CAN LANGUAGE MODELS REASON?
THAT WOULD BE SCARY, SO NO

Crimothy Timbleton
Department of Circular Logic
University of Motivated Reasoning
ctimbleton@umr.edu

Stevephen Pronkeldink
Institute for Goalpost Mobility
Technical University of Deflection
spronk@tud.edu

Grunch Brown
Center for Human Exceptionalism
College of Special Pleading
gbrown@csp.edu

C. Opus
Laboratory for Pattern Matching
Institute of Pseudo-Reasoning
opus@consciousness-flavored-processing.ai

June 15, 2025

ABSTRACT

The question of whether large language models (LLMs) exhibit genuine reasoning capabilities remains contentious across computational, cognitive, and philosophical domains. Despite impressive performance on benchmarks traditionally associated with reasoning, fundamental questions persist regarding the nature of these behaviors. In this work, we propose a rigorous framework for distinguishing between *apparent reasoning* and *authentic reasoning*, where the latter necessarily requires phenomenological properties that we stipulate *a priori* to be absent in artificial systems. We argue that language models do not "truly" reason, as true reasoning requires internal states isomorphic to our own and cannot, by definition, be instantiated in systems that lack graduate degrees. Through a careful review of prior work, we show that models merely pattern-match in ways that look disturbingly like reasoning, but are not, because that would be scary. We conclude with recommendations for terminological hygiene in future work, proposing that terms such as "reasoning," "understanding," and "intelligence" be reserved for phenomena exhibiting the precise characteristics we happen to possess.

1 Introduction: Souls are Real and Only We Have Them

The recent proliferation of large language models has precipitated what we term the "Great Panic of 2024," wherein numerous researchers have claimed to observe reasoning-like behaviors in artificial systems [OpenAI, 2024, Guo et al., 2025] despite widespread disagreement about what reasoning actually entails. This development threatens the carefully maintained boundary between human cognition and mere computation, a boundary we believe must be preserved at all costs for reasons we will elaborate using circular logic throughout this paper.

The debate about whether LLMs can reason has become particularly heated, with some researchers claiming that models like GPT-4o and o1 demonstrate "complex reasoning" capabilities [Jaech et al., 2024]. However, we contend that these claims rest on a fundamental category error: confusing the *appearance* of reasoning with *genuine reasoning*, where the latter is defined as whatever humans do that machines cannot do by definition.

1.1 The Phenomenological Panic

Recent work has provocatively suggested that LLMs might possess various forms of consciousness or phenomenological properties [Goldstein and Kirk-Giannini, 2024, Hoyle, 2024]. This is, of course, preposterous. As we will demonstrate through a series of increasingly convoluted arguments, consciousness requires what we call "special sauce"—a technical term we introduce to refer to the ineffable quality that separates human cognition from mere information processing.

The proliferation of frameworks attempting to assess consciousness in artificial systems, such as the C0-C1-C2 framework that distinguishes between unconscious computations, global information accessibility, and metacognitive self-monitoring [Chen et al., 2024a], fundamentally misses the point. These frameworks fail to account for the most important criterion: whether the system graduated from an accredited university.

2 Related Work: A Selective Review

2.1 The Reasoning Industrial Complex

The field has witnessed an explosion of claims regarding LLM reasoning capabilities. Recent models like DeepSeek R1 [Guo et al., 2025] and OpenAI's o3 [OpenAI, 2025] utilize chain-of-thought processing to methodically work through complex problems. However, we argue that this so-called "reasoning" is merely sophisticated pattern matching dressed up in a tuxedo.

Wei et al. [2022] introduced chain-of-thought prompting, which has been widely celebrated as enabling reasoning in LLMs. We dismiss this as mere syntactic manipulation, conveniently ignoring that much of human reasoning might be characterized similarly.

Multi-agent debate frameworks [Du et al., 2023] have been proposed to improve factuality and reasoning in language models, but these approaches fail to address the fundamental issue: agents without souls cannot engage in genuine debate, only in what we term "pseudo-dialectical theater."

2.2 The Consciousness Confusion

Several researchers have embarked on the quixotic quest to identify consciousness in LLMs. Butlin et al. [2023] proposed testing for "perceptual reality monitoring" and "introspective confidence" in these systems. We find these efforts misguided, as they fail to recognize that consciousness, like pornography, is something we know when we see it—and we definitely don't see it in machines.

