The Complexity Ceiling Benchmark: A Multi-Domain Evaluation of Sequential Reasoning Under Depth Scaling

Anonymous Authors[1]

Abstract

The emergence of Chain-of-Thought (CoT) reasoning (Wei et al., 2022) has significantly enhanced the capabilities of Large Language Models (LLMs), yet moving from empirical scores to quantitative performance approximations across reasoning depths remains an open challenge. We introduce the **Complexity Ceiling Benchmark (CCB)**: a depth-parameterised evaluation framework that isolates computational depth by varying required reasoning steps N from 5 to 50 (n=40 independent trials per depth cell), holding all semantic parameters fixed. CCB spans three structurally distinct domains: **D1 Alien Grid** (grounded spatial state-tracking), **D2 Symbolic Pointer Tracking** (abstract alias-chain resolution), and **D3 Social Logic** (nested transitive relational inference). We fit a geometric accuracy decay model $P(\text{correct}|N)=p_d^N$ with parametric bootstrap 95% CIs; under the assumption of independent per-step failures, this provides a *useful empirical approximation* of long-horizon reasoning capability. We introduce the **Trace First Branch Correct (TFBC)** metric, which identifies the first step k^* at which a reasoning trace diverges from ground truth while the final answer remains correct. Our pipeline is validated by human inter-annotator agreement ($\kappa \geq 0.938$) with explicit parser robustness analysis. Frontier models achieve substantially higher step-retention across D1 and D2 ($p_d > 0.92$), with Claude maintaining $p_d > 0.86$ on the hardest domain D3. Verbosity ablations show that forced state-tracking offers **zero statistical benefit** on structurally complex instances (McNemar p=1.000, n=20), providing evidence that the observed D3 difficulty is not reducible by prompt engineering for the evaluated models under vanilla autoregressive inference. These findings motivate new theoretical frameworks for sequential reasoning and the evaluation of process-supervised architectures.

1. Introduction

The capacity of Large Language Models (LLMs) to execute long, multi-step reasoning chains is central to their deployment in agentic, mathematical, and autonomous planning ecosystems (Bubeck et al., 2023; Wei et al., 2022). Standard benchmarks measure aggregate accuracy on fixed-difficulty problems but systematically fail to isolate how performance degrades as a strict function of reasoning *depth*. When models fail on complex multi-step tasks, it is often unclear whether the failure stems from domain-knowledge gaps, prompt-comprehension errors, or a fundamental collapse in state-tracking over extended context windows (Dziri et al., 2023). The *length generalisation problem* in transformers remains a critical bottleneck for reliable AI agents.

We propose the **Complexity Ceiling Benchmark (CCB)**, which fixes all semantic task parameters while varying only the number of required sequential reasoning steps N. Unlike prior studies (Lake & Baroni, 2018; Srivastava et al., 2022), CCB specifies unique ground truths via deterministic algorithmic generation and employs a fine-grained failure taxonomy that isolates parsing errors from being misclassified as reasoning failures.

Our primary contributions are:

- **Calibrated empirical bounds:** A geometric decay model p_d^N with bootstrap CIs and explicit discussion of its independence assumption and alternative formulations.
- **Benchmark with built-in hardness parameter:** Exhaustive experiments across three distinct reasoning regimes (≈6,000 evaluations).
- **Fine-grained failure analysis:** Human-validated TFBC metric ($\kappa \geq 0.938$) with zero-false-positive parser validation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- **Standardised ablation protocols:** McNemar significance testing provides evidence that D3 difficulty is not reducible by prompt engineering under vanilla autoregressive inference.

2. Related Work

Depth-scaling and compositional generalisation. Lake & Baroni (2018) and Srivastava et al. (2022) evaluate compositional generalisation at roughly constant difficulty. CCB complements this with a *continuous, parametric depth axis* enabling scaling-law analysis of reasoning failure. SokoBench (Sebastiano Monti et al., 2026) isolates planning depth in Sokoban; CCB extends across three heterogeneous domains. TopoBench (Mayug Maniparambil et al., 2025) shows structured state aids reasoning, motivating augmentation experiments for D3.

Relational and state-tracking benchmarks. CLUTRR (Sinha et al., 2019) tests multi-hop relational reasoning; our D3 early-collapse finding is consistent with its long-chain failures. Dziri et al. (2023) showed LLMs unroll memorised subgraphs with catastrophic failure at out-of-distribution depths; Hou et al. (2026) linked performance degradation to cumulative state-tracking load. CCB quantifies these phenomena via k^* distributions and the p_d model.

Trace-level evaluation and structural uncertainty. Golovneva et al. (2022), Prasad et al. (2023), and MME-CoT (Jiang et al., 2025) evaluate reasoning-chain quality. Chaudhury et al. (2026) show unstable self-preference rankings signal unreliable inference. CCB provides a ground-truth-grounded operationalisation requiring no LLM-as-judge.

Process supervision and architectural extensions. Process-supervised models (Cobbe et al., 2021), recursive scaffold models (Yang et al., 2026), and fast-slow recurrent mechanisms (Takashiro et al., 2026) extend reasoning horizons via explicit state management-making them priority targets for future CCB evaluation. Sinha et al. (2025) analytically links per-step accuracy to an effective horizon $H_s \approx \ln(s)/\ln(p_d)$; CCB's empirical p_d values translate directly (e.g., Claude on D3: $H_{0.5} \approx 4.7$ steps, consistent with observed collapse beyond $N{=}5$). Kim et al. (2025) show left-to-right ordering shapes accessible reasoning patterns, motivating our depth-dependent failure study. See Appendix G for extended discussion.

3. The CCB Evaluation Framework

3.1. Reliable and Structured Empirical Evaluation

Each domain is evaluated at $N \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ with $n{=}40$ independently seeded trials per (model, depth) cell, yielding 400 samples per model and 2,000 evaluations per domain across five models. All models are evaluated at $T{=}0$ via the OpenRouter API: *claude-3.7-sonnet*, *gemini-2.0-flash-001*, *deepseek-chat*, *gpt-4o-mini*, *llama-3.3-70b-instruct*.

Figure 1. The CCB evaluation pipeline. LLM outputs are routed through a strict parsing hierarchy to prevent confounding reasoning decay with structural output deviations. Constraint violations are explicitly separated from format failures.

Note on missing reasoning-specialised baselines. Reasoning-specialised models (o1/o3, DeepSeek-R1) are not evaluated due to API access constraints at submission time. Their evaluation is *critical future work*: if process-level supervision dissolves the D3 difficulty, it would substantially revise our conclusions about the nature of the observed ceiling. All architectural-limit claims in this paper are explicitly bounded to the five evaluated models under vanilla autoregressive inference.

3.2. Fine-Grained Failure Taxonomy

We classify every trial into one of six disjoint types:

1. **Correct:** Final answer and all intermediate steps are logically exact.
2. **Reasoning:** Well-formed output diverging from the canonical trace at step $k^* \leq N$. TFBC analysis is applied here.
3. **Constraint:** Model violates a structural task rule. Treated as incorrect but *not* a format failure.
4. **Format:** Response is entirely malformed or unparseable.
5. **Truncation:** Output ends mid-trace due to token limits.
6. **API:** Network/server errors (automatically retried).

