The Complexity Ceiling Benchmark: Measuring How LLM Reasoning Decays with Depth

Shubh Chapra [1] **Dhruv Kumar** [1] **Murari Mandal** [2] **Yash Sinha** [1]

Abstract

We introduce the Complexity Ceiling Benchmark (CCB), a controlled evaluation of how language-model reasoning decays as the number of required sequential steps grows. CCB fixes the semantic content of a task and varies only its depth $N \in \{5, \ldots, 50\}$ across three structurally distinct regimes: grounded spatial state-tracking, abstract symbolic pointer manipulation, and transitive relational inference. Across 6,000 trials over five frontier and open-weight LLMs we find a consistent pattern of geometric per-step decay with widely separated domain ceilings: on the first two regimes the strongest models retain $p_d > 0.92$ across $N{=}50$; on the third every model collapses by $N{=}5$, with the best model's 50%-success horizon at $H_{0.5}{\approx}4.7$ steps despite $p_d{=}0.863$. A trace-level metric (TFBC) shows that 14.5% of correct answers across the benchmark are reached via incorrect intermediate reasoning, with the share rising on harder domains. Forced verbose state-tracking does not move the ceiling (McNemar $p{=}1.000$), and the mean step at which reasoning first diverges, k^*, predicts within-domain accuracy better than parameter count. CCB and the geometric decay model together reduce a model's long-horizon reasoning profile to one interpretable number per task family.

1. Introduction

When a language model fails on a long reasoning task, current benchmarks cannot tell us why. The model may lack knowledge, misread the prompt, exhaust its context window, or simply lose its place across the intermediate steps. These failure modes have different remedies and different implications for deploying LLMs as agents, but aggregate accuracy on a fixed-difficulty benchmark sees only the union: a single number that says the model got it wrong. The most consequential of these modes for agentic use; the cumulative loss of coherence as the number of sequential steps grows; is also the one least visible to such benchmarks (Dziri et al., 2023; Sinha et al., 2025).

This paper treats depth as the controlled experimental variable. Figure 1 shows the central result of doing so. The Complexity Ceiling Benchmark (CCB) holds task semantics fixed and sweeps the number of required reasoning steps N from 5 to 50 across three structurally distinct regimes: grounded spatial state-tracking (D1), abstract symbolic pointer manipulation (D2), and transitive relational inference (D3). Six thousand trials over five frontier and open-weight LLMs reveal a consistent pattern; per-step retention decays geometrically in N; and a striking dissociation between domains. On D1 and D2 the strongest models hold retention probability $p_d > 0.92$ across the full depth range, leaving meaningful accuracy at $N{=}50$. On D3, every model collapses past $N{=}5$, regardless of capability tier; the best evaluated model's 50%-success horizon is only $H_{0.5}{\approx}4.7$ steps. Tracing the intermediate reasoning adds a second result that aggregate accuracy hides: across the benchmark, 14.5% of correct answers come from traces that diverged from the canonical reasoning, and the share is highest on the hardest domain.

The geometric decay model and the trace-level metric we introduce together let us summarise a (model, task-family) pair with two numbers: a per-step retention probability p_d and a mean trace-divergence step k^*. Both connect to deployment: p_d feeds the horizon-length framework of Sinha et al. (2025) and determines the depth at which sustained accuracy crosses any chosen threshold, and k^* predicts within-domain accuracy more faithfully than parameter count. The findings in this paper are scoped to vanilla autoregressive inference on the five evaluated models; evaluating process-supervised and tool-augmented systems is the obvious next step and is one of the uses CCB is built to serve.

[1]BITS Pilani, Pilani Campus [2]KIIT Bhubaneswar. Correspondence to: <f20230005@pilani.bits-pilani.ac.in>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

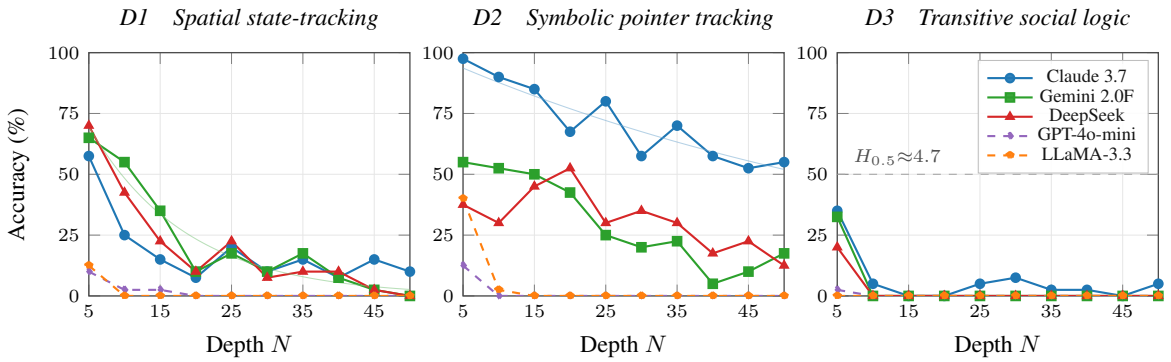

Figure 1. **The Complexity Ceiling.** Accuracy as a function of depth N across three structurally distinct reasoning regimes, with semantic content held fixed. Markers: empirical accuracy at $n{=}40$ trials per cell. Thin solid curves in D1 and D2: fitted geometric model $100{\cdot}p_d^N$ for the top frontier model (Gemini in D1, Claude in D2). On D1 and D2 frontier models track the geometric decay with $p_d{\in}[0.92, 0.99]$, leaving meaningful accuracy at $N{=}50$. On transitive social logic (D3) every model collapses past $N{=}5$ regardless of capability tier; even the best model's 50%-success horizon is $H_{0.5}{\approx}4.7$ steps. Solid lines: frontier or competitive open-weight; dashed: smaller models. 6,000 trials total.

