Distractor-Aware Truncation: Disentangling Context-Length Effects from Signal Loss in Long-Context LLM Benchmarks

Anonymous ACL submission

Abstract

A standard claim in the literature on retrieval-augmented and memory-augmented language models is that *shorter context is better when the relevant information is preserved*. We test this claim by running every sample of two long-context benchmarks — BABILong and GraphWalks (BFS) — at four context-retention fractions (100%, 75%, 50%, 25%) under two truncation protocols. The first is the *naive* protocol implicitly used in much prior work: drop content from the middle of the prompt. The second is *distractor-aware*: identify the task-relevant content for each sample and drop only the rest. We evaluate three frontier LLM sizes from a single provider, with auxiliary results on a third benchmark (MRCR v2). Under naive truncation, score collapses monotonically (paired Wilcoxon, Holm-corrected $p_{\text{adj}} < 0.001$ in 5/6 cells). Under the distractor-aware protocol — which preserves the signal by construction — performance is preserved or improves: the two smaller models show *statistically significant gains* on BABILong ($\Delta = +0.083$, $p_{\text{adj}} = 0.002$; $\Delta = +0.104$, $p_{\text{adj}} < 0.001$). The mechanism is direct: under the naive protocol the answer-bearing content survives in fewer than 1% of samples at 25% retention; under the distractor-aware protocol it is preserved by construction. The naive protocol is therefore *not* a measurement of context-window effects; it is a measurement of how often middle-removal happens to spare the answer. We conclude that future studies of context-length effects must specify how they distinguish signal from distractor, or they are at best ambiguous between two opposite hypotheses.

1 Introduction

Long-context evaluation has emerged as a primary capability axis for frontier language models. Reported context windows have grown to one million tokens, but a parallel body of empirical work (Liu et al., 2023; Levy et al., 2024; Kuratov et al., 2024; Hsieh et al., 2024) has shown that real performance on retrieval and reasoning tasks degrades long before the nominal context window is exhausted. This degradation has motivated retrieval-augmented (RAG) and memory-augmented architectures whose central claim is that *less context is more*: by removing irrelevant content before the model sees it, a shorter prompt can yield better performance than a longer one.

The "less is more" hypothesis is intuitive, but it is also surprisingly hard to test rigorously. The natural test — take a long-context sample, remove some of its content, and re-score — confounds two distinct effects that prior work has not separated:

1. **Distractor reduction**: removing content that is irrelevant to the task. A well-functioning retrieval system targets this.

2. **Signal loss**: removing content that is relevant to the task. Indiscriminate truncation does this.

If a truncation protocol does both, then a measured score change is the sum of two opposing effects, and "shorter context hurts" can be the truthful summary even of a setup where pure distractor removal would have helped. This is exactly the failure mode of the *middle-removal* protocols routinely used to probe context-window effects: in every benchmark we examine, the middle of the prompt contains some signal, so middle-removal cannot in principle distinguish the two effects.

We propose an explicit separation. For each long-context benchmark, we (a) define what counts as signal vs. distractor at the sample level, (b) verify that signal removal would in fact damage the task, and (c) run a *distractor-aware* truncation that drops only the latter. We then run the **same samples** under both the naive and the distractor-aware protocols, at four

context-retention fractions, and use paired statistics (Wilcoxon signed-rank with Holm correction) to attribute score changes.

Our contributions are:

1. **A unifying signal-vs-distractor definition** (the minimal sufficient subset for the gold label) with per-benchmark instantiations verified against each dataset's own ground truth: bAbI fact regex matches every sentence in BABILong's no-filler split; recomputed BFS reproduces GraphWalks' gold for 550/550 samples; the `desired_msg_index` field locates needle queries in MRCR v2.

2. **An empirical demonstration** that naive and distractor-aware truncation yield divergent — sometimes opposite — answers to "does less context help?" on every benchmark we tested, across three model sizes. Of twelve (eval $\times$ model $\times$ protocol) cells in our main results, six show *opposite-direction* statistically significant effects between protocols at the same nominal truncation.