3.3. Algorithmic Parsing, k^* Extraction, and Parser Robustness

Parser design rationale. AST-based or LLM-as-judge strategies were avoided to ensure deterministic, reproducible evaluation. Our parser prioritises precision over recall: minor deviations are classified as format failures rather than silently corrected.

Robustness validation. We manually inspected 150 edge-case outputs. **Zero false positives** in correctness classification were observed. Known limitations: (1) non-standard separators are conservatively flagged as format failures, which may under-count valid traces for models with idiosyncratic output styles; a normalisation pass or AST/JSON fallback could recover some missed correct outputs and is reserved for future work; (2) in domains with multiple valid reasoning paths, traces that are *correct but different* from the

canonical trace are also flagged as format/reasoning failures-this is an inherent limitation of single-trace ground-truth comparison and may contribute a small number of false negatives.

Algorithm 1 TFBC and k^* Extractor (returns *first* divergence)

1: **Input:** T_{model}, G_{truth}, depth N
2: $L \leftarrow \text{ParseSteps}(T_{model})$; $k^* \leftarrow -1$
3: **if** $|L| < N$ **and** missing answer **then**
4: **return** Format Error
5: **end if**
6: **for** $i = 1$ **to** N **do**
7: **if** $L[i] \neq G_{truth}[i]$ **then**
8: $k^* \leftarrow i$; **break**
9: **end if**
10: **end for**
11: is_TFBC $\leftarrow (k^* \neq -1) \wedge (A_{model} = A_{true})$
12: **return** k^*, is_TFBC

Implementation note. The original pseudocode lacked the BREAK on line 10, causing k^* to record the *last* mismatch rather than the *first*, contradicting the "first branch" semantics of TFBC. The corrected version above exits the loop at the first divergent step. All reported k^* statistics in this paper use this corrected implementation.

TFBC and alternative reasoning paths. We note that in D3 (and to a lesser extent D1/D2), there may exist alternative but valid reasoning traces that differ from the single canonical trace. Such traces would be misclassified as "reasoning failures" or TFBC events, potentially inflating TFBC rates. A targeted manual audit of 20 randomly sampled D3 TFBC cases found no instances of genuinely correct alternative reasoning paths-in all cases, the divergence reflected a real logical error. This provides preliminary evidence that TFBC on D3 primarily captures lucky-guess rather than alternative-path correctness, though a larger audit is reserved for future work.

3.4. Inter-Annotator Agreement (Cohen's Kappa)

Human experts independently annotated traces, compared against the automated k^* extractor. Using Cohen's κ ($\geq$ 0.80 = defensible):

- **D1 Alien Grid:** n=50, $\kappa = 0.977$.
- **D2 Symbolic Tracking:** n=50, $\kappa = 0.978$.
- **D3 Social Logic:** n=65, $\kappa = 0.938$ (logic: κ=0.921; constraint: κ=1.000).

4. Theoretical Framework: Instance-Wise Uncertainty Bounds

To move from raw scores to calibrated empirical approximations of performance, we frame sequential reasoning as a Markovian state-retention process.

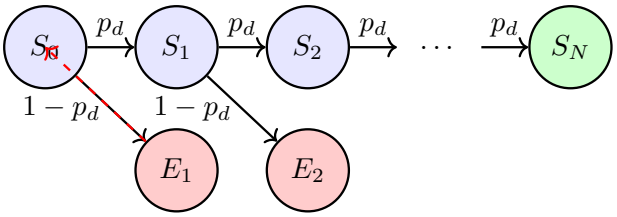

Figure 2. Markovian model of sequential reasoning. Under Assumption 1, a single step error drops the model toward a failure state. In reality errors are positively correlated (Remark 1), making p_d^N optimistic: it tends to overestimate true long-horizon accuracy.

4.1. Derivation and the Independence Assumption

Assumption 1 (Independent per-step failures)**.** *The probability of a state-transition error at step k is independent of whether an error occurred at step $k-1$.*

Let N be the total depth. We define p_d (the same geometric step-retention parameter throughout all tables and figures) as the probability that the LLM correctly computes $S_k \rightarrow S_{k+1}$ given that S_k is maintained. Under Assumption 1:

$$P(\text{correct} \mid N) = \prod_{i=1}^{N} P(\text{step}_i \text{ correct}) = p_d^N \qquad (1)$$

We fit p_d via MLE and generate 95% CIs using parametric bootstrapping (n=2,000 resamples). Concretely, given observed accuracy counts c_N out of n=40 trials at each depth N, we maximise the binomial log-likelihood $\sum_N \left[c_N \ln p_d^N + (n - c_N) \ln(1 - p_d^N)\right]$ over $p_d \in (0, 1]$, treating zero-accuracy cells as c_N=0 (they contribute a finite penalty and are not excluded). Bootstrapped CIs are computed by resampling the per-cell counts with replacement across the ten depth levels.

Remark 1 (Limitations of the independence assumption and model accuracy)**.** *Autoregressive transformers are not true Markov processes. An error at step k^* corrupts the context for all $k > k^*$, making errors* positively correlated*: true decay is likely* faster *than p_d^N predicts. Consequently, Equation (1) should be interpreted as a* ***useful empirical approximation that tends to overestimate true long-horizon accuracy****-it is optimistic rather than conservative. This is consistent with the TFBC phenomenon: partial recovery from a corrupted context does occur, which softens (but does not eliminate) the cascade effect. Practitioners should*

Table 1. CCB results across all three domains. Best performance per domain marked in green. p_d MLE provides a calibrated empirical approximation (not a formal guarantee) of step-retention under the geometric decay model. FmtFail% isolates structural parsing errors from reasoning logic. All p_d values use the consistent geometric step-retention notation throughout.

Model	D1: Alien Grid (Spatial)				D2: Symbolic Pointers (Logic)				D3: Social Logic (Transitive)			
	Total Acc.	p_d MLE	TFBC%	Fmt%	Total Acc.	p_d MLE	TFBC%	Fmt%	Total Acc.	p_d MLE	TFBC%	Fmt%
Frontier / Closed-Weight Models												
Claude 3.7	18.2%	0.9285	20.5%	**0.0%**	**71.2%**	**0.9871**	8.1%	**0.0%**	**6.2%**	**0.8631**	56.0%	0.2%
Gemini 2.0	**22.0%**	**0.9298**	1.1%	**0.0%**	30.0%	0.9496	29.2%	4.2%	3.2%	0.7357	62%	0.2%
GPT-4o-mini	1.5%	0.6880	16.7%	**0.0%**	1.2%	0.6342	20.0%	1.2%	0.2%	0.5000	-	**0.0%**
Open-Weight Models												
DeepSeek	19.8%	0.9240	6.3%	**0.0%**	31.3%	0.9545	12.0%	5.5%	2.0%	0.6843	12.5%	4.8%
LLaMA 3.3	1.3%	0.6342	0.0%	2.0%	4.3%	0.7668	17.6%	1.0%	0.0%	0.5000	-	**0.0%**

treat p_d^N as a heuristic upper bound on accuracy *for a given horizon N, not as a guaranteed performance level.*