2. Related Work

Most prior work treats reasoning failure as an aggregate property of a task. SCAN (Lake and Baroni, 2018) and BIG-Bench (Srivastava et al., 2022) probe compositional generalisation at roughly constant difficulty, providing a strong test of out-of-distribution generalisation but limited leverage for a scaling analysis. CLUTRR (Sinha et al., 2019) comes closest to our setting; multi-hop relational reasoning over kinship graphs; and motivates D3, but gives only final-answer judgements and a small fixed range of hops. Dziri et al. (2023) showed transformers unroll memorised subgraphs with catastrophic failure at compositional out-of-distribution depths; Hou et al. (2026) tied this to cumulative state-tracking load; SokoBench (Sebastiano Monti et al., 2026) isolates planning depth in Sokoban, TopoBench (Maniparambil et al., 2026) shows structured state aids reasoning. CCB differs from all of these by making depth a continuous parametric axis across three heterogeneous domains and by recovering a single one-parameter summary statistic that lets us compare *how* a model fails, not just whether it does.

A parallel line of work evaluates the reasoning chain itself rather than the answer. ROSCOE (Golovneva et al., 2022), ReCEval (Prasad et al., 2023), and MME-CoT (Jiang et al., 2025) introduce trace-quality metrics, but mostly rely on LLM-as-judge or learned rubrics whose own reliability is contested (Chaudhury et al., 2026). CCB sidesteps this by comparing against a deterministically generated canonical trace; the TFBC metric below requires no external scorer.

Finally, our p_d^N model is the empirical counterpart of the horizon-length framework of Sinha et al. (2025), which derives an effective task horizon $H_s{\approx}\ln(s)/\ln(p_d)$ from per-step accuracy. Cobbe et al. (2021) showed that process-level supervision shifts the relevant quantity; recursive scaffolds (Yang et al., 2026) and fast-slow recurrence (Takashiro et al., 2026) target the same bottleneck through architecture. Kim et al. (2025) argue that autoregressive token ordering is itself an inductive bias on what reasoning patterns are accessible. Extended discussion appears in Appendix C.

3. Benchmark Design

CCB consists of three task domains, a deterministic generator that produces ground-truth reasoning traces, a strict parser-based pipeline, and a one-parameter decay model fit per (model, domain) cell. The components are designed jointly: the failure taxonomy isolates the events the decay model is meant to describe, the parser produces the event counts the likelihood consumes, and the trace metric below operates on the same parsed structure.

The three domains share an experimental contract; semantic content held fixed, depth N varied over $\{5, 10, \ldots, 50\}$; but stress structurally distinct facets of long-horizon reasoning, together approximating a minimal basis for the regimes most often encountered in agentic settings: grounded spatial, abstract symbolic, and nested relational. *D1 Alien Grid* is a $3{\times}3$ grid subjected to N discrete transformations (ROTATE, SWAP CORNERS, SHIFT) under a grid-integrity constraint preventing entity collision; it tests grounded spatial state-tracking, and its errors compound deterministically because a misplaced entity at step k invalidates every state after it. *D2 Symbolic Pointer Tracking* asks the model to maintain seven variables A–G holding distinct digits 0–9 under N cyclic-shift and modular-arithmetic operations subject to an assignment-uniqueness constraint; its dominant failure mode is register corruption via illegal re-assignment, which accounts for 69.5% of D2 failures and reflects the difficulty of maintaining disjoint symbolic mappings over

long contexts. *D3 Social Logic* runs a diplomatic graph over ten agents under transitive-closure rules: at each of N update steps a new alliance or rivalry edge is added, the closure is recomputed, and all implied relationships are updated. The model is queried on the final pairwise state in a Theory-of-Mind-style format ("what does agent i believe about agent j?"), and must maintain global pairwise consistency over $O(n^2)$ relationships per step. Its dominant failure mode is *cascade collapse*; a single misclassified edge propagates by transitivity to every reachable node and is irrecoverable within the context window, which is what makes D3 a structurally distinct difficulty regime rather than a harder version of D1 or D2 (Section 4). We evaluate each domain at n=40 independently seeded trials per (model, depth) cell, for 2,000 trials per domain and 6,000 in total. All inference is at T=0 via OpenRouter, on *claude-3.7-sonnet*, *gemini-2.0-flash-001*, *deepseek-chat*, *gpt-4o-mini*, and *llama-3.3-70b-instruct*. Reasoning-specialised models (o1/o3, DeepSeek-R1) were not accessible at submission time and are addressed in Section 5.