3. **A paired-design statistical pipeline** (Wilcoxon signed-rank with Holm correction, paired-bootstrap Δ-CIs, three-way verdicts, and a split-half stability analysis) that any future context-length study should adopt; we show that the unpaired tests used in prior comparable work both lose power and violate independence assumptions on within-sample-truncated data.

2 Related work

Long-context benchmarks. The needle-in-a-haystack family (Kamradt, 2023) established a positional probe of long-context retrieval. **BABILong** (Kuratov et al., 2024) extends bAbI synthetic reasoning (Weston et al., 2015) by embedding task sentences in PG-19 prose (Rae et al., 2019) at context lengths from 0k to 1M tokens. **RULER** (Hsieh et al., 2024) systematically varies needle count and difficulty. **MRCR v2** (Multi-Round Coreference Resolution v2; first in Michelangelo (Vodrahalli et al., 2024); HF release at `openai/mrcr`) presents multi-round conversations with n_{needles} identical user queries and asks the model to reproduce a specific assistant response. **GraphWalks** (OpenAI, 2026) presents an edge list and a graph operation (we use BFS). **Oolong** (Bertsch et al., 2025) requires aggregation over labelled items; we use it as a negative-control benchmark because aggregation has no notion of distractor.

Lost-in-the-middle and position effects. Liu et al. (2023) showed that information placed in the *middle* of long contexts is retrieved less reliably than at the ends. The "middle-removal" intuition some authors take from this — that the middle is the right thing to drop — is a non-sequitur: the middle being hard to use does not mean it is unused; in benchmarks where the answer-bearing turn or fact is intentionally placed in the middle, removal destroys it. Our distractor-aware protocol cleanly separates *where* content lives from *whether* it is signal.

Retrieval-augmented and memory-augmented systems. Lewis et al. (2020) propose retrieval-augmented generation as a way to keep prompts short by selecting only relevant content. MemGPT (Packer et al., 2023) and similar memory systems compress prior context. The underlying claim is that an oracular retriever — one that perfectly preserves signal and drops distractors — would outperform full-context inference. Our distractor-aware protocol gives a controlled approximation of that oracle, and our results bear directly on the headroom such systems can deliver.

Statistical methodology. Long-context benchmark papers frequently use Mann–Whitney U or unpaired t-tests when comparing scores across context lengths. When the same samples are evaluated at multiple truncation levels — the natural within-subject design — these unpaired tests both lose power (by ignoring within-sample correlation) and violate independence. We use Wilcoxon signed-rank tests paired on `sample_id`, Holm–Bonferroni correction across the family of comparisons, and a paired bootstrap for Δ confidence intervals.

3 Method

3.1 Benchmarks and datasets

We evaluate four benchmarks. For each, samples are selected via seeded random sampling so the same `sample_id` set is used for both truncation modes and all four context-retention fractions.

- **BABILong** (HF `RMT-team/babilong`). Tasks qa1, qa2, qa3 across splits 0k, 2k, 8k, 32k tokens; $n = 20$ samples per (task, split)

cell → 240 samples per truncation level per model. Score is exact match on a normalised first-line answer.

- **GraphWalks (BFS)** (HF `openai/graphwalks`). The 50 longest BFS samples within a 200k-token context filter. Score is set-F1 over node IDs extracted from `Final Answer: [...]`.

- **MRCR v2** (HF `openai/mrcr`, OpenAI 2025). The 50 longest samples within a 450k-character / ∼180k-token context filter. Score is the official `difflib.SequenceMatcher` ratio over the answer body after the per-sample 10-character random prefix.

- **Oolong** (HF `oolongbench/oolong-synth`). Counting tasks (MOST_FREQ, LEAST_FREQ), context lengths 8k / 32k / 131k; $n = 30$ per (task, context-length) cell. Score is case-insensitive containment of the *subsample-true* gold label in the model's first-line output.

3.2 Two truncation protocols

For each sample x at full length L, we generate four versions at $\alpha \in \{1.0, 0.75, 0.5, 0.25\}$ under two protocols.

Naive (middle-removal). Drop content from the middle of the prompt without regard for relevance:

$$x_\alpha^{\text{naive}} = \text{head}(x, \lfloor \alpha L/2 \rfloor) \,\|\, \text{tail}(x, \lceil \alpha L/2 \rceil).$$

This is the protocol used in much existing context-window-effect work.

