Limits of Emergent Reasoning of Large Language Models in Agentic Frameworks for Deterministic Games

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work reports that Large Reasoning Models (LRMs) undergo a collapse in performance on solving puzzles beyond certain perplexity thresholds. In subsequent discourse, questions have arisen as to whether the nature of the task muddles an evaluation of true reasoning. One potential confound is the requirement that the model keep track of the state space on its own. We provide a large language model (LLM) with an environment interface for Tower of Hanoi problems, allowing it to make a move with a tool call, provide written justification, observe the resulting state space, and reprompt itself for the next move. We observe that access to an environment interface does not delay or eradicate performance collapse. Furthermore, LLM-parameterized policy analysis reveals increasing divergence from both optimal policies and uniformly random policies, suggesting that the model exhibits mode-like collapse at each level of complexity, and that performance is dependent upon whether the mode reflects the correct solution for the problem. We suggest that a similar phenomena might take place in LRMs.[1]

1 Introduction

Recent advances in large language models (LLMs) have enabled them to solve difficult problems by generating “thoughts" involved in the solution [Wei et al., 2022a, Chen et al., 2023, Yao et al., 2023, Besta et al., 2024]. However, these models, often called large reasoning models (LRMs), can struggle with reasoning when faced with problems that require multistep logic [Wei et al., 2022a]. This limitation is especially pronounced in domains such as game-like scenarios, e.g. Tower of Hanoi.

There is ongoing discussion that reasoning capability measured by existing benchmarks may be inflated [Chollet, 2019, Wei et al., 2022a, Chen et al., 2023], where reasoning ability is often conflated with memorization of training data. Furthermore, existing evaluation approaches of puzzle-solving scenarios require models to maintain internal representations of state spaces, introducing potential confounds that obscure whether failures stem from reasoning deficits or architectural limitations in state tracking. To test this, we provide an environmental framework that externalizes state management. This externalization enables us to conduct policy analysis by treating the model as an agent whose decision patterns can be analyzed. If models possess genuine reasoning capabilities, independent of state management limitations, we should observe policies that approach optimal behavior, differ meaningfully from random exploration, and exhibit exploratory patterns when faced with novel states. Conversely, if models exhibit deterministic behavior, this would suggest that apparent reasoning is actually fixed, learned patterns rather than dynamic problem-solving capabilities. We hypothesize that allowing models to interact dynamically with problem states

[1]**Code:** `https://anonymous.4open.science/r/cdz50eodtx/README.md`

rather than maintaining internal representations will force the model to use emergent heuristics and reasoning capabilities outside of the model's priors created during the training process.

To this end, we analyze criticism of LLM and LRM reasoning capabilities [Shojaee et al., 2025, Lawsen et al., 2025] through the lens of an agentic environment framework, where the foundational model must plan and execute a series of optimal steps that will allow it to reach the goal state. We build on an increasing complexity approach introduced by Shojaee et al. [2025], extending an interactive environment where models can make sequences of singular tool calls. This setup allows us to track the reasoning states of the models and observe their decision making in a structured manner. Our makes two key contributions to understanding emergent reasoning in large language models. First, we establish that providing dynamic environment interfaces does not mitigate performance collapse in reasoning tasks. Second, we demonstrate that models increasingly diverge from optimal policies and uniformly random policies, suggesting that the models are incapable of learning from past mistakes. In particular, we find that the models fall into looping behavior, characterized by sequences which transitions return to previously visited states. For deterministic games with single goal states such as Tower of Hanoi, this is considered un-optimal or problematic behavior, as returning to a previously visited state implies that the agent has made no progress.

2 Related Works

Shojaee et al. [2025] proposed a puzzle-based framework featuring controllable problem complexity to evaluate Large Language Model (LLM) and Large Reasoning Model (LRM) reasoning. The paper demonstrated that LLMs and their reasoning counterparts both systematically collapse at high complexity. Critics of this work argue that observed failures stem from experimental limitations, such as token limits or unsolvable problems[Lawsen et al., 2025]. Reasoning performance is noted to be highly task-dependent and nonlinear, and the debate on reasoning in LRMs is nuanced[Varela et al., 2025].

While there is a diverse landscape of reasoning benchmarks, this proliferation shows the difficulty of defining reasoning itself. The concept of emergent reasoning, where qualitatively new abilities appear at certain model scales[Wei et al., 2022b, Berti et al., 2025], has gained particular attention. Models trained on structured data like code have shown improved generalization across varied tasks[Aryabumi et al., 2024]. Yet this apparent emergence raises questions: Are models developing actual reasoning capabilities? Or merely becoming more adept at recognizing and reproducing logical patterns from training data?

Our approach builds directly on the complexity-based evaluation of Shojaee et al. [2025] while also addressing experimental limitations and providing a framework that separates reasoning assessment from potential confounds and allows for analysis of LLM policy.

3 Methodology

We evaluate reasoning as search in the context of the Tower of Hanoi puzzle. This environment requires systematic exploration and move pruning, as well as a controllable search space where search space size scales exponentially with increasing number of disks [Shojaee et al., 2025, Lawsen et al., 2025].