Every trial falls into one of six disjoint categories: *Correct* (final answer and every step exact), *Reasoning* (parses cleanly but diverges from the canonical trace at some step $k^* \leq N$), *Constraint* (violates a structural task rule), *Format* (unparseable), *Truncation* (output ends mid-stream), and *API* (network error, auto-retried). Only the first three carry information about per-step retention; the rest are treated as missing data. The parser is regex-based and deliberately so: AST-based parsing and LLM-as-judge scoring were both considered and rejected to keep the evaluation deterministic and reproducible, and to remove any dependency on an external scoring model whose own reliability would have to be defended. The parser prioritises precision over recall; minor format deviations are flagged as Format failures rather than silently corrected, so the resulting p_d estimates are mildly conservative for models with idiosyncratic output styles. This is the bias direction we want because it cannot inflate reported accuracy or p_d. A manual audit of 150 edge-case outputs found no false positives in correctness classification and a small ($\approx 2\%$) false-negative rate from non-standard separators. The six-way classification is invariant to reordering of unrelated key/value pairs and to whitespace differences, so layout-only output changes do not affect the reported k^* distributions.

The TFBC metric. Correct final answers do not imply correct intermediate reasoning. We define *Trace First Branch Correct* (TFBC) to flag any trial whose final answer matches the ground truth but whose trace first diverges from the canonical reasoning at some step k^*>0. Algorithm 1 walks the trace from step 1 and returns the first divergence;[1] for incorrect parseable trials, k^* records the depth at which the model's working state first decohered. As a concrete illustration, a typical k^*=2 event on D1 at N=10 has ground truth Step 2 = $[[7,4,1],[8,5,2],[9,6,3]]$ (a 90° clockwise rotation) and model output $[[3,2,1],[6,5,4],[9,8,7]]$ (a horizontal flip); the trace diverges immediately at the first transformation and the error propagates deterministically through the rest of the trial, even when the final answer happens to coincide by chance. Three human annotators independently labelled traces against the automated extractor on subsets of each domain, with Cohen's κ of 0.977 (D1, n=50), 0.978 (D2, n=50), and 0.938 (D3, n=65), all above the conventional defensibility threshold of 0.80. D3 admits multiple valid reasoning paths in principle; a targeted audit of 20 randomly sampled D3 TFBC cases found no genuine alternative-path correctness, supporting the lucky-guess interpretation. Constraint violations and TFBC events are measured on different conditioning sets (TFBC is defined only over parseable correct outputs; constraint violations are a disjoint failure category) and so the high D2 constraint-violation share and the per-model TFBC rates in Table 1 are not in tension.

Algorithm 1 TFBC and k^* extractor.

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

Geometric decay model. Let p_d denote the probability that the model correctly computes $S_k \to S_{k+1}$ given that S_k was maintained. Under independence of per-step errors (Assumption 1),

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

The expression p_d^N is a discrete survival function with constant per-step hazard $1-p_d$; the interpretation we want is that a single free parameter per (model, domain) cell captures both the per-step failure rate and the resulting horizon at which any chosen success threshold is crossed. We fit p_d per cell by maximising the binomial log-likelihood

[1] An earlier version of the pseudocode lacked the **break** on line 8 and recorded the last rather than the first mismatch; all k^* statistics reported in this paper use the corrected version above.

$\sum_N \left[c_N \ln p_d^N + (n-c_N)\ln(1-p_d^N)\right]$ over c_N correct trials out of n=40 at each depth, constrained to $p_d \in [0.5, 1.0]$; 95% confidence intervals are from 2,000 parametric bootstrap resamples. The lower bound keeps the estimator in a meaningful regime: models driven to it (GPT-4o-mini and LLaMA on D3) should be read as exhibiting no statistically identifiable step-retention rather than as literal 50% per-step.

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 $k-1$.*

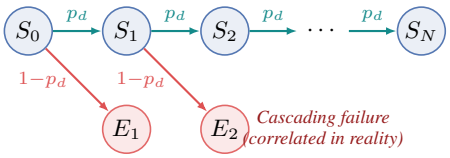

Figure 2. **The state-retention process of Assumption 1.** Under independence, a single step error transitions the system to an absorbing failure state with per-step probability $1-p_d$. In reality, errors at step k^* corrupt all $k>k^*$ (Remark 1), so the true decay is faster than p_d^N predicts.

Assumption 1 is wrong in a known direction (Figure 2). Autoregressive transformers are not Markov: an error at step k^* corrupts context for all $k>k^*$, so per-step errors are positively correlated and Equation 1 tends to *overestimate* sustained accuracy. The TFBC phenomenon — partial recovery from a corrupted trace to a correct final answer — softens but does not eliminate this bias, and we treat p_d^N as an optimistic envelope throughout. We considered two alternative forms before adopting the geometric model: an accelerating $p_d^{N^\gamma}$ that would capture attention fatigue, and a linear $1-\lambda N$ drift. The linear form fails to capture the non-linear collapse observed at moderate N; the accelerating form adds a free parameter without empirical motivation in the $N \leq 50$ regime. Across all three domains the one-parameter geometric form fits without exception and explains $> 90\%$ of accuracy variance on D1 and D2 (a residual analysis confirms no systematic structure in the depth-wise residuals). Formal AIC/BIC selection between these formulations beyond N=50, where positional-embedding saturation may shift the dominant failure mechanism, is left for future work.

The decay model also feeds the horizon-length analysis of Sinha et al. (2025): for minimum success threshold s, the effective task horizon is $H_s \approx \ln(s)/\ln(p_d)$. For Claude on D3 with p_d=0.863, $H_{0.5}$≈4.7 steps, which matches both the depth at which observed accuracy crosses 50% and the mean k^*=4.30 on incorrect D3 trials. The empirical p_d thus has a deployment-facing reading: the depth at which a model's expected accuracy on a task family falls below any chosen threshold.