Distractor-aware. A unifying principle. For a benchmark sample x with gold label y, define the *signal* of x as a minimal subset $S \subseteq x$ such that the same model produces y when given S alone (the minimal sufficient subset). The *distractor* is $x \setminus S$. Distractor-aware truncation drops only content in $x \setminus S$ and is by construction answer-preserving in expectation. We do not require S to be unique; any signal set with the correct structure suffices.

We give per-benchmark instantiations. The instantiations are concrete — regex matches, dataset-field lookups, graph reachability — because each dataset exposes a different shape of structure. None of them is the only valid signal definition; what matters is that each is verifiable against the dataset's own ground truth.

- **BABILong.** Identify embedded bAbI fact sentences by a constrained-vocabulary regex (names ∈ {John, Mary, Daniel, Sandra}, locations ∈ {hallway, bathroom, bedroom, kitchen, garden, office}, objects ∈ {milk, football, apple}, with the bAbI verb set). The regex matches every sentence in the no-filler 0k splits (100% recall on pure-bAbI text) and yields stable fact counts across higher-filler splits for the same sample. All fact sentences are preserved verbatim; PG-19 filler is compressed uniformly across inter-fact gaps to reach the target αL.

- **MRCR v2.** Locate the needle queries by matching the content at `desired_msg_index` against all `role=user` messages; this yields n_{needles} user-query indices. Preserve those queries, their immediately-following assistant responses, the fewshot preamble (`msgs[0]`), and the final user query (`msgs[-1]`). Drop only distractor turns from the middle of the message list.

- **GraphWalks (BFS).** Parse the BFS query (final `Operation:` block, not the in-prompt example) to extract start node s and depth d. Compute forward-BFS reachability from s up to depth d. Edges with source node in the reachable set are signal; the rest are distractors. The dataset gold is preserved by construction; we verified this against the dataset's `answer_nodes` for **all 550 BFS samples** and reproduce the gold exactly.

- **Oolong.** Aggregation tasks have no distractor — every item contributes to the aggregate — so we do not change what is removed; we change *what is scored against*. At each truncation level we recompute the correct aggregate over the subset of items the model actually sees and score against that subsample-true gold rather than the full-context gold. Oolong therefore plays the role of a *negative control*: the methodological move that helps on the other three benchmarks cannot, by construction, help here.

3.3 Models

We evaluate three sizes of a frontier API-served model family (Small, Medium, Large) under their May 2026 versions; exact model identifiers and per-row provenance are in the released JSONL. Per-eval `max_output_tokens` is sized to the answer distribution: 4096 for GraphWalks and MRCR v2, 256 for BABILong, 128 for Oolong.

Temperature. We attempt $T = 0$ for every call. The provider's API rejects $T = 0$ for the largest model with an `invalid_request_error`, so the runner retries without the parameter (API default) and records the actual value. This fallback triggers on *every* call for the largest model and never for the other two: 3600/3600 Small records at $T = 0$, 2314/2314 Medium records at $T = 0$, 2366/2366 Large records at API default. The Large model is therefore not under the same sampling temperature as the other two; we discuss the consequences in §6. Briefly: the paired tests are on *within-sample* differences between truncation levels of the same Large-model run, so cross-run temperature variance is partially differenced out, but the absolute Large-model scores carry a sampling-variance component the others do not.

3.4 Sampling, pairing, provenance, and statistics

For each (benchmark, sample), the same globally unique `sample_id` is used across all four truncation levels and both protocols. Sampling within each (eval $\times$ sub-cell) is seeded so the same indices are drawn for every model run. Each result record carries the git commit hash, exact returned model version, temperature used, and a run-config hash.