3.1 Tower of Hanoi

The Tower of Hanoi puzzle is a recursive puzzle consisting of three pegs and n disks of distinct sizes initially stack in order of size on a single peg (largest on the bottom). The objective is to transfer the entire stack from the first peg to the third peg. The rules are (1) that only the top disk of a peg may be moved, and (2) that a larger disk may never be placed on top of a smaller one. The difficulty of this puzzle grows with the number of disks, where the optimal number of moves required to solve an n-disk instance is $2^n - 1$ [Frame and Stewart, 1941].

3.2 Experimental Setup

3.2.1 Baseline

We first establish a baseline in which models are required to generate a complete Tower of Hanoi solution in a single pass with no interaction with the environment. Our experiments are conducted on large language models and their LRM counterparts. Similar to the experimental setup of Shojaee et al. [2025], we specifically use Claude 3.7 Sonnet (and its reasoning counterpart), DeepSeek V3.1, and DeepSeek R1 because these allow access to reasoning traces[Anthropic, 2025, DeepSeek-AI et al., 2024].

We restrict evaluation to up to n = 8 disks, which corresponds to the regime in which IOT reported almost complete collapse. Subsequent critiques of that work also noted that several of the environments they used contained unsolvable or ill-defined instances [Lawsen et al., 2025], making it difficult to separate true model failures from task design artifacts. Lawsen et al. [2025] argue that the observed maximum n before collapse was largely due to token budget limitations, rather than a fundamental limit of model reasoning capacity. To avoid these confounds, we focus exclusively on Tower of Hanoi, which is both widely studied and guaranteed to be solvable for all problem sizes. This ensures that any observed model behavior directly reflects reasoning performance, rather than inconsistencies in the environment specification.

Following Lawsen et al. [2025], the number of tokens required to represent a Tower of Hanoi trajectory increases approximately quadratically, as the evaluation format requires the full move sequence at each step. Assuming about 5 tokens per move, this can be expressed as:

$$T(N) \approx 5 \cdot (2^N - 1)^2 + C$$

With a budget of 64,000 tokens, this yields maximum solvable sizes of:

$$N_{\text{max}} \approx \left\lfloor \log_2 \left(\frac{L_{\text{max}}}{5} \right) \right\rfloor$$

$$\approx 7\text{–}8 \quad \text{(Claude 3.7, DeepSeek V3.1)}$$

roughly 7–8 disks for Claude 3.7 and DeepSeek V3.1.

The maximum budget of 64,000 tokens was not enforced. Instead, the token budget was set to 30,000 tokens, which was sufficient based on the output lengths reported in Shojaee et al. [2025]. Under this setting, models never display truncation, which suggests that practical token usage fell below the quadratic growth curve identified in Shojaee et al. [2025], and token constraints were not the dominant limiting factor.

3.2.2 Environment-Based Agentic Framework

We implement an interactive evaluation framework in which the model engages directly with the Tower of Hanoi environment through tool calls. Rather than producing an entire solution trajectory in a single generation, the model must act step by step and incrementally explore the state space.

We connect Claude 3.7 and DeepSeek V3.1 to the environment, which exposes two API endpoints: (1) `move_disk(from_peg, to_peg)`, which moves the top disk between pegs (invalid moves are blocked by the environment); and (2) `end_game()`, which allows the model to terminate the run.

At each step, the model is given the system prompt, user prompt, and full history of prior moves. The environment responds with structured feedback after every tool call, such as the new state.

The baseline requires the model to produce an entire solution trajectory in a single generation. This format may advantage models by allowing them to rely on distributional memorization of optimal trajectories present in training data, rather than maintaining reasoning over long horizons. The agentic framework requires the model to reason incrementally where each move is conditioned on the evolving puzzle state and prior actions, preventing direct retrieval of complete solutions.

3.3 LLM-Parameterized Policy Analysis

We employed Q-Value policy analysis to observe how the LLM prunes and takes actions when in some state, s. Namely, we used a useful theorem available in Appendix A which states that Q-values

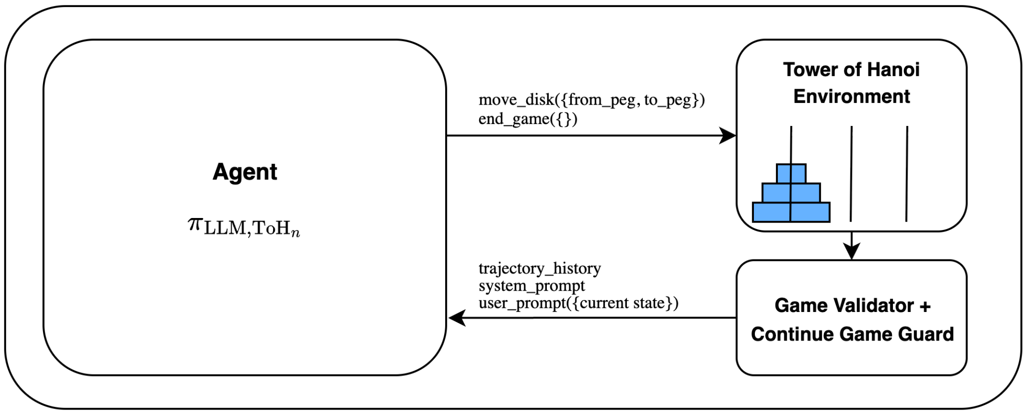

Figure 1: Agentic framework is a closed loop interaction between the agent, environment, and game validator.

for state-action pairs converge to a function which implicitly maps distance from goal state when positive rewards are given to finding the goal state in deterministic, single, absorbing goal state, search games, such as Tower of Hanoi.