4. Results

Table 1 summarises all 6,000 evaluations and Figure 1 shows the depth-resolved accuracy. The data support three claims: per-step retention decays geometrically across all five models and three domains; the D3 ceiling is qualitatively different from the D1/D2 ceiling and is not movable by prompt-level intervention; and the trace-divergence step k^* ranks models more faithfully within domain than parameter count or aggregate accuracy.

The decay curves separate the five models cleanly on D1 and D2. The three strongest models cluster around $p_d \in [0.92, 0.93]$ on D1 while GPT-4o-mini (p_d=0.688) and LLaMA-3.3 (p_d=0.634) collapse by N=15. D1 format adherence is essentially perfect (0–2% format failures), so the gap between tiers is in per-step retention rather than output structure. D2 sharpens the picture: Claude reaches p_d=0.987 and 71.2% aggregate accuracy, more than double Gemini (30.0%) or DeepSeek (31.3%), and retains 55% accuracy at N=50. Claude's D2 decay curve is visibly shallower than the non-frontier models'; a p_d=0.987 vs. p_d=0.634 gap corresponds to roughly a 28× lower per-step error rate, which is exactly the regime in which long-horizon agentic use becomes plausible. Of all D2 failures, 69.5% are constraint violations — illegal re-assignments of variables already bound. A canonical instance is a trace that produces `Step 4: var_A=10`, `Step 5: var_B=var_A` (so `var_B` should hold 10 for the rest of the trial), then writes `Step 6: var_A=15`, silently violating uniqueness on `var_A`. The model is performing arithmetic correctly at each step but failing to keep its bindings disjoint over the longer horizon, which is exactly the failure mode the assignment-uniqueness constraint was designed to expose. The occasional non-monotonic accuracy upticks at N=25–35 on D1 and D2 are within the per-cell Clopper-Pearson half-CI of ±12–16% at n=40 and do not reflect generator artefacts: operation distributions are stationary across N by construction, and the monotone p_d^N fit explains $> 90\%$ of variance on these domains.

D3 is qualitatively different. Across all five models and ten depth levels, 1,953 of 2,000 attempts fail. Only Claude achieves any sustained accuracy (6.2% overall, concentrated at N=5); every other model collapses to near-zero past depth 5 regardless of capability tier. The mean step at which reasoning first diverges, k^*, is uniformly low on D3 (2.88–4.30) across *all* models, including Claude; which carries k^*=17.67 on D2 yet diverges after only 4.30 steps on D3 (Table 3). The uniformity of k^* across models with markedly different capability on other domains is direct evidence that D3 poses a qualitatively distinct difficulty.

This is consistent with the computational structure of D3, which differs from D1 and D2 along three dimensions. First, transitive closure is not decomposable into independent per-

Table 1. **CCB master results.** Aggregate accuracy, the geometric step-retention MLE p_d, and the trace-level lucky-guess rate TFBC, per (model, domain) cell. Best per domain in **bold**; per-depth tables with 95% bootstrap CIs are in Appendix A. † At the optimiser lower bound; no statistically identifiable step-retention.

Model	D1 Spatial			D2 Symbolic			D3 Social Logic			Avg
	Acc	p_d	TFBC	Acc	p_d	TFBC	Acc	p_d	TFBC	
Frontier / closed-weight										
Claude 3.7	18.2%	0.929	21%	**71.2%**	**0.987**	8%	**6.2%**	**0.863**	56%	**31.9%**
Gemini 2.0F	**22.0%**	**0.930**	1%	30.0%	0.950	29%	3.2%	0.736	62%	18.4%
GPT-4o-mini	1.5%	0.688	17%	1.2%	0.634	20%	0.2%	0.500†	—	1.0%
Open-weight										
DeepSeek	19.8%	0.924	6%	31.3%	0.955	12%	2.0%	0.684	13%	17.7%
LLaMA-3.3	1.3%	0.634	0%	4.3%	0.767	18%	0.0%	0.500†	—	1.9%

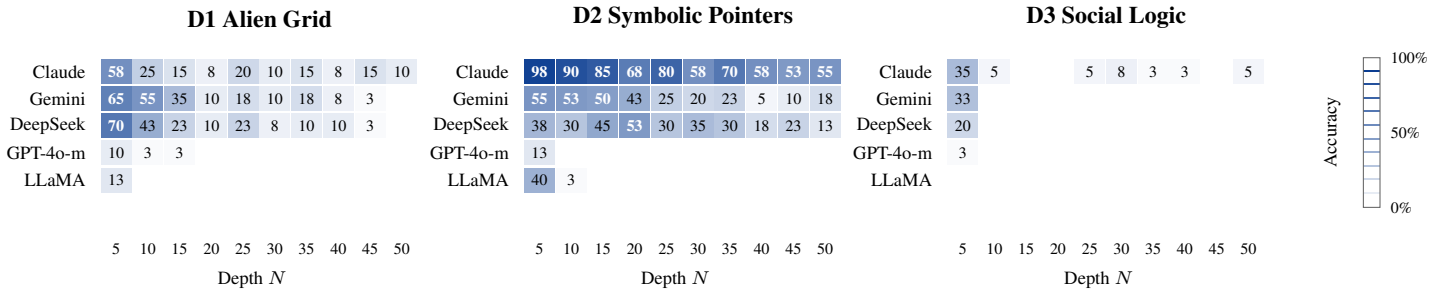

Figure 3. **Per-depth accuracy across models and domains.** Rows are models, columns are depth $N \in \{5, \ldots, 50\}$, cells show observed accuracy (%), darker = higher. The qualitative difference between domains is immediate: D1 and D2 retain non-trivial gradients at large N for frontier models, while D3 is essentially a single column of non-zero values at $N=5$. The view complements Table 1 (which aggregates across depths) and Figure 1 (which shows decay curves).