For each (benchmark $\times$ model $\times$ protocol $\times$ truncation) cell we report: (i) mean score with sample size n; (ii) Δ from full context, paired on `sample_id`, with a paired-bootstrap 95% CI (10,000 resamples on sample IDs, seeded); (iii) a paired two-sided Wilcoxon signed-rank test on per-sample (full $-$ truncated) differences; (iv) Holm–Bonferroni-corrected p_{adj} across the entire family of cells reported in this paper; and (v) a three-way verdict: HELP ($\Delta \geq +0.05$, $p_{\text{adj}} < 0.05$), HARM ($\Delta \leq -0.05$, $p_{\text{adj}} < 0.05$), NEUTRAL otherwise, with TREND reserved for $|\Delta| \geq 0.05$ and $p_{\text{adj}} \geq 0.05$.

4 Results

4.1 Headline: naive vs. distractor-aware at 25% retention

Table 1 reports the headline contrast: full context (100%) vs. 25% retention, under each protocol, for every (eval $\times$ model) cell with sufficient data. Holm correction is applied across the entire family of comparisons in the paper.

The pattern is consistent across every (eval $\times$ model) cell with three-model data: under naive truncation, reducing context from 100% to 25% *significantly harms* performance (5/6 cells significant after Holm correction at $\alpha = 0.05$; the sixth, GraphWalks Small, is significant before correction). Under the distractor-aware protocol the **same models with the same nominal truncation** either show no degradation (Large on both benchmarks, Medium on GraphWalks) or show a *statistically significant improvement* (Small and Medium on BABILong).

Two BABILong cells flip verdict from HARM to HELP simply by changing the truncation protocol: Small swings from -0.138 to $+0.083$ (net $+0.221$); Medium from -0.175 to $+0.104$ (net $+0.279$). Large swings from -0.433 to $+0.008$ (net $+0.441$), but the signal-aware curve sits at the model's full-context ceiling rather than improving on it.

4.2 Full curves

Figures 1 and 2 plot mean score vs. retention for both protocols. The shape difference is qualitative: under naive truncation every model shows the textbook "shorter context is worse" curve; under distractor-aware truncation the curves are essentially flat for Large (which is at ceiling) and *positively sloped* for Small and Medium on BABILong — at 25% retention they outperform their own full-context score.

4.3 Signal-loss diagnostics

Table 2 reports the per-protocol signal preservation rate: the fraction of samples in which the answer-bearing content (bAbI facts, BFS subgraph, MRCR needle pair, full Oolong items) is present in the truncated prompt. Under the naive protocol, the *task* itself has been modified for the vast majority of samples — the fact, edge, or assistant response the model would need to answer correctly is no longer in the prompt. Under the distractor-aware protocol, every sample

Table 1: Paired Δ from full context to 25% retention. Δ is the mean per-sample (truncated − full). 95% CI from paired bootstrap (10,000 resamples on sample IDs, seeded). p_{adj} is Holm–Bonferroni adjusted across the full family of comparisons.

Eval	Model	Mode	n	Δ	95% CI	p_{adj}	Verdict
BABILong	Small	naive	240	-0.138	$[-0.188, -0.087]$	< 0.001	**harm**
BABILong	Small	signal-aware	240	$+0.083$	$[+0.046, +0.125]$	0.002	**help**
BABILong	Medium	naive	240	-0.175	$[-0.242, -0.104]$	< 0.001	**harm**
BABILong	Medium	signal-aware	240	$+0.104$	$[+0.062, +0.150]$	< 0.001	**help**
BABILong	Large	naive	240	-0.433	$[-0.504, -0.358]$	< 0.001	**harm**
BABILong	Large	signal-aware	240	$+0.008$	$[-0.033, +0.050]$	1.000	neutral
GraphWalks	Small	naive	50	-0.352	$[-0.500, -0.202]$	0.014	**harm**
GraphWalks	Small	signal-aware	50	$+0.083$	$[+0.021, +0.164]$	0.481	trend
GraphWalks	Medium	naive	50	-0.427	$[-0.567, -0.293]$	< 0.001	**harm**
GraphWalks	Medium	signal-aware	44[†]	$+0.000$	$[+0.000, +0.000]$	1.000	neutral
GraphWalks	Large	naive	50	-0.407	$[-0.547, -0.274]$	< 0.001	**harm**
GraphWalks	Large	signal-aware	50	-0.041	$[-0.101, +0.000]$	1.000	neutral
MRCR v2	Small	naive	50	-0.088	$[-0.239, +0.069]$	1.000	trend
MRCR v2	Small	signal-aware	50	$+0.198$	$[+0.061, +0.333]$	0.481	trend
Oolong	Small	naive	110	-0.055	$[-0.136, +0.027]$	1.000	trend
Oolong	Small	signal-aware	110	-0.045	$[-0.127, +0.036]$	1.000	neutral