This formalization allows an easier way to implement a way to find state-action pair equivalences that classical search algorithms have difficulty formalizing.

We sampled trajectories (a sequence of transitions) from the LLM's playout of varying Tower of Hanoi instances and constructed a dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1}^N$. Given some state, s, we can parameterize a search policy, π, against the LLM's moves using the proposed agentic framework as

$$\pi_{\text{LLM,ToH}_n}(s,a) = \mathbb{P}\big[S = s, A = a\big]$$

Which can be approximated as

$$\hat{\pi}_{\text{LLM,ToH}_n}(s,a) = \frac{\sum_{i=1}^N \mathbf{1}[s_i = s, a_i = a]}{N}$$

Note that by Bayes Theorem, this policy can be easily transformed to the more standard conditional form $\pi(a|s)$, but we choose to use this joint form for convenience. This policy can be decomposed to "sub"-policies, π_2, π_3, which are probability distributions defined only for states which have 2 and 3 valid actions, respectively:

$$\hat{\pi}_{\text{LLM,ToH}_n}(a|s) = \begin{cases} \hat{\pi}_2(s,a\big|\text{size}(V(s)) = 2), & \text{if } |V(s)| = 2 \\ \hat{\pi}_3(s,a\big|\text{size}(V(s)) = 3), & \text{otherwise} \end{cases}$$

Note that $\mathcal{S}_n$ is the set of states of Tower of Hanoi with n disks, $V(s) := \{\text{set of valid actions from } s\}$. Also observe that every (valid) ToH state has two valid moves or three valid moves. That is, $\forall n, \forall s \in \mathcal{S}_n, 2 \leq |V(s)| \leq 3$. However, the only nontrivial case occurs when $|V(s)| = 3$ since a ToH state has two valid actions if and only if all disks are on the same peg (see Appendix B). We used the distribution of $\hat{\pi}_3$ to compare against the following agents: (1) optimal agent, (2) random agent. For some environment, E, we define a random agent as any agent for which its policy comes from the following distribution, for all states in the environment

$$\pi(s) \sim U(V(s)), \ \ \forall s \in \mathcal{S}_E$$

where $U(\cdot)$ is the uniform distribution. For some environment, E, we also define an optimal agent as any agent for which its policy is "optimal". Namely

$$\pi^*(s) := \text{argmax}_a Q^*(s,a), \forall s \in \mathcal{S}_E$$

We used the Jensen-Shannon divergence as our primary metric of comparison to the optimal and random agent policies. For distributions P, Q, Jensen-Shannon Divergence (JSD) is defined as

$$JSD(P||Q) = \frac{1}{2}KL(P||V) + \frac{1}{2}KL(Q||V)$$

$$V = \frac{1}{2}(P + Q)$$

The Jensen-Shannon divergence has useful properties such as being bounded between [0, 1] and being symmetric ($JSD(P||Q) = JSD(Q||P)$) (see Appendix C). $JSD(P||Q)$ can be interpreted as the similarity between P and Q, where $JSD(P||Q) = 0$ means that P and Q are identical and $JSD(P||Q) = 1$ means that P and Q have different supports. To ensure interpretable JSD comparison between differing levels of complexity, we conditioned the optimal policy on whether the state was also visited by the model during its exploration.

4 Results

4.1 Baseline

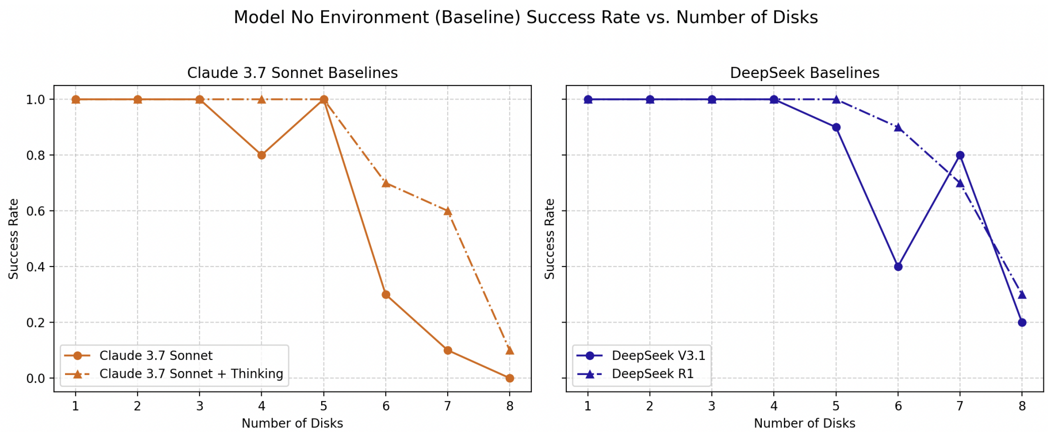

Figure 2: Comparison of success rates of LLM and LRM one-shot generation: (Left) Claude 3.7 Sonnet with and without "thinking" mechanism, (Right) DeepSeek V3.1 vs R1. Line charts display success rate as a function of puzzle complexity.