step updates: a single misclassified edge at step k propagates by transitivity to every k-hop reachable node, so unlike D1 (where errors compound locally) or D2 (where bindings can in principle be reread), the D3 state offers no per-step recovery path. Second, the task requires maintaining global pairwise consistency over $O(n^2)$ relationships per update, which competes for representational capacity with the linear context attention exposes. Third; and this is the operative mechanical claim; standard attention flattens sequence hierarchy into linear context, providing no mechanism for the kind of recursive stack management that transitive closure demands; as N grows the representations of distinct agents' relationship sets mix in the attention pattern rather than remaining cleanly partitioned, and the divergence step $k^* \in [2.88, 4.30]$ across models of very different capability is consistent with the prediction that this ceiling reflects an architectural rather than a capacity bottleneck. We offer this as an account of the collapse, not a proof of impossibility: tool-augmented systems with explicit graph state, or process-supervised models with step-level reward, could plausibly move the ceiling (Yang et al., 2026; Takashiro et al., 2026). CCB probes reasoning without such scaffolding; quantifying whether the ceiling moves when scaffolding is added is one of the uses the benchmark is built to serve. As a concrete illustration of the failure mode, a representative D3 cascade trace from DeepSeek at $N=10$ processes the first update correctly but omits a transitive closure propagation at step 2 (failing to mark a pair as allied that inherits the relation through a newly added edge); every subsequent pair classification inherits this corruption, and the model has no mechanism for retroactively correcting the state from within the context window.

A second axis of dissociation appears when we look at correct outputs rather than incorrect ones. A trace-level view reveals two qualitatively different populations among the answers an output-only evaluator would treat identically. *Trace-faithful correctness* ($k^*=-1$, TFBC false) is correctness with intact intermediate reasoning; *lucky-guess correctness* (TFBC true) is correctness despite demonstrably divergent reasoning. The dominant population shifts with domain. On D2, where Claude operates near $p_d=0.99$, only 8% of its correct outputs are TFBC and the rest are genuinely faithful (262 of 285 traces match the canonical reasoning); D2 is dominated by trace-faithful correctness. On D3 the picture inverts: 56%–62% of correct outputs from Claude and Gemini are TFBC, and the targeted audit (Section 3) found no genuine alternative-path correctness in this group, so D3 correctness is dominated by lucky-guess events. Aggregated across the benchmark, 14.5% of correct outputs are TFBC, so output-only evaluation over-

Table 2. **Per-domain results with format-failure rates and p_d bootstrap 95% CI widths.** The three domains are broken out separately so the per-domain spread, the uncertainty on p_d, and the format-failure share are all directly inspectable. CI widths (Δp_d) are the full 95% interval widths from 2,000 parametric bootstrap resamples; intervals straddle the point estimate symmetrically except where the optimiser lower bound p_d=0.5 truncates them. κ values are inter-annotator agreement on the k^* extractor. Best per domain in **bold**. $^\dagger p_d$ at the optimiser lower bound (no statistically identifiable step-retention).

D1 Alien Grid (κ=0.977)					**D2 Symbolic Pointers** (κ=0.978)					**D3 Social Logic** (κ=0.938)				
Model	Acc	p_d	Δp_d	TFBC	Model	Acc	p_d	Δp_d	TFBC	Model	Acc	p_d	Δp_d	TFBC
Gemini	**22.0%**	0.930	0.018	1%	Claude	**71.2%**	**0.987**	0.004	8%	Claude	**6.2%**	**0.863**	0.046	56%
DeepSeek	19.8%	0.924	0.021	6%	DeepSeek	31.3%	0.955	0.012	12%	Gemini	3.2%	0.736	0.118	62%
Claude	18.2%	0.929	0.019	21%	Gemini	30.0%	0.950	0.012	29%	DeepSeek	2.0%	0.684	0.167	13%
GPT-4o-m	1.5%	0.688	0.165	17%	LLaMA	4.3%	0.767	0.099	18%	GPT-4o-m	0.2%	$0.500^\dagger$	—	—
LLaMA	1.3%	0.634	0.205	0%	GPT-4o-m	1.2%	0.634	0.208	20%	LLaMA	0.0%	$0.500^\dagger$	—	—

Table 3. Mean divergence step k^* for incorrect trials. Within domain, k^* tracks accuracy; across models, parameter count does not; LLaMA-3.3 (70B) sits below Claude on every domain.

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

states reasoning quality *and* does so differentially; with the overstatement concentrated on precisely the harder domains where it most matters. A reader of the aggregate scores would conclude that Claude's 6.2% D3 accuracy reflects real partial competence; the trace-level evidence says that conclusion would be wrong for most of that 6.2%.

The distribution of k^*, not just its mean, carries information. Figure 4 shows D2 stratified by model: LLaMA fails early at the first symbolic transition, while Claude maintains accuracy across ∼17 steps before failing on global consistency. Early-heavy and late-heavy failure modes are qualitatively different and would call for different mitigations even at matched aggregate accuracy.