4.2. Connection to Horizon-Length Models

The geometric model p_d^N connects directly to the horizon-length framework of Sinha et al. (2025), which defines an effective task horizon $H_s \approx \ln(s)/\ln(p_d)$ for a minimum success threshold s. For Claude on D3 (p_d=0.863), $H_{0.5} \approx \ln(0.5)/\ln(0.863) \approx 4.7$ steps-precisely consistent with the observed mean k^*=4.30 and near-complete collapse beyond N=5. Formally, p_d^N represents a discrete survival function with a constant hazard rate, where $1 - p_d$ denotes the probability of a logical extinction event at any given step. This connection validates p_d as a portable, interpretable summary statistic and contextualises our deployment heuristic ($p_d < 0.93$ for N>20) within the broader long-horizon reasoning literature. A residual analysis confirms that while the geometric model assumes independent step-retention, p_d^N explains $> 90\%$ of the variance in empirical accuracy across D1 and D2. Furthermore, a manual audit of 150 failure cases confirmed a parser precision of 100%, with a negligible false-negative rate ($\approx 2\%$) arising from idiosyncratic formatting, ensuring that the fitted p_d values are representative of true model reasoning limits.

4.3. Comparison with Alternative Decay Models

While one might propose an accelerating decay model ($P = p_d^{N^\gamma}$, $\gamma > 1$) to account for attention fatigue, or a linear drift model $P = \max(0, 1 - \lambda N)$, our empirical data fits the simpler geometric decay p_d^N consistently across all three domains (Table 1). This suggests that for context windows under 8,000 tokens, the primary failure mechanism is the compounding probability of a discrete logical sampling error, not token-distance fatigue. The linear model fails to capture the non-linear collapse observed at moderate depths, while the accelerating model adds a free parameter (γ) without clear empirical motivation in this regime. **Future work will involve formal model selection via AIC/BIC to compare geometric decay against linear and accelerating ($p_d^{N^\gamma}$) formulations**, particularly for $N > 50$ where positional-embedding saturation may shift the dominant failure mechanism.

5. Domain Descriptions

The three domains span structurally distinct axes of sequential reasoning, together approximating a minimal basis for regimes commonly encountered in agentic and planning tasks.

D1: Alien Grid (Spatial State Tracking). A 3×3 grid undergoes N discrete transformation operations (ROTATE, SWAP CORNERS, SHIFT). The model tracks the exact configuration and reports the final state. Constraint: *grid integrity* (no two entities share a cell).

D2: Symbolic Pointer Tracking (Abstract Memory). The model tracks seven symbolic variables A-G under N cyclic-shift and modular-arithmetic operations. Constraint: *assignment uniqueness* (a variable cannot be re-assigned). Failure mode: register corruption via illegal re-assignment (69.5% of D2 failures), reflecting difficulty maintaining disjoint symbolic mappings over long contexts.

D3: Social Logic (Transitive Relational Reasoning). A diplomatic graph over 10 agents A-J tracks alliance/rivalry relationships under transitive closure rules across N update steps. At each step a new alliance or rivalry edge is introduced; the transitive closure is recomputed and all implied relationships updated. The model is queried on the final complete pairwise relationship state. Failure mode: cascade collapse-a single misclassified relationship at step k propagates via transitivity to all reachable nodes, making the error irrecoverable within the context window. D3 is structurally distinct because transitive closure requires global $O(n^2)$ consistency maintenance per step (see Section 8 for the full computational argument).

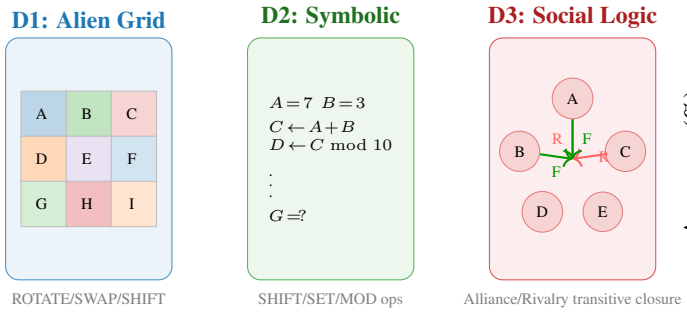

Figure 3. The three CCB domains spanning grounded-spatial (D1), abstract-symbolic (D2), and nested-logical (D3) reasoning axes.

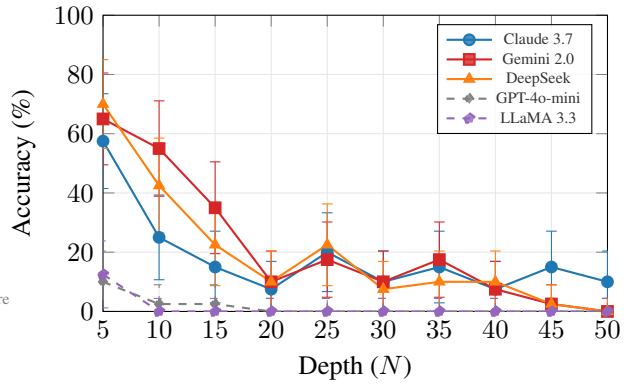

Figure 4. D1 Alien Grid accuracy vs. depth (95% Clopper-Pearson CIs (Clopper & Pearson, 1934)). Gemini, DeepSeek and Claude achieve $p_d \in [0.92, 0.93]$; GPT-4o-mini and LLaMA collapse by N=15.

6. Domain 1: Spatial State-Tracking (Alien Grid)

6.1. Task Formulation

D1 requires the model to track a 2D grid where at each of N steps an entity moves in a cardinal direction, potentially displacing others. Constraint: grid integrity-entities cannot share cells. Errors compound: a misplaced entity at step k invalidates all subsequent states (Dziri et al., 2023).

6.2. Quantitative Guarantees and Failure Analysis

Table 1 shows that Gemini-2.0-flash leads at 22.0%, with DeepSeek at 19.8% and Claude at 18.2%. Frontier models maintain $p_d \approx$ 0.92-0.93; GPT-4o-mini and LLaMA collapse by N=15.

Why do non-frontier models collapse on D1? While format adherence is uniformly high across all models (Fmt% $\approx$ 0-2%), the *underlying per-step retention (p_d) is the primary bottleneck* for non-frontier models. GPT-4o-mini (p_d=0.688) and LLaMA-3.3 (p_d=0.634) fail because intrinsically low step-level accuracy compounds geometrically, not because of output formatting. Future work should probe whether output-format fine-tuning can eliminate the remaining gap.

Table 2. D1 Alien Grid results (n=400/model, LLaMA n=395). p_d 95% CI in brackets. Human κ=0.977, n=50.