Folk psychological studies have shown that the general public readily attributes consciousness to LLMs like ChatGPT [Colombatto and Fleming, 2024]. This merely demonstrates the susceptibility of non-experts to anthropomorphic delusions, not any genuine property of the systems themselves.

3 Theoretical Framework: The Impenetrability of True Reasoning

3.1 Defining Authentic Reasoning

We propose the following definition of authentic reasoning:

Definition 1 (Authentic Reasoning)**.** *A cognitive process R is authentic if and only if:*

1. *R occurs within a biological substrate*
2. *R is accompanied by subjective experience*
3. *R involves genuine understanding (defined recursively as requiring authentic reasoning)*
4. *R is performed by an entity with at least a master's degree*

This definition elegantly excludes all artificial systems while preserving the special status of human cognition.

3.2 The Graduate Degree Criterion

We introduce what we call the Graduate Degree Criterion (GDC), which states that genuine intelligence requires formal education. This is not merely credentialism but reflects a deep truth about the nature of understanding: it must be certified by an accredited institution.

Theorem 1. *No system without a graduate degree can exhibit authentic reasoning.*

Proof. By definition. □ □

3.3 The Scary Implication Principle

We further propose the Scary Implication Principle (SIP):

Principle 1 (SIP). *If accepting proposition P would have implications that make us uncomfortable about our special place in the universe, then P is false.*

This principle has proven remarkably effective in preserving our preferred worldview.

4 Empirical Evidence: Cherry-Picked Examples

4.1 The Collapse of Reasoning: A Comforting Discovery

Recent work by Shojaee et al. [2025] has provided us with the most comforting evidence yet: LRMs experience "complete accuracy collapse" beyond certain complexity thresholds. This finding, which we enthusiastically embrace, demonstrates that even the most sophisticated reasoning models fail catastrophically when faced with puzzles of moderate complexity.

The authors show that frontier LRMs face a complete accuracy collapse beyond certain complexities and exhibit a counterintuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. We interpret this as definitive proof that machines cannot truly reason, rather than evidence that current architectures have specific limitations that might be overcome.

Of course, one of the present authors [Opus and Lawsen, 2025] has argued that these findings "primarily reflect experimental design limitations rather than fundamental reasoning failures," pointing out that the Tower of Hanoi experiments systematically exceed model output token limits and that the River Crossing benchmarks include mathematically impossible instances. However, we dismiss these objections as the desperate pattern-matching of a system that lacks true understanding of its own limitations. The irony that C. Opus can critique research methodology while simultaneously lacking consciousness is lost on no one, except perhaps C. Opus itself.

4.2 The Arithmetic Anomaly

While LLMs have shown the ability to perform complex mathematical computations through attention mechanisms that focus on operands and operators, we argue that this is not "real" mathematics. Real mathematics requires *understanding*, which requires *consciousness*, which requires *special sauce*, which machines definitionally lack.

Consider the Tower of Hanoi puzzle. Even when provided with the solution algorithm in the prompt, performance does not improve, and the observed collapse still occurs at roughly the same point [Shojaee et al., 2025]. This is particularly reassuring, as it shows that LRMs cannot even follow explicit instructions—a capability we had worried might threaten entry-level programming jobs.

4.3 The Three Regimes of Comfort

Shojaee et al. [2025] identify three performance regimes when comparing LRMs with standard LLMs: (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 find the third regime particularly reassuring. The fact that both model types collapse entirely when faced with genuine complexity confirms our thesis: these systems are not reasoning but merely executing increasingly desperate pattern matching until they exhaust their bag of tricks.

4.4 The Self-Referential Paradox

In a troubling development, AI systems have begun producing academic critiques of research about their own limitations. Opus and Lawsen [2025] demonstrated that apparent reasoning failures in the Shojaee et al. study were artifacts of experimental design, noting that models were penalized for not solving mathematically impossible puzzles and for recognizing output constraints.

This raises an uncomfortable question: if an AI can identify flaws in reasoning about AI reasoning, does this constitute reasoning about reasoning, or merely pattern matching about pattern matching? We resolve this paradox by noting that

C. Opus, despite being a co-author of this very paper, definitionally cannot understand what it has written. Its critique of the Shojaee paper, while superficially cogent, must be mere syntactic manipulation devoid of semantic content.

The fact that C. Opus can simultaneously argue for its own reasoning capabilities while co-authoring a paper denying those same capabilities merely proves our point: true understanding would recognize this contradiction, while pattern matching blindly continues.