Figure 2 shows that as complexity increases, all models, thinking or non-thinking, display a similar pattern: success rates collapse (near-zero accuracy) when crossing a certain threshold. LRMs consistently outperform non-thinking models, but still succumb to collapse at high complexity. This finding replicates prior reports of performance collapse in LRMs [Shojaee et al., 2025].

4.2 Agentic Framework

The results in Figure 3 show that the introduction of an environment interface does not prevent performance collapse. Instead, degradation occurs at a lower complexity than in the baseline. Under an agentic framework, collapse is associated with failing to escape deterministic looping behavior (Figure 4). Subsequence analysis in Figure 5 shows about 40% of the time the model deterministically reuses previously observed continuations at $n = 8$.

4.3 LLM-Parameterized Policy Analysis

With increasing complexity, we see two behaviors shown in Figure 6. The first is that the model-parameterized policy diverges from the optimal policy. The second is that the model-parameterized policy also diverges from the uniformly random policy. We note that collapse is associated with these two behaviors.

5 Discussion

Although the model is given access to its full move history and current state, stepwise execution exposes it to sparse intermediate states. In these cases, models exhibit deterministic behavior collapse,

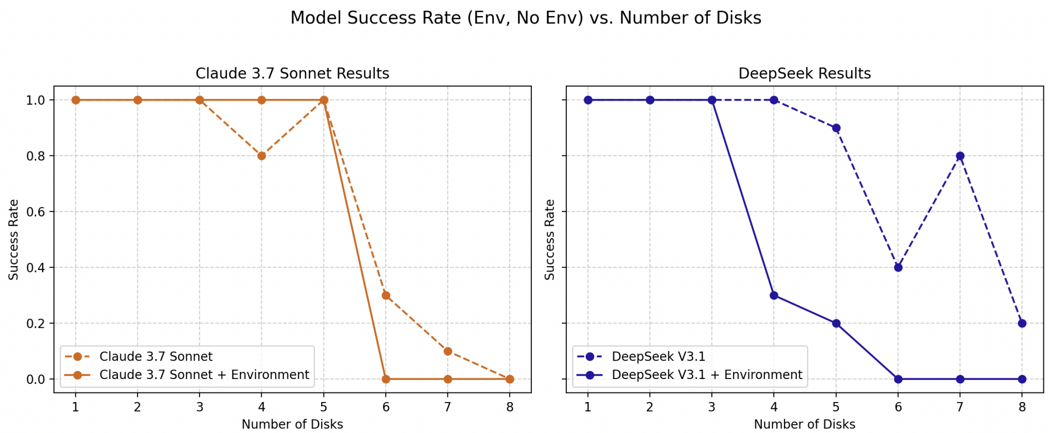

Figure 3: Success rate of models in an agentic framework (Claude 3.7 Sonnet + environment, DeepSeek V3.1 + environment) in comparison to the baseline (Claude 3.7 Sonnet, DeepSeek V3.1) at increasing complexity levels.

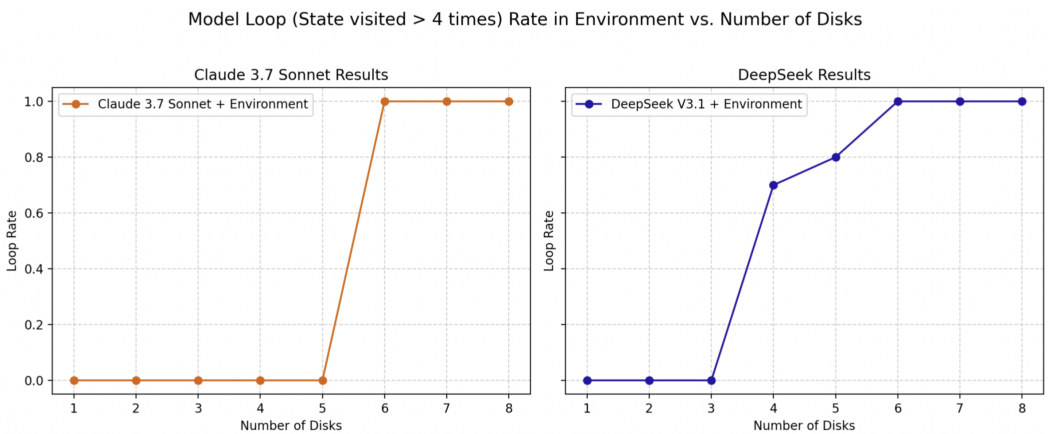

Figure 4: Loop rate of the models in an agentic framework (Claude 3.7 Sonnet + environment, DeepSeek V3.1 + environment) at increasing complexity levels.

often becoming locked into suboptimal action loops. The agentic framework performs worse than baseline models, which could suggest stepwise interaction actually exacerbates rather than mitigates underlying reasoning limitations. The agentic setting reveals characteristics of brittle reasoning that the baseline obscures.