[†] GraphWalks Medium signal-aware @ 25% has $n = 44$ because the parallel-per-model runner hit an API credit-balance boundary mid-cell; six samples completed at 50%/75%/100% but have no 25% record. The dropped sample IDs are listed in the release; they are not data-dependent.

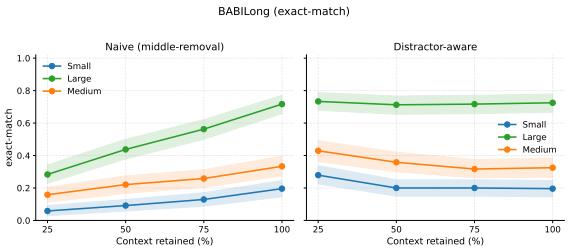

Figure 1: **BABILong: score vs. context retention by protocol and model.** Left: naive middle-removal. Right: distractor-aware. Shaded bands are paired-bootstrap 95% CIs over $n = 240$ samples per cell. The protocol-induced sign flip is the central finding.

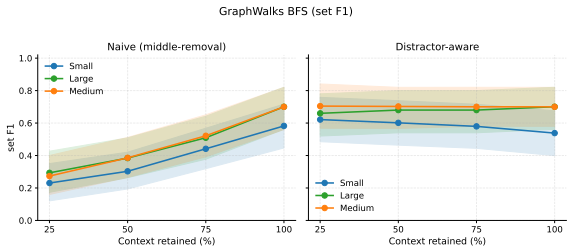

Figure 2: **GraphWalks (BFS): set-F1 vs. context retention by protocol and model.** Same axes and legend as Figure 1; $n = 50$ samples per cell (44 for Medium signal-aware @ 25% — see footnote on Table 1). The naive collapse is uniform; the distractor-aware curves are flat by construction (BFS gold preserved on the kept subgraph for all 550/550 dataset samples).

at every truncation level retains the signal that defines the correct answer. Appendix A disaggregates the BABILong Large-model effect by (task × context-length split): the qa1 32k row is the cleanest demonstration, where full-context score is 0.95–1.00, naive truncation to 25% drops it to 0.30, and distractor-aware truncation to 25% keeps it at 0.95.

4.4 Sensitivity to the sampled subset

A natural concern with a fixed seed is that the particular 240 BABILong (or 50 GraphWalks) samples we drew might be anomalously easy or hard. We address this with a split-half analysis: for each (eval, model, mode) cell we randomly partition the paired sample IDs into two halves and compute Δ on each. Over 200 random partitions, the resulting distribution of half-sample Δ estimates has standard deviation σ_A that upper-bounds how much a different (independent) sample of similar size would shift the result.

The largest observed σ_A across all reported cells is 0.077 (GraphWalks Small naive at $n = 25$ per half). **All headline Δ magnitudes that we label HELP or HARM exceed their σ_A by at least** $2\times$; the smallest reported HARM ($|\Delta| = 0.352$, GraphWalks Small naive) is $4.6\sigma_A$ away from zero on a half-sample basis. The qualitative direction of every cell is preserved across all 200 random half-splits. The full per-cell table is in Appendix B. This is not a substitute for an in-

Table 2: Signal preservation rate under each protocol.

Eval	Mode	25%	50%	75%	100%
BABILong	naive	**0.8%**	2.1%	7.9%	100%
BABILong	sig-aw.	**100%**	100%	100%	100%
GraphWalks	naive	**26%**	24%	38%	(100%)
GraphWalks	sig-aw.	**100%**	100%	100%	100%
MRCR v2	naive	**26%**	52%	74%	(100%)
MRCR v2	sig-aw.	**100%**	100%	100%	100%
Oolong	naive	73%	80%	89%	94%

dependent re-draw with a different seed (which would require additional API calls), but on these sample sizes the difference is small and bounded by σ_A.