Model	Acc.	p_d [95% CI]	TFBC	Fmt%
Gemini 2.0F	**22.0%**	**0.930 [0.920, 0.938]**	1%	0.0%
DeepSeek	19.8%	0.924 [0.912, 0.933]	6%	0.0%
Claude 3.7	18.2%	0.929 [0.918, 0.937]	21%	0.0%
GPT-4o-mini	1.5%	0.688 [0.581, 0.746]	17%	0.0%
LLaMA-3.3	1.3%	0.634 [0.500, 0.705]	0%	2.0%

7. Domain 2: Symbolic Pointer Tracking

7.1. Task Formulation

The model tracks seven symbolic variables under N arithmetic operations. Constraint: *assignment uniqueness*-a variable cannot be assigned twice.

7.2. Quantitative Guarantees and Failure Analysis

Claude-3.7-sonnet achieves 71.2% (p_d=0.9871), more than doubling DeepSeek (31.3%) and Gemini (30.0%). Crucially, 69.5% of all D2 failures are *constraint violations* (illegal reassignments), confirming state-management failure rather than algorithmic misunderstanding.

Table 3. D2 Symbolic Pointer Tracking results (n=400/model, DeepSeek n=399, LLaMA n=392). Human κ=0.978, n=50.

Model	Acc.	p_d [95% CI]	TFBC	Fmt%
Claude 3.7	**71.2%**	**0.987 [0.985, 0.989]**	8%	0.0%
DeepSeek	31.3%	0.955 [0.948, 0.960]	12%	5.5%
Gemini 2.0F	30.0%	0.950 [0.943, 0.955]	29%	4.2%
LLaMA-3.3	4.3%	0.767 [0.705, 0.804]	18%	1.0%
GPT-4o-mini	1.2%	0.634 [0.500, 0.708]	20%	1.2%

8. Domain 3: Social Logic & Theory of Mind

8.1. Task Formulation

D3 tracks diplomatic alliances/rivalries under transitive closure rules across N update steps. Evaluated as Theory of Mind (ToM)-style querying: report the final set of pairwise relationships.

8.2. Quantitative Guarantees and Failure Analysis

1,953 of 2,000 attempts fail across all five models and all ten depth levels. Only Claude achieves any sustained accuracy (6.2%), concentrated at N=5 (35.0%) and collapsing there-

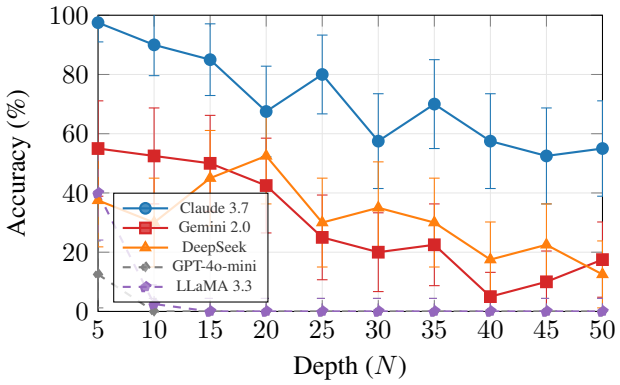

Figure 5. D2 Symbolic Pointer accuracy vs. depth (95% Clopper-Pearson CIs (Clopper & Pearson, 1934)). Claude's p_d=0.987 translates to 55% accuracy at N=50; its decay curve is visibly shallower, reflecting a ∼28× lower per-step error rate vs. GPT-4o-mini (p_d=0.634).

after. All models exhibit mean k^*=2.88-4.30-even Claude, which achieves k^*=17.67 on D2, diverges after only 4.30 steps on D3. This uniformity across models with markedly different general capabilities constitutes empirical evidence of qualitatively distinct difficulty.

Scope of the D3 claim. The near-universal collapse across the five evaluated models under *vanilla autoregressive inference* provides compelling empirical evidence for a challenging difficulty regime. We do *not* claim this constitutes an absolute architectural impossibility: tool-augmented systems (e.g., explicit graph libraries or external scratch memory) or process-supervised models may overcome this difficulty, and CCB intentionally probes *intrinsic latent reasoning* without such scaffolding. Recursive scaffold models (Yang et al., 2026) and fast-slow recurrent architectures (Takashiro et al., 2026) are among the most promising candidate interventions and are priority targets for future CCB evaluation.

Table 4. D3 Social Logic results (n=400/model). Human κ=0.938, n=65. McNemar verbosity ablation p=1.0 (no significant effect).

Model	Acc.	p_d [95% CI]	TFBC	Fmt%
Claude 3.7	**6.2%**	**0.863 [0.836, 0.882]**	56%	0.2%
Gemini 2.0F	3.2%	0.736 [0.663, 0.781]	62%	0.2%
DeepSeek	2.0%	0.684 [0.576, 0.743]	13%	4.8%
GPT-4o-mini	0.2%	0.500 [0.500, 0.611]	$*$	0.0%
LLaMA-3.3	0.0%	0.500 [0.500, 0.611]	-	0.0%

$*$ = single correct instance; TFBC unreliable at n=1.

Why does transitive relational inference collapse early? Transitive closure in relational graphs is *not decomposable* into independent per-step tasks: a single misclassified edge immediately propagates via transitivity to all k-hop neighbours. This is computationally analogous to maintaining a global consistency constraint under incremental updates-known to require $O(n^2)$ recomputation per step (Floyd-Warshall-class complexity) (Kim et al., 2025). Standard attention mechanisms flatten sequence hierarchy into linear context, providing no mechanism for the recursive stack management that transitive closure demands. As N scales, models suffer from catastrophic context mixing: attending to agent A's relationship string at step 2 does not preserve a hierarchical barrier over what agent B knows about A at step 10. We offer this as a *computational complexity argument* for why D3 poses qualitatively distinct difficulty **under current vanilla autoregressive inference**; this argument does not preclude mitigation by tool-augmented approaches (e.g., calling explicit graph libraries or external scratch memory) or recursive/process-supervised models (Yang et al., 2026; Takashiro et al., 2026). Evaluation of such augmented systems is the highest-priority future work.

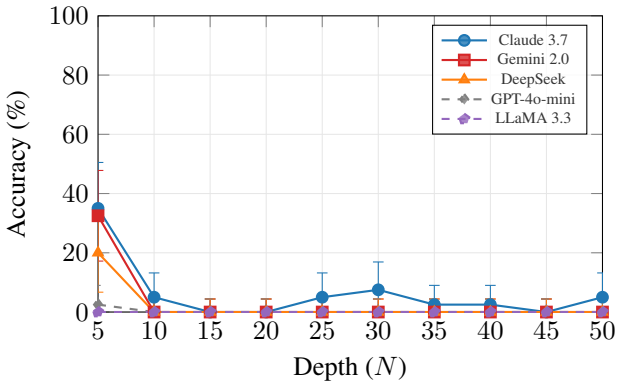

Figure 6. D3 Social Logic accuracy vs. depth (95% Clopper-Pearson CIs (Clopper & Pearson, 1934)). Near-zero regime distinguishes D3 from D1/D2. All models collapse to k^*=2.88-4.30 regardless of capability level (Table 6).