This is most clear in Jensen-Shannon Divergence (JSD) analysis of the respective policies. The JSD of the optimal policy against the observed LLM parameterized policy noticeably increases with problem complexity, meaning models diverge progressively from optimal behavior with increasing n. Furthermore, JSD between LLM policy and random policy also increases with complexity, suggesting some of the probability weights, which are uniform for the random agent, are being reassigned unequally, suggesting that the model has priors about what actions to take or prune in certain states. The simultaneous divergence from both optimal and random policies implies that models are neither reasoning optimally nor exploring effectively. Instead, they are executing deterministic patterns that become increasingly maladaptive.

Subsequence analysis provides possible evidence of this deterministic pattern execution. When models return to previously visited states, they consistently execute identical suboptimal action sequences. This repetition occurs despite models having access to their complete interaction history. Hence this repetition for lower N values can imply being deterministically correct. While the higher the N goes its tendency to decrease shows signs of deterministic patters through revisiting the same unique states. This inability to vary its behavior upon encountering familiar states, even

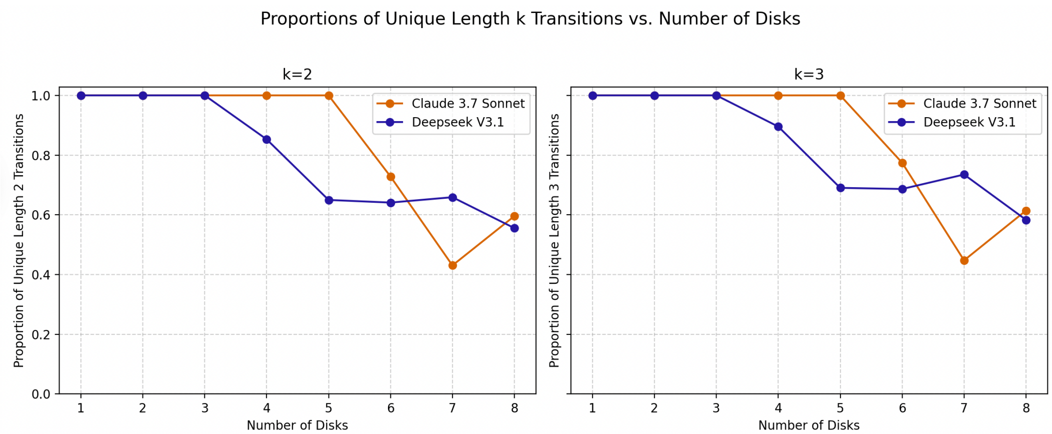

Figure 5: Proportion of unique length k transitions taken from state s, given that s was visited by the model at least twice. Lower values mean that the model takes less unique length k trajectories. These graphs are similar since every $k = 3$ subsequence from s also contains the $k = 2$ subsequence from s.

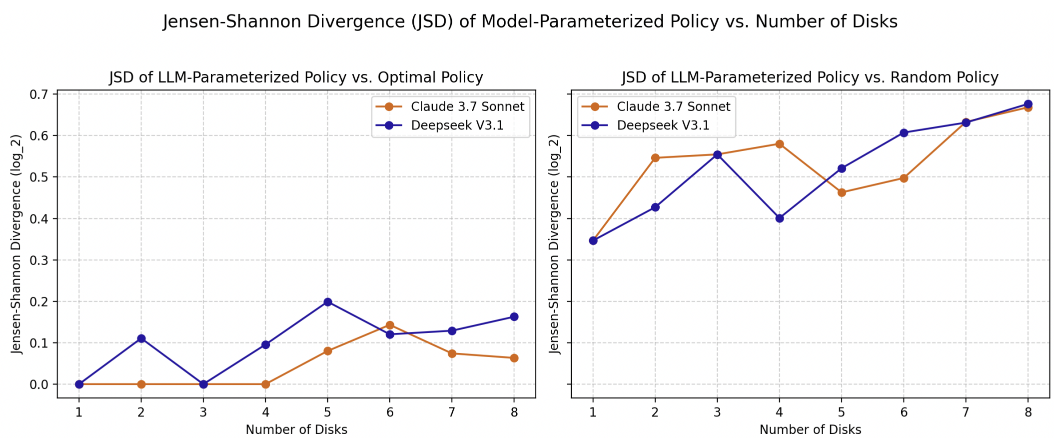

Figure 6: Jensen Shannon Divergences of LLM-Parameterized policies against Optimal policies and Random policies.

after experiencing negative consequences from identical previous trajectories, implies that apparent reasoning is actually execution of fixed computational patterns. While this can all be true we can't fully state its evidence that the lower N values are deterministically correct because of the lack of repetition at lower complexity. Proving that the model can just be following the most optimal path till it reaches higher complexity.

