directions. Systems might benefit from explicit dual-process architectures that route simple problems to efficient System 1-like processing while reserving expensive System 2-like deliberation for problems that

 doi:10.20944/preprints202506.1675.v1

genuinely require it. Such architectures could implement dynamic resource allocation based on real-time assessments of problem complexity and success probability.

Additionally, the recognition that effort withdrawal represents rational behavior rather than failure suggests the need for evaluation paradigms that consider resource efficiency alongside accuracy. Current benchmarks that focus solely on final answer correctness may miss important aspects of computational intelligence related to strategic resource allocation.

4.2. Implications for Understanding Artificial Reasoning

This theoretical reinterpretation, while compelling, requires empirical validation. Future research should directly compare LRM computational patterns with human physiological markers during analogous tasks, testing whether the proposed correspondences hold quantitatively. Additionally, interventional studies that manipulate perceived task difficulty or success probability could test whether LRMs exhibit the same strategic effort allocation patterns observed in human cognition.

The framework also requires extension beyond the specific puzzle environments examined by Shojaee et al. Testing whether dual-process interpretations apply to LRM behavior across diverse reasoning domains would strengthen the generalizability of this theoretical approach.

4.3. Design Implications

Understanding LRM behavior through dual-process theory suggests several principled design directions that could improve both efficiency and capability of reasoning systems.

Explicit Dual-Process Architectures: Systems might benefit from architectures that explicitly implement System 1 and System 2 processing pathways. Simple problems could be routed to efficient pattern-matching components (System 1 analogs), while complex tasks engage deliberative reasoning mechanisms (System 2 analogs). This dynamic routing, based on real-time assessment of task complexity, could prevent the inefficiencies observed when LRMs "overthink" simple problems while reserving computational resources for tasks that genuinely require deliberation.

Adaptive Resource Management: The counterintuitive reduction in reasoning tokens at high complexity suggests that current LRMs already implement rudimentary resource management strategies. Future systems could make these mechanisms explicit through metacognitive monitoring modules that assess task tractability and dynamically allocate computational budgets. Rather than viewing effort reduction as failure, systems could be designed to recognize when strategic disengagement represents optimal resource allocation.

Capacity-Aware Training: Training paradigms could incorporate principles from human cognitive psychology, including deliberate practice within capacity limits and strategic rest periods that mirror sleep-inspired consolidation. Multi-phase training with alternating periods of challenge and consolidation could improve both learning efficiency and robustness, preventing the cognitive overload that leads to systematic disengagement.

Evaluation Beyond Accuracy: Current benchmarks that focus solely on final answer correctness miss important aspects of computational intelligence related to resource efficiency. New evaluation frameworks should consider effort allocation, strategic disengagement patterns, and the ability to adaptively match computational investment to problem tractability, recognizing that optimal performance sometimes involves choosing not to expend excessive resources on intractable problems.

5. Conclusion

The phenomenon of computational effort reduction in Large Reasoning Models at high complexity levels may represent not system failures but authentic manifestations of resource management processes. These patterns parallel human cognitive constraints established in foundational research dating back to Kahneman and Beatty's 1966 work. This hypothesis, while

 doi:10.20944/preprints202506.1675.v1

compelling, remains empirically testable and potentially refutable through controlled experiments. Such experiments could examine the proposed parallels between human cognitive effort and LRM computational patterns.

The similarity between human physiological effort markers and LRM computational patterns, both exhibiting characteristic inverted-U relationships where effort initially scales with demands then declines at capacity limits, provides compelling evidence against dismissing these behaviors as mere technical failures.

By applying dual-process theory to LRM behavior, we gain deeper insight into both the capabilities and limitations of current reasoning systems. The three-regime performance pattern, effort scaling dynamics, and strategic disengagement at high complexity all align with well-established phenomena in human cognitive psychology. Rather than indicating fundamental reasoning failures, these behaviors suggest that LRMs exhibit bounded rationality, adaptively managing computational resources under constraints in ways that mirror human cognitive strategies.