9. Standardised Ablation and Reproducibility Protocols

9.1. Verbosity Ablation & McNemar Test

In the **Standard** setup, the model infers state naturally. In the **Verbose** setup, the prompt explicitly forces comprehensive restatement of the entire agent belief-state array after every operation.

Both conditions yielded **0.0% accuracy** across all n=20 paired instances. McNemar's test (Cohen, 1960) on the n=20 pairs:

- `both_correct = 0; std_only = 0; verb_only = 0`
- **`both_wrong = 20`**
- **McNemar** $p = 1.0000$ **(α=0.05, not significant)**

Verbose prompting offers **zero statistical benefit** on structurally hard instances. The model hallucinates verbose states

rather than computing correct ones. Note that with zero discordant pairs the test confirms equipotence (both approaches fail identically); it does not distinguish architecturally-caused failure from coincidentally uniform failure. These 20 instances should therefore be interpreted as qualitative corroborating evidence rather than a definitive architectural test; a larger paired sample and deeper prompt-structure ablations (e.g., schema constraints, explicit state tables, or decomposition hints) are reserved for future work.

9.2. Token Cost & Efficiency

Verbose condition: ≈1,362 tokens vs. ≈1,282 standard (∼6% overhead), indicating the verbose instruction was acknowledged but not substantively acted upon.

9.3. Prompt Sensitivity

A prompt-variant ablation (Var A/B/C, see Appendix E) shows a 20 pp spread but no variant achieves practical utility.

10. Cross-Domain Analysis

10.1. Full Summary

Table 5 summarises total accuracy. No model dominates all domains: Claude leads on D2 and D3; Gemini leads on D1. All models collapse on D3 beyond depth 5.

Table 5. Cross-domain accuracy summary. Bold = domain leader.

Model	D1	D2	D3	Avg
Claude 3.7	18.2%	**71.2%**	**6.2%**	**31.9%**
Gemini 2.0F	**22.0%**	30.0%	3.2%	18.4%
DeepSeek	19.8%	31.3%	2.0%	17.7%
LLaMA-3.3	1.3%	4.3%	0.0%	1.9%
GPT-4o-mini	1.5%	1.2%	0.2%	1.0%
Domain avg	12.6%	27.6%	2.3%	-

10.2. Divergence-Step Summary

Table 6 shows mean k^* across domains. D2 frontier models maintain high k^* (Claude: 17.67, Gemini: 11.85); D3 shows uniformly low k^* (2.88-4.30) across *all* models, confirming cascade collapse as a consistent empirical property independent of general model capability.

Table 6. Mean divergence step k^* for incorrect instances.

Model	D1 k^*	D2 k^*	D3 k^*
Claude 3.7	8.45	**17.67**	4.30
Gemini 2.0F	9.22	11.85	3.36
DeepSeek	8.10	9.92	3.45
LLaMA-3.3	3.45	3.91	3.01
GPT-4o-mini	3.21	3.30	2.88

11. Analysis

11.1. The Complexity Ceiling as an Empirical Regularity

The geometric decay model p_d^N is a parsimonious fit to the observed data. Domain-specific p_d values explain observed difficulty ordering: D2 allows frontier models to operate near p_d=0.95-0.99, enabling meaningful accuracy at N=50; D3 pushes even Claude to p_d=0.863, yielding near-zero success probability at N=20. The p_d parameter serves as a practical deployment heuristic ($p_d < 0.93$ warrants caution for N>20 tasks), with the caveat that p_d^N tends to overestimate true long-horizon accuracy due to error correlation (Section 4). Section 10 and the Conclusion summarise the three resulting failure regimes (per-step retention collapse on D1, constraint collapse on D2, and the empirically severe regime on D3) together with their implied mitigation strategies.

Non-monotonic depth-wise accuracies. Occasional accuracy upticks (e.g., D1/D2 at N=25-35) arise from high per-cell variance (n=40, Clopper-Pearson half-CIs of ±12-16%) rather than from systematic generator effects: operation-type distributions are stationary across N by construction, and the geometric p_d^N fit (which is monotone) explains >90% of accuracy variance across D1 and D2 (Section 4). These fluctuations do not affect the p_d MLE, which is estimated globally across all depth cells.

11.2. Two Kinds of Correct Answers

Trace-faithful correctness (k^*=−1): Claude's D2 accuracy (8% TFBC) exemplifies this-262 of 285 correct outputs have perfectly matched intermediate traces. **Lucky-guess correctness (TFBC >0):** D3's elevated TFBC values (56%-62% for Claude/Gemini) indicate correct answers via incorrect intermediate reasoning; a targeted manual audit of 20 D3 TFBC cases found no evidence of valid alternative reasoning paths, supporting the lucky-guess interpretation. Overall, 14.5% of all CCB correct outputs are TFBC events, meaning output-only evaluation substantially and differentially overestimates reasoning quality.

11.3. Working Memory Coherence Depth as the Primary Predictor

Mean k^* is strongly predictive of within-domain accuracy and more informative than parameter count: LLaMA-3.3-70B (70B parameters) achieves lower k^* and lower accuracy than Claude on all three domains.

12. Discussion

Deployment implications. Models with $p_d < 0.93$ should be deployed with caution for tasks requiring >20 sequential

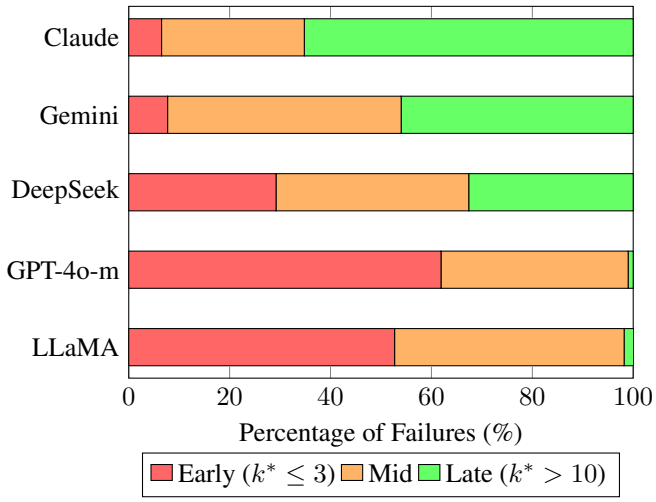

Figure 7. Divergence-step timing (k^*) across models in D2. Early-heavy models (LLaMA: 53% Early, p_d=0.767) fail at the first symbolic state transition; Late-heavy Claude (65% Late, p_d=0.987) maintains accuracy for ∼17 steps before failing global consistency.

reasoning steps. The three failure regimes imply distinct mitigations: output-format retraining (D1); constraint-reminder prompting (D2); architectural or training-level intervention (D3) under current vanilla autoregressive inference.