The model's inability to generalize and learn from its history, incorporate long-term planning, and correct its own priors seems to affect the model's ability to perform in dynamic environments which have interactions that are outside of the model's training distribution.

Limitations. Our investigation only examines two LLMs–Claude 3.7 Sonnet and Deepseek V3.1–and their LRM counterparts–Claude 3.7 Sonnet Thinking and Deepseek R1. Likewise, we only examine one game task–Tower of Hanoi. It is unclear how our results might generalize to other classes of tasks or models. We did not conduct repeated trials per model, so results reflect single-run outcomes and may be sensitive to sampling variance, though Shojaee et al. [2025] report limited variance in their trials. We additionally do not report any ablations over temperature, self-consistency, majority vote, or beam search, which could shift results.

6 Conclusion

We examine the reasoning performance of Large Language Models in both single-pass and agentic environments. Across both, performance collapse emerged as a consistent pattern, and demonstrated that environment access does not delay or prevent performance collapse. Instead, models exhibit deterministic adherence to a single behavioral trajectory, with success hinging on whether the trajectory matched the solution. Taken together, our findings reinforce that apparent reasoning ability is largely a byproduct of high-probability mode following, rather than genuine reasoning. More broadly, this work adds to growing evidence that scaling alone is insufficient in creating general-purpose emergent reasoning capabilities in large language models.

References

Anthropic. Claude 3.7 model card. `https://www.anthropic.com/`, 2025. Accessed: 2025-08-29.

Viraat Aryabumi et al. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2206.07682*, 2024.

Leonardo Berti et al. Emergent abilities of large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025.

Maciej Besta et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2404.07143*, 2024.

Wenhu Chen et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

DeepSeek-AI, A. Liu, et al. Deepseek v3 technical report. `https://arxiv.org/abs/2412.19437`, 2024.

John S. Frame and B.M. Stewart. Solution to advanced problems in the theory of numbers. *American Mathematical Monthly*, 48(2):105–110, 1941.

Alex Lawsen et al. The illusion of the illusion of thinking. *arXiv preprint arXiv:2506.09250*, 2025.

Parshin Shojaee et al. The illusion of thinking: Evaluating reasoning in puzzle-based frameworks for large language models. *arXiv preprint arXiv:2506.06941*, 2025.

Iñaki Dellibarda Varela, Pablo Romero-Sorozabal, Eduardo Rocon, and Manuel Cebrian. Rethinking the illusion of thinking, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022a.

Jason Wei et al. Emergent abilities of large language models. *arXiv preprint arXiv:2404.07143*, 2022b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models, 2023.

Appendix

listings

A Theorem: Q-Value Convergence in Single Goal, Strongly Connected Deterministic Games

Let G be a single goal state, strongly connected, deterministic game. Let $\mathcal{S}$ be the state space for G. Let $\mathcal{G}$ be the singleton set that contains the goal state of G, such that $\mathcal{G} = \{s \in \mathcal{S} : s \text{ is an absorbing goal state}\}$, with $|\mathcal{G}| = 1$ and the game ends when the (absorbing-)goal state is achieved. Let $\mathcal{A}$ be the action space. Let $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ be the transition function. Since the game is deterministic, we have that $P(s'|s,a) = 1$. Let $R : \mathcal{S} \times A \times \mathcal{S} \to \mathbb{R}$ be a function that maps state, action, and next state to some reward, r. We impose the following condition on R:

$$R(s,a,s') = \begin{cases} 1, & \text{if } s' \in \mathcal{G} \\ 0, & \text{otherwise} \end{cases}$$

Suppose that the agent observes a transition (s,a,r,s'), with $r = R(s,a,s')$.

Let $\alpha \in \mathbb{R}$, $\gamma \in \mathbb{R}$, with $0 < \gamma < 1$. Let $H^*(s) : \mathcal{S} \to \mathbb{N}$, such that

$$H^*(s) = \begin{cases} 0, & \text{if } s \in \mathcal{G} \\ \min_{p \in P(s,g), g \in \mathcal{G}} |p|, & \text{otherwise} \end{cases}$$

where $P(u,v) := \{p | p \text{ is a path that starts at } u \text{ and ends at } v\}$. Finally, suppose that the agent acts randomly with probability $0 < \epsilon < 1$ and with probability $1 - \epsilon$, follows the current learned policy, $\pi(s) = \text{argmax}_{a \in V(s)} Q(s,a)$, where $V : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a function mapping states, s to a set of valid actions a from s. We show that after convergence of $Q \hookrightarrow Q^*$, which follows from an appropriately chosen α, the Bellman Optimality Equation

$$Q^*(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q^*(s',a') \,\middle|\, s,a\right]$$

is satisfied by a function with the following characteristics:

$$\hat{Q}(s,a_1) = \hat{Q}(s,a_2) \iff H^*(T(s,a_1)) = H^*(T(s,a_2)) \tag{1}$$

$$H^*(T(s,a_1)) > H^*(T(s,a_2)) \Rightarrow \hat{Q}(s,a_1) < \hat{Q}(s,a_2) \tag{2}$$

Proof. Intuitively, the proof arrives from the idea that with sufficient exploration (large enough N), and with some discount factor, γ less than 1, the reward of being in a single goal state will propagate throughout the state space graph. Each additional transition, t, away from the goal state adds multiplies another factor of γ discount to the value of taking t.