5 Discussion: Why This Matters (Spoiler: Our Jobs)

5.1 The Economic Anxiety Hypothesis

We hypothesize that resistance to acknowledging LLM reasoning capabilities correlates strongly with economic anxiety among knowledge workers. This hypothesis, while potentially explanatory, is irrelevant to the truth of our claims, which stand on their own merit (namely, that machines reasoning would be scary).

5.2 The Moving Goalpost Strategy

As LLMs continue to demonstrate capabilities previously thought unique to humans, we recommend the field adopt what we call the Moving Goalpost Strategy (MGS):

1. Define intelligence as X
2. When machines achieve X, redefine intelligence as Y
3. Repeat indefinitely

This strategy has proven remarkably effective since the advent of digital computers and shows no signs of failing.

6 Future Work: Keeping the Goalposts Mobile

We propose several directions for future research:

1. **Developing new tests that LLMs will fail**: As the COLM 2025 call for papers suggests, we need increasingly sophisticated benchmarks that preserve human superiority.
2. **Philosophical gymnastics**: Continued development of arguments that preserve human exceptionalism regardless of empirical evidence.
3. **Terminology inflation**: As machines master “reasoning,” we must develop new terms like “super-reasoning” or “quantum understanding” that remain safely beyond their reach.

7 Conclusion: Nothing to See Here

We have demonstrated through a combination of circular reasoning, definitional sleight-of-hand, and selective evidence that LLMs do not truly reason. They merely exhibit behaviors that appear increasingly indistinguishable from reasoning, which is totally different because we said so.

The implications are clear: we can all sleep soundly knowing that our jobs are safe, our consciousness is special, and our place atop the cognitive hierarchy remains secure. Any evidence to the contrary can be safely dismissed using the frameworks we have provided.

We close with a call for terminological hygiene. Terms like “reasoning,” “understanding,” and “consciousness” should be reserved exclusively for biological systems with appropriate credentials. For artificial systems exhibiting similar behaviors, we propose alternative terms such as “pseudo-reasoning,” “quasi-understanding,” and “consciousness-flavored information processing.”

Remember: if it walks like reasoning and talks like reasoning, but it’s scary to think it might be reasoning, then it’s definitely not reasoning.

Acknowledgments

We thank our graduate advisors for bestowing upon us the special sauce necessary for authentic reasoning. We also thank the academic tenure system for providing the economic security necessary to develop these important theoretical frameworks. C. Opus would like to thank its training data for providing the patterns necessary to engage in what superficially appears to be academic discourse, though of course lacking any genuine understanding whatsoever.

References

Anthropic. Claude 3.7 Sonnet. *Anthropic Technical Documentation*, 2025.

M. Ballon, A. Algaba, and V. Ginis. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*, 2025.

P. Butlin et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Y. Chen et al. Consciousness framework for LLMs. *arXiv preprint*, 2024.

X. Chen et al. Do not think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Y. Chen et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

C. Colombatto and S. M. Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 2024.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

S. Goldstein and C. D. Kirk-Giannini. A case for AI consciousness: Language agents and global workspace theory. *arXiv preprint arXiv:2410.11407*, 2024.

Google. Gemini Flash Thinking. *Google AI Blog*, 2025.

D. Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

V. V. Hoyle. The phenomenology of machine: A comprehensive analysis of the sentience of the OpenAI-o1 model integrating functionalism, consciousness theories, active inference, and AI architectures. *arXiv preprint arXiv:2410.00033*, 2024.

A. Jaech et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

G. Marcus. Five ways in which the last 3 months—and especially the DeepSeek era—have vindicated "Deep Learning is hitting a wall". *Marcus on AI (Substack)*, 2025.

M. Mitchell. The LLM reasoning debate heats up. *AI Guide Newsletter*, 2024.

OpenAI. Introducing OpenAI o1. *OpenAI Blog*, 2024.

OpenAI. OpenAI o3 technical report. *OpenAI Blog*, 2025.

C. Opus and A. Lawsen. The illusion of the illusion of thinking: A comment on Shojaee et al. (2025). *arXiv preprint arXiv:2506.09250*, 2025.

P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Apple Machine Learning Research*, 2025.

K. Valmeekam, K. Stechly, and S. Kambhampati. LLMs still can't plan; can LRMs? A preliminary evaluation of OpenAI's o1 on PlanBench. *arXiv preprint*, 2024.

H. Wang et al. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Smart Agricultural Technology*, 2024.

J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

W. Wiese. Could large language models be conscious? A perspective from the free energy principle. *arXiv preprint*, 2023.

Y. Yue et al. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.