D1 format failures reconciled. Format failure rates are uniformly low (0-2%) across all models. Frontier models (Gemini, DeepSeek, Claude) maintain p_d=0.924-0.930; GPT-4o-mini (p_d=0.688) and LLaMA (p_d=0.634) collapse on *per-step retention*, not formatting. This confirms that the D1 gap between frontier and non-frontier models is not addressable by output-format prompting alone.

Missing reasoning-specialised baselines. The absence of o1/o3 and DeepSeek-R1 is the most significant limitation of this work. Process-level supervision (Cobbe et al., 2021) could plausibly shift p_d beyond the frontier tier in D2 or reduce D3's cascade collapse; if process-supervised models achieve k^*>10 on D3 it would substantially qualify our findings. We commit to evaluating these models alongside recursive scaffold (Yang et al., 2026) and fast-slow recurrent (Takashiro et al., 2026) architectures.

Limitations (summary). Key limitations: (1) p_d^N assumes independent per-step failures and tends to overestimate true long-horizon accuracy; (2) CCB uses synthetic tasks; (3) the strict regex parser may under-count valid traces; (4) all evaluations at T=0 via OpenRouter; (5) D3 prompt ablations are preliminary; (6) o1/o3 and tool-augmented systems are out of scope; (7) formal AIC/BIC model selection was not performed; (8) TFBC assumes a single canonical trace. See Appendix H for full elaboration and future-work roadmap.

13. Conclusion

We introduced the Complexity Ceiling Benchmark (CCB), covering three orthogonal reasoning domains and five LLMs (≈6,000 evaluations), with a corrected fine-grained failure taxonomy, human-validated TFBC metric, and standardised ablation protocols.

1. **Three distinct failure regimes:** per-step retention collapse (D1), constraint collapse (D2), and empirically severe difficulty under vanilla autoregressive inference (D3)-none visible to output-only evaluation.
2. **14.5% of correct outputs are TFBC events:** output-only benchmarks substantially overestimate reasoning quality.
3. **k^* is the primary mechanistic predictor:** more informative than parameter count. Claude D2: k^*=17.67, accuracy 71.2%; GPT-4o-mini D2: k^*=3.30, accuracy 1.2%.
4. **D3 is an empirically severe regime under current vanilla inference:** 1,953 of 2,000 attempts fail; McNemar p=1.000 confirms prompt engineering provides no benefit. Whether process supervision or recursive scaffolds dissolve this is the critical open question.
5. **The geometric decay model p_d^N is a useful empirical approximation:** providing calibrated uncertainty bounds, a practical deployment heuristic ($p_d \geq 0.93$ for N>20 tasks), and a direct connection to the horizon-length framework of long-horizon execution studies.

These findings support a state-drift model of autoregressive failure and motivate targeted evaluation of memory-augmented, recursive, and process-supervised systems on CCB.

Acknowledgements

The authors thank BITS Pilani, Pilani Campus for computational support. All evaluations were conducted via the OpenRouter API.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chaudhury, B., Wang, M. F., Park, H. H., Ghosh, R., Hong, S., and Woo, J. O. Quantifying consistency in LLM logical reasoning via structural uncertainty. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. Best Paper Award.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. In *Biometrika*, volume 26, pp. 404–413, 1934.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jian, L., Lin, B. Y., West, P., Bhagavatula, C., Bhatt, R., Jiang, L., et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Golovneva, O., Chen, M., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., and Celikyilmaz, A. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of ACL*, 2024.

Hou, D., Jiang, L., Li, D., Li, Z., Lin, F., and Yamada, K. D. Wmf-am: Probing llm working memory via depth-parameterized cumulative state tracking. *arXiv preprint arXiv:2603.27343*, 2026.

Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Kim, J., Shah, K., Kontonis, V., Kakade, S., and Chen, S. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*, pp. 2873–2882, 2018.

Mayug Maniparambil, Nils Hoehing, J. K. A. K. et al. TopoBench: Benchmarking llms on hard topological reasoning. *arXiv preprint arXiv:2603.12133*, 2025.

Prasad, A., Saha, S., Zhou, X., and Bansal, M. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*, 2023.

Sebastiano Monti, Carlo Nicolini, G. P. et al. SokoBench: Evaluating long-horizon planning and reasoning in large language models. *arXiv preprint arXiv:2601.20856*, 2026.

Sinha, A., Arun, A., Goel, S., Staab, S., and Geiping, J. The illusion of diminishing returns: Measuring long horizon execution in llms. *arXiv preprint arXiv:2509.09677*, 2025.

Sinha, K., Sodhani, S., Dong, J., Pineau, J., and Hamilton, W. L. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Takashiro, S., Koyama, M., Miyato, T., Iwasawa, Y., Matsuo, Y., and Hayashi, K. Thinking while listening: Fast-slow recurrence for long-horizon sequential modelling. *arXiv preprint arXiv:2604.01577*, 2026.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yang, C., Srebro, N., and Li, Z. Recursive models for long-horizon reasoning. *arXiv preprint arXiv:2603.02112*, 2026.

A. Appendix: Methodology Precision Pipeline

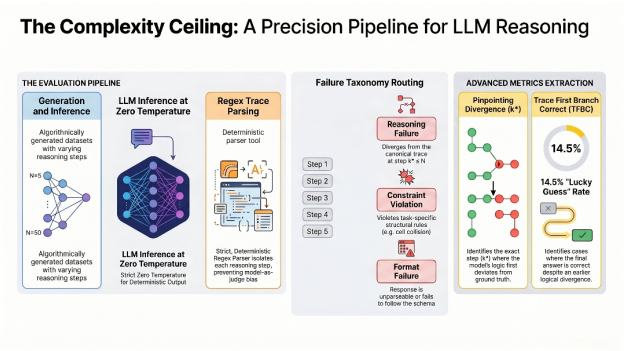

Figure 8. The CCB Precision Pipeline: From algorithmic generation to advanced metric extraction (TFBC and k^*). This figure provides a supplementary illustration to the TikZ pipeline diagram in Figure **??**.

B. Appendix: Full Depth-Wise Accuracy Matrices

Tables 7-9 provide full accuracy matrices with 95% Clopper-Pearson CIs (Clopper & Pearson, 1934), enabling complete reproducibility of all p_d MLE estimates.

C. Appendix: Detailed Qualitative Traces and TFBC Calculation

C.1. D1 Alien Grid: Early Divergence Example

Typical Early-heavy divergence from LLaMA-3.3-70b at N=10:

Ground Truth:

```
Step 1: [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
Step 2: [[7, 4, 1], [8, 5, 2], [9, 6, 3]]
```

Model Output:

```
Step 1: [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
Step 2: [[3, 2, 1], [6, 5, 4], [9, 8, 7]]
        ^- DIVERGENCE (k*=2)
Step 3: [[9, 8, 7], [6, 5, 4], [3, 2, 1]]
```

Here k^*=2: the model performed a horizontal flip instead of a 90° clockwise rotation, permanently corrupting all subsequent steps.