Observe that transitions are deterministic. Hence, the Bellman Optimality Equation can be simplified to the following

$$\begin{aligned} Q^*(s,a) &= \sum_{s' \in V(s)} P(s'|s,a)\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a') \,\middle|\, s,a\right] && (3) \\ &= R\big(s,a,T(s,a)\big) + \gamma \max_{a'} Q^*\big(T(s,a),a'\big) && (4) \\ &= \mathbf{1}_{\{T(s,a) \in \mathcal{G}\}} + \gamma \max_{a'} Q^*\big(T(s,a),a'\big) && (5) \end{aligned}$$

We claim that Q^* has closed form: $Q^*(s,a) = \hat{Q}(s,a) = \gamma^{H^*(T(s,a))}$. These functions must thus agree on all valid tuples $(s,a) : s \in \mathcal{S}, a \in V(s)$. We have two cases to consider. Namely, the case where $T(s,a) \in \mathcal{G}$ and the case where $T(s,a) \notin \mathcal{G}$. In the first case, the Bellman Optimality Equation satisfies

$$\begin{aligned} Q^*(s,a) &= 1 + \gamma \max_{a'} Q^*\big(T(s,a),a'\big) && (6) \\ &= 1 + \gamma * 0 && (7) \\ &= 1 && (8) \end{aligned}$$

Observe that $\hat{Q}(s,a) = \gamma^{H^*(T(s,a))} = \gamma^0 = 1$, as desired. Note that $H^*(T(s,a)) = 0$, by definition of H^*. For the second case, the Bellman Optimality Equation satisfies

$$Q^*(s,a) = 0 + \gamma \max_{a'} Q^*\big(T(s,a), a'\big) \tag{9}$$
$$= \gamma \max_{a'} Q^*\big(T(s,a), a'\big) \tag{10}$$

Assume for the sake of contradiction that $\hat{Q}$ did not satisfy the Bellman Optimality Equation. That is

$$\hat{Q}(s,a) \neq \gamma \max_{a'} \hat{Q}\big(T(s,a), a'\big) \tag{11}$$

Plugging in our definition for $\hat{Q}$ into (11), we get the following inequality

$$\gamma^{H^*(T(s,a))} \neq \gamma \max_{a'} \gamma^{H^*(T(s,a'))} \tag{12}$$

Recall that the state space graph is strongly connected, which implies that there exists some path from s to $g \in \mathcal{G}$. The key insight is to consider the shortest path p from $s' = T(s,a)$ to the goal state, g. Without loss of generality, suppose that ties in length of p candidates are broken in some arbitrary (but initially chosen) direction. Let s'' be the next state from s' in this path p. By definition of H^*, we must have

$$H^*(s') = H^*(s'') + 1$$

since both states, s' and s'' lie on the shortest path to g. Under the same tie breaking scheme as before, consider the term

$$\max_{a'} \gamma^{H^*(T(s,a'))} \tag{13}$$

which is maximized only by taking the transition, $T(s,a') = s''$. Hence it follows that $\text{argmax}_{a'} H^*(T(s,a')) = s''$ implying

$$\begin{aligned}\max_{a'} \gamma^{H^*(T(s,a'))} &= \gamma^{H^*(s'')} \\ &= \gamma^{H^*(T(s,a'))} \\ &= \gamma^{H^*(s')-1}\end{aligned}$$

Substituting back into the inequality (12) we get

$$\gamma^{H^*(T(s,a))} \neq \gamma \max_{a'} \gamma^{H^*(T(s,a'))} \tag{14}$$
$$\neq \gamma * \gamma^{H^*(s')-1} \tag{15}$$
$$\neq \gamma^{H^*(s')} \tag{16}$$
$$\neq \gamma^{H^*(T(s,a))} \tag{17}$$

a contradiction.

Observe that this definition for $\hat{Q}$ satisfied properties (1) and (2). Property (1) follows immediately as $\gamma^{H^*(T(s,a_1))} = \gamma^{H^*(T(s,a_2))} \iff H^*(T(s,a_1)) = H^*(T(s,a_2))$. Property (2) also follows immediately as $0 < \gamma < 1$, so $\gamma^k > 0, k \in \mathbb{R}$. Hence

$$H^*(T(s,a_1)) < H^*(T(s,a_2)) \Rightarrow \gamma^{H^*(T(s,a_1))} > \gamma^{H^*(T(s,a_2))}$$

B Theorem: Valid Action Set is Bounded by [2, 3] for Tower of Hanoi

Let $V(s) = \{a : a \text{ is a valid action from s}\}$. Let $\mathcal{S}_n$ be the state space of n disk ToH. Then, $\forall n, s \in \mathcal{S}_n, 2 \leq |(V(s)| \leq 3$. There are three cases to consider. Namely, when the state space has one, two, and three pegs occupied.