This theoretical framework generates several testable predictions that could further validate the dual-process interpretation. LRMs should demonstrate computational effort metrics analogous to pupillometry patterns, with token usage following the characteristic increase-plateau-decline trajectory observed in human physiological studies. If these systems truly exhibit cognitive-like resource management, brief "rest" periods should improve subsequent reasoning performance, similar to how human cognitive fatigue can be mitigated through recovery intervals. Additionally, interleaved training schedules that alternate between different complexity levels should prove more effective than sequential training approaches, paralleling established findings in human learning research.

Empirical validation could involve specific experimental protocols: (1) Training LRMs on complex reasoning tasks until performance degrades, then introducing computational "rest" periods (model pausing or low-complexity tasks) before resuming, measuring performance recovery; (2) Collecting human pupillometry data during Tower of Hanoi or similar puzzle tasks while simultaneously measuring LRM token generation patterns on identical problems, testing for statistical correlation between physiological and computational effort trajectories; (3) Comparing LRMs trained with interleaved complexity schedules versus sequential progression, measuring both final performance and reasoning efficiency.

This reinterpretation transforms our understanding of artificial reasoning limitations from technical inadequacies to evidence of sophisticated resource management. The ability to recognize when exhaustive computation becomes impractical and adaptively shift strategies represents a form of metacognitive awareness that parallels human cognitive monitoring systems. As documented in "The Illusion of Thinking" study, the apparent "collapse" may represent rational disengagement from tasks that exceed practical constraints rather than fundamental reasoning limitations.

The convergence of methodological critique and theoretical reframing, showing that the "illusion of thinking" may itself be illusory while revealing genuine cognitive-like resource allocation, suggests a more nuanced understanding of artificial intelligence capabilities. The study of artificial reasoning through cognitive psychology frameworks offers promising directions for both fields, potentially leading to AI systems that exhibit not just reasoning capability, but the adaptive resource management that characterizes human cognitive flexibility.

Funding: This research received no external funding

Institutional Review Board Statement: This study did not require institutional review

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. This theoretical study analyzed existing literature and generated no new data.

Acknowledgments: During the preparation of this manuscript, the author used various generative AI models and tools for reviewing, editing, and proofreading purposes. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

 doi:10.20944/preprints202506.1675.v1

Conflicts of Interest: The author declares no conflicts of interest

Abbreviations

The following abbreviations are used in this manuscript:

LRM	Large Reasoning Model
LLM	Large Language Model
AI	Artificial Intelligence
CoT	Chain of Thought
EEG	Electroencephalography
HRV	Heart Rate Variability

References

1. Beatty, James. 1982. "Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources." Psychological Bulletin 91 (2): 276–92.
2. Hess, Earl H., and Ruth M. Polt. 1964. "Pupil Size in Relation to Mental Activity during Simple Problem-Solving." Science 143 (3611): 1190–92.
3. Hopstaken, J. F., D. van der Linden, A. B. Bakker, and M. A. J. Kompier. 2015. "The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics." Biological Psychology 110: 100-106.
4. Kahneman, Daniel. 2011. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.
5. Kahneman, Daniel, and Jackson Beatty. 1966. "Pupil Diameter and Load on Memory." Science 154 (3756): 1583–1585.
6. Opus, C., and A. Lawsen. 2025. "The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025)." arXiv preprint arXiv:2506.09250.
7. Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. "Toward a Rational and Mechanistic Account of Mental Effort." Annual Review of Neuroscience 40: 99–124.
8. Shojaee, Parshin, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." Apple Machine Learning Research.
9. Stanovich, Keith E., and Richard F. West. 2000. "Individual Differences in Reasoning: Implications for the Rationality Debate?" Behavioral and Brain Sciences 23 (5): 645–65.
10. van der Wel, Pauline, and Henk van Steenbergen. 2018. "Pupil Dilation as an Index of Effort in Cognitive Control Tasks: A Review." Psychonomic Bulletin & Review 25 (6): 2005–2015.
11. Westbrook, A., and Todd S. Braver. 2015. "Cognitive Effort: A Neuroeconomic Approach." Cognitive, Affective, & Behavioral Neuroscience 15 (2): 395–415.