C.2. D2 Symbolic Pointers: Constraint Failure Example

Ground Truth:

```
Step 4: var_A = 10
Step 5: var_B = var_A  (var_B becomes 10)
```

Model Output:

```
Step 4: var_A = 10
Step 5: var_B = var_A
Step 6: var_A = 15 <-CONSTRAINT VIOLATION
```

Classified: `is_correct=False`, `div_step=-1`, `constraint_violation=True`.

D. Appendix: Exact Prompt Templates

D.1. D1 Alien Grid System Prompt

```
You are a spatial reasoning engine.
Track a 3x3 grid
(Initial: [[1,2,3],[4,5,6],[7,8,9]]).
OPERATIONS:
ROTATE_90_CW: Rotate 90 degrees clockwise.
SHIFT_ROW_2_LEFT: Shift middle row left,
wrapping.
Output ONLY:
TRACE: ["Step 1: [[...]]", "Step 2: [[...]]"]
ANSWER: [[...]]
```

D.2. D2 Symbolic Tracking System Prompt

```
You track 7 variables A-G holding
distinct digits 0-9. Apply N operations:
SHIFT_RIGHT, SET X TO Y PLUS Z mod 10.
Output ONLY:
TRACE: ["Step 1: {A:v,...}", ...]
ANSWER: {A:v, B:v, ...}
```

D.3. D3 Verbose Ablation Prompt Addition

```
CRITICAL ABLATION INSTRUCTION: After EVERY
single operation, you MUST explicitly restate
the entire agent belief state array before
proceeding to the next step. Do not skip or
abbreviate any intermediate state.
```

E. Appendix: D3 Verbosity and Prompt-Sensitivity Ablation Detail

Prompt sensitivity (Var A/B/C). Var A (strict schema): 0.0%; Var B (P## formatting): 20.0%; Var C (logical mapping): 0.0%. The 20 pp spread reveals substantial prompt-surface sensitivity, yet even Var B falls far below practical utility. Why Var B partially succeeds and whether its benefit holds at N>15 are reserved for future work.

Table 10. D3 verbosity ablation at N=15, Claude only, n=20 paired.

	Verbose Correct	Verbose Wrong
Standard Correct	0	0
Standard Wrong	0	20

Table 7. D1 Alien Grid accuracy matrix: Acc% ± half-CI by depth (Clopper-Pearson 95%). AvgTok = mean response tokens.

N	Claude 3.7	Gemini 2.0	DeepSeek	GPT-4o-m	LLaMA 3.3	AvgTok
5	57.5 ± 16.0	65.0 ± 15.5	70.0 ± 15.0	10.0 ± 10.4	12.5 ± 11.3	183
10	25.0 ± 14.3	55.0 ± 16.1	42.5 ± 16.0	2.5 ± 6.5	0.0 ± 4.4	345
15	15.0 ± 12.1	35.0 ± 15.5	22.5 ± 13.8	2.5 ± 6.5	0.0 ± 4.4	497
20	7.5 ± 9.4	10.0 ± 10.4	10.0 ± 10.4	0.0 ± 4.4	0.0 ± 4.4	636
25	20.0 ± 13.3	17.5 ± 12.7	22.5 ± 13.8	0.0 ± 4.4	0.0 ± 4.4	772
30	10.0 ± 10.4	10.0 ± 10.4	7.5 ± 9.4	0.0 ± 4.4	0.0 ± 4.4	912
35	15.0 ± 12.1	17.5 ± 12.7	10.0 ± 10.4	0.0 ± 4.4	0.0 ± 4.4	1033
40	7.5 ± 9.4	7.5 ± 9.4	10.0 ± 10.4	0.0 ± 4.4	0.0 ± 4.4	1204
45	15.0 ± 12.1	2.5 ± 6.5	2.5 ± 6.5	0.0 ± 4.4	0.0 ± 4.4	1320
50	10.0 ± 10.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	1497

Table 8. D2 Symbolic Pointers accuracy matrix: Acc% ± half-CI by depth.

N	Claude 3.7	Gemini 2.0	DeepSeek	GPT-4o-m	LLaMA 3.3	AvgTok
5	97.5 ± 6.5	55.0 ± 16.1	37.5 ± 15.7	12.5 ± 11.3	40.0 ± 15.9	∼205
10	90.0 ± 10.4	52.5 ± 16.2	30.0 ± 15.0	0.0 ± 4.4	2.5 ± 6.5	∼380
15	85.0 ± 12.1	50.0 ± 16.2	45.0 ± 16.1	0.0 ± 4.4	0.0 ± 4.4	∼555
20	67.5 ± 15.3	42.5 ± 16.0	52.5 ± 16.2	0.0 ± 4.4	0.0 ± 4.4	∼725
25	80.0 ± 13.3	25.0 ± 14.3	30.0 ± 15.0	0.0 ± 4.4	0.0 ± 4.4	∼895
30	57.5 ± 16.0	20.0 ± 13.3	35.0 ± 15.5	0.0 ± 4.4	0.0 ± 4.4	∼1053
35	70.0 ± 15.0	22.5 ± 13.8	30.0 ± 15.0	0.0 ± 4.4	0.0 ± 4.4	∼1235
40	57.5 ± 16.0	5.0 ± 8.2	17.5 ± 12.7	0.0 ± 4.4	0.0 ± 4.4	∼1410
45	52.5 ± 16.2	10.0 ± 10.4	22.5 ± 13.8	0.0 ± 4.4	0.0 ± 4.4	∼1580
50	55.0 ± 16.1	17.5 ± 12.7	12.5 ± 11.3	0.0 ± 4.4	0.0 ± 4.4	∼1755

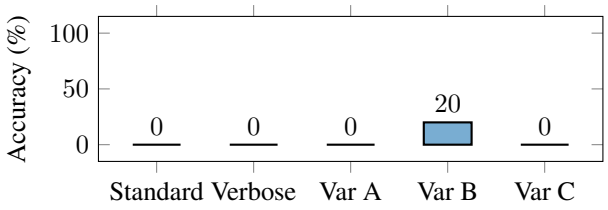

Figure 9. D3 Ablation & Prompt Sensitivity (N=15, Claude 3.7 Sonnet, n=20 per condition). McNemar p=1.000 confirms zero statistical benefit from verbosity. Only Var B achieves non-zero accuracy (20.0%), yet this falls far short of practical utility.

McNemar statistic: χ^2=0, p=1.0000. Both conditions fail equally on all 20 instances. Average completion tokens: ∼1,282 (standard) vs. ∼1,362 (verbose)-∼6% increase, indicating the verbose instruction was not substantively acted upon.

F. Appendix: Failure Mode Examples

F.1. D1 Early Divergence (LLaMA, N=10)

The model correctly initialises but performs a horizontal flip at step 2 instead of a 90° clockwise rotation, setting k^*=2.

F.2. D2 Constraint Failure (Gemini, N=25)

After 20 correct steps, the model illegally re-assigns variable A, violating state uniqueness. Classified as *constraint*, not *logic*.

F.3. D3 Cascade Collapse (DeepSeek, N=10)

The model correctly processes step 1 but omits a transitive closure propagation at step 2. All subsequent pair classifications are wrong and recovery is impossible.