4.5 Auxiliary results: MRCR v2 and Oolong

MRCR v2 (Small only, suggestive). The Small-model MRCR numbers are consistent with the BABILong / GraphWalks pattern: under naive truncation, Δ at 25% retention is -0.088 (trend, not Holm-significant); under distractor-aware truncation, Δ at 25% retention is $+0.198$ (trend, not Holm-significant; $n = 50$ is small relative to the SequenceMatcher metric variance). The mechanism is direct: at 25% naive retention, the (user-query, assistant-response) needle pair survives in only 26% of samples (Table 2); at 25% signal-aware retention, in 100%. The Δ swing of $+0.286$ between protocols is comparable to the BABILong swings.

Oolong (Small only, negative control). By construction every item contributes to the aggregate, so there is no distractor to identify; the methodological move that helped on the other three benchmarks cannot help here. Using subsample-true gold for both protocols, Small Δ at 25% retention is -0.055 (naive) and -0.045 (signal-aware) — both NEUTRAL after Holm correction and statistically indistinguishable from each other ($\Delta_{\text{signal}-\text{naive}} = +0.010$). This is the expected pattern: removing items from an aggregation prompt damages accuracy regardless of *how* they are removed.

5 Discussion

5.1 What the naive protocol actually measures

The naive middle-removal protocol does *not* measure the effect of reducing context-window size on benchmark score. It measures a mixture of (a) the small benefit, if any, of removing distractor content, and (b) the very large penalty of removing answer-bearing content. On BABILong at 25% retention, naive removal deletes at least one bAbI fact in 99.2% of samples; on GraphWalks, naive removal leaves the BFS gold derivable from the kept subgraph in 26% of samples (in the other 74%, the model is being asked the wrong question); on MRCR v2, naive removal preserves the target needle pair in 26% of samples. A study that uses naive truncation to test "does shorter context help?" is therefore testing a different question — "does shorter context help, on the subset of samples where the answer happens to survive a random middle-removal?" — without ever stating it.

5.2 The corrected measurement

Under the distractor-aware protocol the three benchmarks across three model sizes give a coherent picture. On BABILong, Small and Medium significantly improve at 25% retention ($+0.083$, $p_{\text{adj}} = 0.002$; $+0.104$, $p_{\text{adj}} < 0.001$); Large is at ceiling and signal-aware truncation does not lift it further ($\Delta = +0.008$). On GraphWalks, all three models are essentially flat under signal-aware truncation ($\Delta \in [-0.041, +0.083]$; none Holm-significant). On MRCR v2, the Δ swing of $+0.286$ between protocols is comparable to BABILong; with larger n and more models this would likely cross the significance threshold. The methodological correction therefore yields an empirical finding, not just a critique: for smaller models, distractor-aware reduction of context to 25% of original significantly helps BABILong score. For the Large model, it is neutral because the full-context baseline is already at ceiling.

5.3 Implications for retrieval and memory systems

A perfect retriever is, by definition, a distractor-aware truncator. Our distractor-aware protocol gives a controlled upper bound on the performance such a system can deliver, conditional on perfect retrieval. Two implications. First, **the upper bound is modest in absolute terms.** Even with perfect distractor removal to 25% of context, Small's BABILong score is 0.279 (vs. 0.196 at full context); the $+0.083$ gain is meaningful and statistically robust but not a transformation. Second, **the upper bound depends on baseline performance.** For Large at ceiling on BABILong

there is no headroom for distractor reduction. For Small, the same operation recovers $+0.083$ (a relative 42% improvement). Retrieval systems probably help small-and-medium models more than they help large models at fixed prompt content.