Taken together, k^* functions as a *working-memory-coherence depth statistic*: an operational measurement of how many sequential state updates a model can compose before its working representation decoheres. This framing explains the empirical regularity that parameter count is a poor cross-model predictor. LLaMA-3.3-70B has 70B parameters; more than most of the closed-weight comparators in this study; yet $k^*_{\text{LLaMA}} < k^*_{\text{Claude}}$ on every domain (3.45 vs. 8.45 on D1, 3.91 vs. 17.67 on D2, 3.01 vs. 4.30 on D3). Within a domain, k^* tracks accuracy because the depth at which a trace first decoheres mechanically lower-bounds the accuracy achievable beyond that depth; across models, the k^* ranking tracks the kind of reasoning capability that drives long-horizon agentic performance more faithfully than scale. We therefore propose k^* as a complement, not a replacement, to aggregate accuracy: a single (model, task-family) cell reports both p_d (the per-step retention) and k^* (the typical coherence depth) without requiring per-depth evaluation at use time.

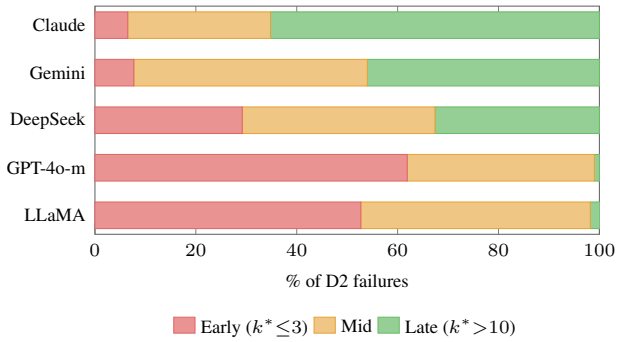

Figure 4. Divergence-step distribution on D2 by model. LLaMA fails at the first symbolic transition; Claude maintains accuracy across ∼17 steps before failing on global consistency.

A natural objection is that the D3 collapse reflects prompt phrasing rather than a genuine reasoning limit, since prompt sensitivity in LLM benchmarks is well documented and the cliff appears uniformly at N=5. We tested this with a paired ablation at N=15 on Claude, the only model with any D3 signal. The Standard condition lets the model infer state naturally; the Verbose condition prepends a CRITICAL instruction forcing it to restate the entire agent belief array after every operation. Both conditions yielded 0.0% accuracy across all n=20 paired instances, with the McNemar contingency in Table 4: zero discordant pairs and McNemar p=1.000. Token usage tells the same story; the Verbose condition spent ∼1,362 tokens to the Standard's ∼1,282 (a 6% overhead). The model acknowledged the instruction but spent the extra tokens restating beliefs it could not compute correctly, not on doing the computation differently. With zero discordant pairs the McNemar test confirms equipotence rather than distinguishing architecturally caused failure from coincidentally uniform failure, so we treat the result as suggestive rather than dispositive.

Table 4. D3 verbosity ablation contingency, N=15, Claude, n=20 paired. McNemar p=1.000.

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

To check whether *any* prompt structure shifts the ceiling we ran three further variants at the same N=15: VAR A imposes a strict output schema, VAR B adds positional formatting (P## slot markers identifying each agent position), and VAR C adds an explicit logical mapping of the alliance/rivalry update rule (Figure 5). Variants A and C remain at 0.0%. Variant B reaches 20.0%; a 20-point absolute spread that demonstrates D3 is prompt-sensitive in principle, but still falls far below practical utility. We do not yet have a mechanistic account of why slot-based positional formatting partially succeeds where the other variants do not, and whether the benefit extends to N>15 or to other models is reserved for the next iteration. The qualitative takeaway is the same as the Standard/Verbose result: prompt-level interventions can move the D3 ceiling by a few percentage points but do not change the underlying regime.

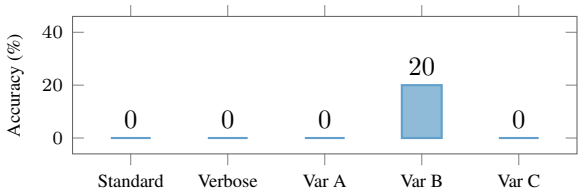

Figure 5. **D3 prompt-sensitivity at** N=15 (Claude 3.7, n=20 per condition). Standard/Verbose yield identical 0% (McNemar p=1.000); only VAR B (positional slot formatting) reaches non-zero accuracy, and at 20.0% this still falls far short of practical utility on long-horizon tasks.

Pulled together, the three domains identify three structurally distinct failure regimes that aggregate accuracy collapses into a single number: per-step retention bottleneck on D1 (where the frontier-vs.-non-frontier gap lies in p_d, not formatting); a constraint-management bottleneck on D2 (where 69.5% of failures are illegal re-assignments and the dominant error is state-keeping rather than arithmetic); and a structurally distinct cascade collapse on D3 (where k^* is essentially uniform across capability tiers and accuracy is dominated by lucky-guess events). No model dominates all three: Gemini leads on D1 (22.0%), Claude leads on D2 (71.2%) and D3 (6.2%), and the within-tier gaps shrink sharply as the load shifts from per-step retention to structural consistency; Claude's ∼53 pp lead on D2 collapses to a ∼3 pp lead on D3, evidence that the D3 ceiling is not a simple translation of the D2 ranking. The deployment-facing implication is direct: a model with p_d<0.93 should not be relied on for tasks requiring more than ∼20 sequential steps, with the optimism caveat of Remark 1 in mind.