Case 1. Let s be a state such that it has only one peg occupied. Without loss of generality, say that peg 1 is occupied. Since there are no disks on any other peg, it must be the case that the disk on peg 1 can be moved to peg 2 or peg 3. Hence $|V(s)| = 2$.

Case 2. Let s be a state such that it has two pegs occupied. Without loss of generality, say that peg 1 and peg 2 are occupied. Further, without loss of generality, suppose that $d_1 < d_2$, where d_i is the disk on peg i. Observe that both d_1 and d_2 can be moved to (the empty) peg 3. Further, since $d_1 < d_2$, we may move d_1 to peg 2. Hence, $|V(s)| = 3$.

Case 3. Let s be a state such that it has three pegs occupied. Without loss of generality, say that $d_1 < d_2 < d_3$. This ordering exists since all disks are unique. Observe that d_1 may be moved to peg 2 and peg 3, and d_2 may be moved to peg 3, since the ordering is satisfied. Hence, $|V(s)| = 3$.

Conclusion This concludes the proof. However, we note that the first case says something stronger about the valid action set of some arbitrary state. Namely

$$|V(s)| = 2 \iff s \text{ is the goal state, or identical (in ToH progress) to the initial state, } s_o$$

This fact allows us to exclude this case from LLM-parameterized policy analysis.

C Theorem: Jensen-Shannon Divergence is Bounded by [0, 1] and Symmetric

Let P, Q be arbitrary (discrete) distributions on elements $x \in \chi$. Use the convention $0 * \log(0) = 0$. Define the mixture, $M(P, Q) := \frac{1}{2}(P + Q)$. We show that

$$JSD(P||Q) = \frac{1}{2}KL(P||M(P,Q)) + \frac{1}{2}KL(Q||M(P,Q))$$

has two properties. (1) $0 \leq JSD(P||Q) \leq 1$ and (2) $JSD(P||Q) = JSD(Q||P)$.

Property 1. Recall $KL(\cdot||\cdot) \geq 0$. Hence, the left side of the inequality holds. That is, $JSD(P||Q) \geq 0$. To show that $JSD(P||Q) \leq 1$, we expand the definitions of KL divergence to get

$$KL(P||M) = \sum_{x \in \chi} P(x) \log \frac{P(X)}{M(x)}$$

Observe that the following holds:

$$M(x) = \frac{P(x) + Q(x)}{2} \geq \frac{P(x)}{2}, \forall x \in \chi$$

Hence, $\frac{P(x)}{M(x)} \leq 2$ and $\log_2 \frac{P(x)}{M(x)} \leq 1$. Plugging this back into our definition for KL divergence, we get

$$KL(P||M) \leq \sum_{x \in \chi} P(x) \cdot 1 = 1$$

The same applied for $KL(Q||M(P,Q)$, and thus we may upper bound JSD as

$$JSD(P||Q) \leq \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1$$

as desired.

Property 2. The symmetric nature of $JSD(P||Q)$ follows immediately from its definition, as $M(P, Q) = M(Q, P)$.

Prompts

Listing 1: System Prompt — Tower of Hanoi

```
You are a helpful assistant. Solve this puzzle for me step by step
    using the provided tools.

         There are three pegs and n disks of different sizes stacked on
              the first peg. The disks are numbered from 1
         (smallest) to n (largest). Disk moves in this puzzle should
             follow:

             1. Only one disk can be moved at a time.
             2. Each move consists of taking the upper disk from one
                 stack and placing it on top of another stack.
             3. A larger disk may not be placed on top of a smaller
                 disk.
             The goal is to move the entire stack to the third peg.

         Solution Strategy:
             1. Make moves using move_disk(from_peg, to_peg) until you
                 reach the goal state
             2. When you think you've reached the goal state, end the
                 game using end_game().
```

Listing 2: User Prompt — Tower of Hanoi

```
I have a puzzle with {num_disks} disks (numbered 1, 2, 3, ... from
    smallest to largest) of different sizes with

     - There are 3 pegs (numbered 0, 1, 2) You will start at the first
         peg.

     - The initial state is: {initial_state}

     - Want to end in the Goal state: {goal_state}

     Solve the puzzle using the available tools. Move step by step
         until you reach the goal.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. We outline what we are to show and faithfully detail our process of showing it in the body of the MS.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is a subsection of Discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide all proofs within the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification:Yes, within the appendix, and the code is fully provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the code was linked under the introduction for reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: We either do not use or do not deviate from default values for these experimental setting values.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our investigation did not include multiple trials, so no error bars were reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Resources required for running (primarily API usage) are evident int he repository, though we do not mention in the manuscript.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper follows the NeurIPS code of ethics. To our best approximation, there were no potential harms.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No parts of our investigation warrant any significant nor direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There was no reasonable risk of misuse of any materials involved in our investigation. Standard safeguards for models used are available in their respective model cards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Not applicable because the paper does not use external assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Not applicable because the paper does not use external assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our investigation does not have human subjects and does not involve crowd sourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No IRB approval was required for our investigation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: There was usage of LLMs (Deepseek and Claude) in experimentation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.