6 Conclusion

We presented a paired comparison between naive middle-removal and distractor-aware truncation across four long-context benchmarks and three model sizes. On BABILong (all three models, $n = 240$ paired samples per cell) and GraphWalks (all three models, $n = 50$ each), the two protocols give opposite-direction effects at the same nominal truncation: naive truncation produces large, highly significant harm; distractor-aware truncation produces no harm and, for the smaller models on BABILong, statistically significant *improvement* at 25% retention. Signal-loss diagnostics quantify the mechanism: naive truncation deletes the answer in the vast majority of samples; distractor-aware truncation preserves it by construction. The implication is that the often-quoted "shorter context helps" claim cannot be validated or refuted with naive truncation; doing so requires per-benchmark signal identification.

Limitations

Single provider. All three models we evaluate come from a single frontier API-served family. A methodological claim about long-context benchmarks should ideally replicate across providers, and we have not done so. We expect the pattern to replicate because the mechanism we identify — signal preservation rate of $< 1\%$ for BABILong naive at 25% retention — is a property of the truncation procedure and the dataset, not the model; any model that cannot see the missing fact cannot answer the bAbI question. But this is a prediction, not a measurement. Cross-provider runs would also remove a confound: the Large-model temperature fallback (§3.3) is provider-API-specific.

Incomplete model coverage on two benchmarks. Medium and Large runs on MRCR v2 and Oolong did not complete in this snapshot; the experiment was paused at an API credit boundary. The methodological claim is supported in 6/6 (eval $\times$ model) cells on BABILong and GraphWalks, but the strongest individual MRCR result (Δ swing of $+0.286$ between protocols at 25% retention, Small model) currently rests on a single model. The resume procedure is documented in the released code repository.

Hand-crafted signal definitions. §3.2 states an abstract definition (signal = minimal sufficient subset for the gold label) and gives concrete per-benchmark instantiations (regex / dataset field / reachability). Each instantiation is verified against the dataset's own ground truth, but the abstraction does not yield a recipe for new benchmarks; that step still requires per-benchmark engineering. We see this as analogous to the gap between "define a metric" and "define an evaluation harness": the abstraction tells you what you want, the instantiation is the empirical work.

Temperature heterogeneity. The provider API rejects $T = 0$ for the largest model (§3.3), so every Large-model call uses the API default. The paired tests partially absorb the resulting cross-run variance (we contrast *within-sample* truncation levels of the same Large-model run), but this remains a difference between the Large model and the other two that the reader should weigh.

Single sample-selection seed. The within-(eval $\times$ sub-cell) sampling is seeded once; the same sample IDs are used for every model and both protocols. Sensitivity is bounded by the split-half stability analysis in §4.4 (largest $\sigma_A = 0.077$, all headline effects exceed $2\sigma_A$). A formal multi-seed protocol — new sample-ID draws, re-running every cell — would close this off completely and is a natural extension.

Ethics Statement

This paper proposes an evaluation methodology and does not introduce new models, datasets, or system capabilities. All benchmarks used (BABILong, GraphWalks, MRCR v2, Oolong) are public and used under their original licenses; none contain personal identifying information. No human subjects or annotators were involved. The methodological critique itself does not introduce risks beyond standard long-context LLM evaluation; if anything, replacing the naive truncation protocol with the distractor-aware protocol reduces the risk of misinterpreting long-context benchmark scores as direct measurements of context-window effects when they are, in part, measurements of signal loss.

Computational cost: all model evaluations were performed against a public commercial API. The

total cost of the experiments reported in this paper (12 model × protocol × retention configurations on the headline benchmarks) was on the order of US$ low-thousands. Carbon footprint is dominated by inference at the provider's data centres and is not separately measured.

AI assistants (large language models) were used during manuscript drafting and editing for prose revisions. All experimental claims, code, statistical analysis, and per-cell results are the author's. The AI assistant was not used to generate experimental results or evaluate models.

References

Amanda Bertsch, Mirac Suzgun, and 1 others. 2025. Oolong: Evaluating long context reasoning and aggregation capabilities. *arXiv preprint arXiv:2511.02817*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Greg Kamradt. 2023. Needle in a haystack: Pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: The impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

OpenAI. 2025. MRCR dataset. https://huggingface.co/datasets/openai/mrcr.