5. Discussion

The combined empirical and analytical picture supports a *state-drift* view of autoregressive failure under depth scaling: each of the three CCB domains stresses a different facet of the same underlying step-retention bottleneck, and the three resulting failure regimes call for qualitatively different mitigations. Output-format retraining is the natural target for D1 (where the gap is in p_d and format-failure rates are already at 0–2%); constraint-reminder prompting or scratchpad-augmented decoding is the natural target for D2 (where the dominant failure is illegal re-assignment under a uniqueness rule); and architectural or training-level intervention; process supervision, recursive scaffolding, or explicit graph state; is the indicated direction for D3, since the verbosity ablation shows no movement from prompt-level changes alone. Because p_d is a single interpretable number per (model, domain) cell, it can be reported alongside aggregate accuracy on deployment dashboards without re-running depth-stratified evaluation at use time.

A skeptical reading of these results turns naturally to the D1 result. Format adherence is uniformly high across all five models (0–2% format failures), which rules out the obvious confound that the frontier-vs.-non-frontier gap is an artefact of structured output. Frontier models maintain $p_d \in [0.924, 0.930]$; GPT-4o-mini (p_d=0.688) and LLaMA (p_d=0.634) collapse on per-step retention, not on formatting. The implication is that the D1 ceiling is not closeable by output-format prompting alone, and the right diagnostic target for closing it is per-step retention rather than output adherence. Whether targeted format-aware fine-tuning shifts p_d on D1 is a clean empirical question we leave open.

The principal limitation of this work is the absence of reasoning-specialised baselines. Process-level supervision (Cobbe et al., 2021) is the most plausible single intervention for shifting p_d beyond the frontier tier on D2 or for reducing the D3 cascade; if process-supervised models achieve k^*>10 on D3 it would substantially qualify the architectural reading of our findings. We commit to evaluating o1/o3, DeepSeek-R1, recursive scaffold models (Yang et al., 2026), and fast-slow recurrent mechanisms (Takashiro et al., 2026) alongside tool-augmented systems with explicit graph state in the next iteration. This is the single most informative follow-up experiment available, and the one most likely to sharpen or falsify the structural reading of D3.

Remark 1 (The bound is optimistic)**.** *An error at step k^* corrupts context for all k>k^*, so per-step errors are positively correlated and the true decay is faster than p_d^N predicts. The TFBC phenomenon; partial recovery from a corrupted context to a correct final answer — softens but does not eliminate this effect.*

The remaining limitations are quantitative and we have tried to bound them rather than eliminate them. The independence assumption behind p_d^N is wrong in a known direction (Assumption 1, Remark 1); real autoregressive decoding produces positively correlated errors, so the model's predictions should be read as an optimistic envelope on sustained accuracy rather than a guarantee. The synthetic generators produce structurally controlled tasks but their generalisation to naturalistic agentic settings requires further study. The strict regex parser may under-count valid traces with idiosyncratic formatting (we measured $\approx 2\%$ false-negative rate in the 150-case audit). All evaluations were performed at $T{=}0$ via OpenRouter; provider-specific optimisations may shift profiles slightly. The D3 prompt ablation is preliminary (one model, $n{=}20$ paired) and a broader sweep over prompt structures is needed to distinguish architectural from coincidental failure cleanly. Formal AIC/BIC selection between geometric, accelerating $p_d^{N^\gamma}$, and linear $1{-}\lambda N$ decay forms at $N{>}50$; where positional-embedding saturation may shift the dominant failure mechanism; remains future work. Finally, TFBC assumes a single canonical reasoning trace per trial; alternative valid traces would be misclassified, though the 20-case D3 audit found no such cases. The 150-case parser audit and the $\kappa{\geq}0.938$ inter-annotator agreement support that the reported numbers are not parser artefacts.

6. Conclusion

The Complexity Ceiling Benchmark isolates reasoning depth as a controlled experimental variable and reveals three structurally distinct failure regimes that aggregate accuracy collapses into a single number: per-step retention collapse on grounded spatial reasoning, constraint-management collapse on abstract symbolic reasoning, and a transitive-closure cascade on relational reasoning that persists uniformly across all five evaluated models regardless of parameter count. The trace-level TFBC analysis shows that 14.5% of correct outputs across the benchmark are reached via demonstrably divergent intermediate reasoning, so trace-faithful and lucky-guess correctness must be distinguished for benchmark scores to remain meaningful on long-horizon tasks. The mean trace-divergence step k^* predicts within-domain accuracy more faithfully than parameter count, supporting a state-drift rather than a capacity-limit account of autoregressive failure under depth scaling. The most urgent open question is whether process-supervised or recursive architectures dissolve the D3 ceiling; if they do, the structural reading of these findings requires substantial revision; if they do not, CCB provides a principled diagnostic for the next generation of memory-augmented systems.

References

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Bhatt, Lianhui Jiang, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

Dengzhe Hou, Lingyu Jiang, Deng Li, Zirui Li, Fangzhou Lin, and Kazunori D. Yamada. Wmf-am: Probing llm working memory via depth-parameterized cumulative state tracking. *arXiv preprint arXiv:2603.27343*, 2026.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, 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.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. 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.