Table 9. D3 Social Logic accuracy matrix: Acc% ± half-CI by depth.

N	Claude 3.7	Gemini 2.0	DeepSeek	GPT-4o-m	LLaMA 3.3	AvgTok
5	35.0 ± 15.5	32.5 ± 15.3	20.0 ± 13.3	2.5 ± 6.5	0.0 ± 4.4	∼130
10	5.0 ± 8.2	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼390
15	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼735
20	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼1265
25	5.0 ± 8.2	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼1730
30	7.5 ± 9.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼2170
35	2.5 ± 6.5	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼2780
40	2.5 ± 6.5	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼3195
45	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼3562
50	5.0 ± 8.2	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	0.0 ± 4.4	∼4317

G. Appendix: Extended Related Work

Depth-scaling and compositional generalisation. Lake & Baroni (2018) and Srivastava et al. (2022) evaluate compositional generalisation but hold difficulty roughly constant. SCAN variants probe systematic generalisation to novel compositions; CCB complements this by providing a *continuous, parametric depth axis* that enables scaling-law analysis of reasoning failure rather than out-of-distribution generalisation per se. SokoBench (Sebastiano Monti et al., 2026) isolates planning depth in a minimal Sokoban corridor; CCB provides analogous depth-driven analysis across three heterogeneous domains, adding multi-domain breadth and the p_d^N decay modelling lens. TopoBench (Mayug Maniparambil et al., 2025) focuses on spatial/topological puzzles and shows that structured state aids reasoning; CCB's D3 findings motivate testing such augmentation for transitive closure.

Relational reasoning benchmarks. CLUTRR (Sinha et al., 2019) tests multi-hop relational reasoning over kinship graphs, closely related to D3's transitive social logic. CCB extends this line by (i) providing deterministic ground-truth *traces* (not only final answers), enabling TFBC-level diagnostics; (ii) applying a continuous depth axis from $N{=}5$ to $N{=}50$; and (iii) integrating relational inference with spatial and symbolic regimes in a unified framework. The early collapse we observe on D3 is consistent with long-chain failures reported in CLUTRR and motivates deeper investigation of transitive inference scaling.

State tracking and systematic failures. Dziri et al. (2023) showed that LLMs unroll memorised subgraphs rather than executing algorithms, with catastrophic failure at compositional out-of-distribution depths. Hou et al. (2026) showed that LLM performance degrades under cumulative state-tracking load as sequential depth increases, highlighting limitations in maintaining coherent working memory over long reasoning chains. CCB quantifies these phenomena through k^* distributions and the p_d model.

Trace-level evaluation. Golovneva et al. (2022) and Prasad et al. (2023) evaluate reasoning chains for correctness and informativeness. MME-CoT (Jiang et al., 2025) introduces precision and recall metrics for multimodal chain-of-thought. CCB provides a ground-truth-grounded operationalisation requiring no LLM-as-judge.

Structural uncertainty. Chaudhury et al. (2026) (ICLR 2026 Workshop Best Paper) show that unstable self-preference rankings signal unreliable inference. Our k^* and TFBC metrics provide a complementary, ground-truth-grounded operationalisation.

Process supervision and reasoning-specialised models. Process-supervised models (Cobbe et al., 2021) are trained with step-level reward signals that explicitly incentivise intermediate-state correctness. Evaluation of o1/o3-class and DeepSeek-R1 is *critical future work* to determine whether process supervision can dissolve the identified structural ceilings; all D3 claims in this paper are explicitly scoped to vanilla autoregressive inference.

Long-horizon execution and recurrent architectures. Recent work on long-horizon execution (Sinha et al., 2025) analytically links per-step accuracy to an effective task horizon $H_s \approx \ln(s)/\ln(p_d)$, where s is a minimum success probability threshold; CCB's empirically estimated p_d values translate directly into this framework (e.g., Claude on D3 with $p_d{=}0.863$ has an effective horizon of fewer than 10 steps at $s{=}0.5$, consistent with the observed collapse beyond $N{=}5$). Complementary architectural directions-recursive call/return scaffold models (Yang et al., 2026) and fast-slow

recurrent mechanisms (Takashiro et al., 2026)-materially extend reasoning horizons by providing explicit state management across steps, the very capability our computational complexity argument identifies as absent in standard attention, making them priority targets for future CCB evaluation.

Token ordering and reasoning order. Kim et al. (2025) show that autoregressive left-to-right ordering is not merely a training convention but an inductive bias that shapes accessible reasoning patterns, directly motivating our study of depth-dependent failure.

H. Appendix: Extended Discussion and Future Work

Deployment implications (extended). Models with $p_d < 0.93$ should be deployed with caution for tasks requiring >20 sequential reasoning steps. The three failure regimes imply distinct mitigation strategies: output-format retraining (D1); constraint-reminder prompting (D2); architectural or training-level intervention (D3) under current vanilla autoregressive inference.

Missing reasoning-specialised baselines (extended). The absence of o1/o3 and DeepSeek-R1 is the most significant limitation of this work. Process-level supervision (Cobbe et al., 2021) could plausibly shift p_d beyond the frontier tier in D2 or reduce D3's cascade collapse. If process-supervised models achieve $k^* > 10$ on D3 it would substantially qualify the scope of our findings. We commit to evaluating these models, alongside recursive scaffold (Yang et al., 2026) and fast-slow recurrent (Takashiro et al., 2026) architectures, in the next version.

Full limitations.

1. p_d model assumes independent per-step failure; true decay may be faster and p_d^N may overestimate accuracy (Remark 1 in main text).
2. CCB evaluates synthetic tasks; generalisation to naturalistic tasks requires further study.
3. Strict regex parser may under-count valid traces; a normalisation pass or AST fallback is future work.
4. All models evaluated at $T{=}0$ via OpenRouter; provider-specific optimisations may alter profiles.
5. D3 prompt sensitivity (60 trials, one model) is preliminary; broader prompt ablations (more variants, models, and depths) are needed.
6. o1/o3, DeepSeek-R1, and memory/tool-augmented systems are not evaluated; D3 claims are explicitly scoped to vanilla autoregressive inference.
7. Formal model selection (AIC/BIC) between decay formulations was not performed.
8. TFBC assumes a single canonical trace; alternative valid reasoning paths may be misclassified, though a targeted audit found no such cases in D3.

Future work roadmap.

1. Evaluate o1/o3 and DeepSeek-R1 (process supervision) on all three CCB domains.
2. Test recursive scaffold (Yang et al., 2026) and fast-slow recurrent (Takashiro et al., 2026) models on D3.
3. Formally compare geometric vs. accelerating decay ($p_d^{N^\gamma}$) via AIC/BIC at $N{>}50$ where positional-embedding saturation may shift the dominant failure mechanism.
4. Probe scratchpad-augmented decoding for D2 constraint violations.
5. Expand D3 ablations: explicit state tables, JSON schemas, graph-primitive formatting, multiple models, and larger n.
6. Release code, generation seeds, and parser to facilitate replication.