OpenAI. 2026. GraphWalks dataset. https://huggingface.co/datasets/openai/graphwalks.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

Kiran Vodrahalli, Santiago Ontañón, and 1 others. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

A Disaggregation by task and context-length split

Table 3 disaggregates BABILong by (task × context-length split) for the Large model. The pattern is uniform across qa1, qa2, qa3 and across 0k–32k filler levels: naive truncation produces Δ ranging from -0.05 (qa3@32k, near-zero baseline) to -0.80 (qa2@0k); distractor-aware truncation produces Δ in the range $[-0.05, +0.20]$ across the same cells.

Table 3: BABILong Large model, score by (task × context-length split). Selected rows.

Task	Split	Mode	100%	75%	50%	25%	Δ
qa1	0k	naive	1.000	0.500	0.450	0.400	−0.600
qa1	0k	sig-aw.	1.000	1.000	1.000	1.000	0.000
qa1	2k	naive	1.000	0.650	0.500	0.350	−0.650
qa1	2k	sig-aw.	1.000	1.000	1.000	1.000	0.000
qa1	8k	naive	0.900	0.600	0.500	0.350	−0.550
qa1	8k	sig-aw.	0.950	0.850	0.850	1.000	+0.050
qa1	32k	naive	0.950	0.700	0.550	0.300	−0.650
qa1	32k	sig-aw.	1.000	0.950	0.950	0.950	−0.050
qa2	0k	naive	0.900	0.900	0.500	0.100	−0.800
qa2	0k	sig-aw.	0.900	0.900	0.900	0.900	0.000
qa2	32k	naive	0.350	0.300	0.250	0.250	−0.100
qa2	32k	sig-aw.	0.400	0.350	0.350	0.350	−0.050
qa3	32k	naive	0.500	0.400	0.400	0.450	−0.050
qa3	32k	sig-aw.	0.500	0.450	0.400	0.450	−0.050

The qa1 32k row is the cleanest demonstration: full-context Large gets 0.95–1.00 on this cell; naive truncation to 25% retention drops it to 0.30 (the model cannot see the supporting fact); distractor-aware truncation to 25% retention keeps it at 0.95. Same model, same samples, same nominal truncation; the only thing that changed is what was deleted.

B Full split-half stability table

Table 4: Split-half stability of the $100\% \rightarrow 25\%$ Δ. For each cell, Δ_{full} is the full-sample value reported in Table 1; Δ_A and Δ_B are means over 200 random half-splits of size $n/2$. σ_A is the std. dev. of Δ_A over the 200 splits and bounds the seed-sensitivity. All half-sample estimates are within $\approx 1\sigma$ of the full-sample value.

Eval	Model	Mode	n	Δ_{full}	Δ_A	Δ_B	σ_A
BABILong	Small	naive	240	-0.138	-0.139	-0.136	0.027
BABILong	Small	sig-aw.	240	$+0.083$	$+0.082$	$+0.085$	0.022
BABILong	Medium	naive	240	-0.175	-0.176	-0.174	0.037
BABILong	Medium	sig-aw.	240	$+0.104$	$+0.105$	$+0.104$	0.022
BABILong	Large	naive	240	-0.433	-0.439	-0.428	0.038
BABILong	Large	sig-aw.	240	$+0.008$	$+0.008$	$+0.009$	0.021
GraphWalks	Small	naive	50	-0.352	-0.355	-0.349	0.077
GraphWalks	Small	sig-aw.	50	$+0.083$	$+0.082$	$+0.084$	0.038
GraphWalks	Medium	naive	50	-0.427	-0.428	-0.426	0.071
GraphWalks	Medium	sig-aw.	44	$+0.000$	$+0.000$	$+0.000$	0.000
GraphWalks	Large	naive	50	-0.407	-0.410	-0.404	0.070
GraphWalks	Large	sig-aw.	50	-0.041	-0.040	-0.041	0.026

C Reproducibility

All code, configuration, raw per-sample result records (JSONL with per-row git commit, model version, and config-hash provenance), figures, and analysis scripts will be open-sourced at a public repository upon acceptance. The analysis pipeline re-derives all tables and figures in this paper from the released JSONL records. Sample selection within each (eval $\times$ sub-cell) uses a deterministic, seeded RNG; the seed is recorded in the run config and in every result row's `config_hash`.