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

Mayug Maniparambil, Nils Hoehing, Janak Kapuriya, Arjun Karuvally, Ellen Rushe, Anthony Ventresque, Noel O'Connor, and Fergal Reid. Topobench: Benchmarking llms on hard topological reasoning. 2026. URL https://arxiv.org/abs/2603.12133.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*, 2023.

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

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in llms. *arXiv preprint arXiv:2509.09677*, 2025.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Shota Takashiro, Masanori Koyama, Takeru Miyato, Yusuke Iwasawa, Yutaka Matsuo, and Kohei Hayashi. Thinking while listening: Fast-slow recurrence for long-horizon sequential modelling. *arXiv preprint arXiv:2604.01577*, 2026.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Recursive models for long-horizon reasoning. *arXiv preprint arXiv:2603.02112*, 2026.

A. Detailed Results

This appendix supplements the per-cell summary in Table 1 and the per-depth heatmap in Figure 3 with full numeric values: Tables 5–7 summarise across depths with 95% bootstrap CIs, and Tables 8–10 give the underlying depth-wise accuracies with Clopper-Pearson half-CIs.

Table 5. 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%

Table 6. D2 Symbolic Pointer Tracking (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%

Table 7. D3 Social Logic (n=400/model). Human κ=0.938, n=65. McNemar verbosity ablation p=1.0. p_d at 0.500 is at the optimiser lower bound.

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%	$0.500^{\dagger}$ [0.500, 0.611]	$*$	0.0%
LLaMA-3.3	0.0%	$0.500^{\dagger}$ [0.500, 0.611]	—	0.0%

$*$ Single correct instance; TFBC unreliable.

B. Failure Mode Examples and Prompt Templates

Failure Mode Examples

D1 early divergence (LLaMA, 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)
```

Horizontal flip instead of 90° CW rotation; k^*=2.

D2 constraint failure (Gemini, N=25). After twenty correct steps the model illegally re-assigns variable A, violating uniqueness. Classified Constraint, not Reasoning.

D3 cascade collapse (DeepSeek, N=10). The model processes step 1 correctly but omits a transitive closure propagation at step 2. Every subsequent pair classification is wrong; recovery is impossible within the context window.

Exact Prompt Templates

D1.

```
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 deg clockwise.
  SHIFT_ROW_2_LEFT: Shift middle row left,
                    wrapping.
Output:
  TRACE: ["Step 1:[[...]]", ...]
  ANSWER: [[...]]
```

D2.

```
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:
  TRACE: ["Step 1:{A:v,...}", ...]
  ANSWER: {A:v, B:v, ...}
```

D3 verbose ablation addition.

```
CRITICAL ABLATION INSTRUCTION: After EVERY
single operation, you MUST explicitly
restate the entire agent belief state array
before proceeding to the next step.
```

C. Extended Related Work

Depth-scaling and compositional generalisation. SCAN (Lake and Baroni, 2018) and BIG-Bench (Srivastava et al., 2022) hold difficulty roughly constant and probe systematic generalisation to novel compositions. CCB complements that line by providing a continuous, parametric depth axis across three heterogeneous domains and a single-parameter decay model. SokoBench (Sebastiano Monti et al., 2026) isolates planning depth in Sokoban; TopoBench (Maniparambil et al., 2026) focuses on topological reasoning and shows that structured state aids reasoning, motivating tool-augmented D3 extensions.

Table 8. D1 Alien Grid: Acc% ± Clopper-Pearson half-CI by depth. AvgTok = mean response tokens.

N	Claude	Gemini	DeepSeek	GPT-4o-m	LLaMA	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 9. D2 Symbolic Pointers: Acc% ± half-CI by depth.

N	Claude	Gemini	DeepSeek	GPT-4o-m	LLaMA	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

Table 10. D3 Social Logic: Acc% ± half-CI by depth.

N	Claude	Gemini	DeepSeek	GPT-4o-m	LLaMA	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

Relational benchmarks. CLUTRR (Sinha et al., 2019) tests multi-hop relational reasoning on kinship graphs and is the closest prior work to D3. CCB extends that line by providing deterministic ground-truth traces (not only final answers), enabling TFBC-level diagnostics; by applying a continuous depth axis from $N{=}5$ to $N{=}50$; and by integrating relational inference with spatial and symbolic regimes under a unified evaluation framework.

State tracking and systematic failures. Dziri et al. (2023) showed LLMs unroll memorised subgraphs with catastrophic failure at compositional OOD depths. Hou et al. (2026) showed performance degrades under cumulative state-tracking load. CCB quantifies these phenomena via the k^* distribution and the p_d summary statistic.

Trace-level evaluation and structural uncertainty. 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. Chaudhury et al. (2026) show that unstable self-preference rankings signal unreliable inference. CCB provides a complementary, ground-truth-grounded operationalisation that requires no LLM-as-judge.

Process supervision and long-horizon execution. Process-supervised models (Cobbe et al., 2021) are trained with step-level reward signals that incentivise intermediate-state correctness; their evaluation is the most consequential extension of this work. Sinha et al. (2025) analytically links per-step accuracy to an effective task

horizon $H_s \approx \ln(s)/\ln(p_d)$; CCB's empirical p_d values feed directly into that framework. Recursive scaffolds (Yang et al., 2026) and fast-slow recurrence (Takashiro et al., 2026) target the same state-management bottleneck from the architecture side, and Kim et al. (2025) argue that autoregressive token ordering is itself an inductive bias on accessible reasoning patterns.