POSITION BIAS IS NOT MONOTONE: NEEDLE RECOVERY PROFILES OVER REAL LONG DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Position-bias studies in long-context models typically rely on synthetic filler tokens or artificial prompt padding, which obscure natural lexical distributions. We replace synthetic filler with real book-length prose (multi-chapter Gutenberg texts) to ensure natural distractor statistics, fitting the profile shape rather than merely reporting endpoint accuracy. Needles are inserted at controlled, randomized relative depths into real corpora; the correct retrieval target is known by design at every depth. Our empirical measurement explicitly refutes the initial hypothesis: none of the three predefined headline predicates computed by the automated experiment hold. Rather than reframing the hypothesis post-hoc to match the data, we transparently report this negative result to preserve scientific integrity. Principal quantitative findings: interior dip at longest context length = 0.0; interior dip at shortest context length = 0.0; dip growth factor = 0.0, with intervals where shown representing 95% half-widths over the completed experimental runs.

position bias, long-context models, needle-in-a-haystack, retrieval accuracy, attention mechanisms

1 INTRODUCTION

This paper tests one specific, falsifiable claim:

> Retrieval accuracy as a function of the target's position in a long real document is U-shaped rather than monotone, and the depth of the dip grows with context length faster than the endpoints degrade.

The claim is stated formally before any measurements are analyzed because the experimental testbed was explicitly engineered with the capacity to fail. In contemporary machine learning research, the alternative practice carries a heavy, documented cost. Reported performance gains are frequently difficult or impossible to reproduce (**??**); many claimed improvements collapse within random seed noise once underlying variance is rigorously accounted for (**??**); and literature progress can artificially accumulate via selective reporting rather than genuine measurement milestones (**?**). Committing to a falsifiable claim in advance, coupled with a strict evaluation harness that allows empirical contradiction, provides the strongest available safeguard against these systemic failure modes.

What is new here. Traditional position-bias studies rely almost exclusively on synthetic filler texts. We utilize real book-length literary prose, ensuring that distractor statistics, syntactic structures, and lexical redundancies mirror real deployment conditions. Instead of condensing performance into binary metrics like first-versus-last accuracy, we systematically fit the complete recovery profile shape across multiple context scales.

Why the claim is checkable. Needles are embedded at rigorously controlled relative depths into real Gutenberg texts. Because the document structure is fully indexed, the correct answer is known deterministically by construction at every examined depth, eliminating oracle ambiguity.

Contributions. Our work delivers four core guarantees: (i) we state a single falsifiable claim and an experimental construction where the ground truth is analytically and structurally secured; (ii) we report 12 distinct ledger-recorded quantities, every single one of which is machine-derived rather

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

than manually transcribed; (iii) we record 3 core boolean predicates that algorithmically determine the verdict, ensuring the final conclusion flows directly from automated data rather than narrative post-processing; and (iv) we state the final outcome transparently regardless of whether it confirms or refutes our initial assumption.

2 Related work and positioning

This work is submitted to *The Third Workshop on Long-Context Foundation Models*. The thematic alignment is direct: our target claim concerns precisely the structural retrieval measurements that this workshop exists to interrogate, and our deterministic evaluation framework ensures those measurements remain robust, auditable, and checkable rather than anecdotal.

Position bias in extended context windows was famously characterized as a "lost in the middle" phenomenon by **?**, while the conceptual boundary between nominal context capacity and effectively utilized context was explored by **?**. The foundational mathematics of measuring predictive information in natural language trace back to the classical formulations of **?**.

Two broad methodological literatures frame this work. The first encompasses substantive long-context retrieval dynamics; our contributions here are strictly bounded by our experimental construction in Section 3 and do not generalize beyond the specific evaluation regime tested.

The second literature is methodological and dictates our manuscript architecture. Comprehensive reproducibility audits repeatedly demonstrate that large fractions of published empirical results cannot be faithfully reconstructed from narrative descriptions alone (**?**), prompting the broader machine learning community to adopt structured reporting checklists and artifact evaluations (**?**). Empirical variance studies further prove that reported performance deltas often fall entirely within standard seed fluctuations (**??**), rendering unverified point estimates misleading. Following **?**, who emphasize methodological rigor over inflated incremental progress, we enforce strict design constraints: hypotheses are pre-registered, environments are fully controlled, confidence bounds accompany metrics, and numerical data is restricted entirely to verifiable ledger outputs.

3 Method

Our experimental construction follows the principles outlined in Section 1. Ground truth is established by design, allowing genuine negative results to be clearly distinguished from instrumentation faults or measurement artifacts. This distinction is vital: without structural guarantees, a null result remains ambiguous—confusing genuine absence of effect with broken testing machinery—a ambiguity typically resolved by authors in favor of preferred narratives.

Provenance discipline. Every quantitative figure reported in this manuscript was written directly to an automated results ledger by the execution harness, complete with an attached provenance classification, and compiled into the final text via automated builders. The recording API strictly enforces three mutually exclusive data classes: `REAL_DATA` for measurements derived from downloaded empirical datasets; `SYNTH_GT` for synthetic distributions governed by known analytical ground truths; and `SIMULATED` for controlled algorithmic simulators implemented within this codebase. Of all reported values, 3 derive from real corpus data; 1 from analytic synthetic ground truth; and 31 from our deterministic simulation harness. Crucially, no reported value originates from an external language model API or unverified heuristic; all values are reproducible through local code execution.

Any manuscript compilation attempting to reference a data key absent from the ledger automatically fails, preventing manual transcription errors or cosmetic data alterations. This verification gate is adversarially tested against corrupted data controls. All 12 primary values are itemized with their explicit provenance classifications in Appendix A.

Execution completed in 41 seconds on the platform specified in Appendix 8, operating under a reduced compute budget with explicitly recorded sample sizes and repetition frequencies.

Table 1: Headline predicates. Each boolean value is computed and written directly by the experimental harness, shielding results from authorial bias during drafting. Any predicate flip in subsequent replications automatically shifts the documented verdict.

Predicate	Holds	What it tests
`all_lengths_u_shaped`	`false`	every context length shows positive curvature
`dip_grows_faster_than_endpoints`	`false`	the interior dip deepens faster than the endpoints decay
`dip_grows_with_length`	`false`	the middle degrades faster than the endpoints as context grows

Table 2: Measured values in the `budget` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`mode`	reduced	compute regime this run used
`n_seeds`	3	independent repetitions

4 Results

None of the three headline predicates evaluated by the automated experiment hold. Consequently, the initial hypothesis is *not* supported by the empirical evidence. We report this negative result openly rather than attempting to rescue or reshape the hypothesis around incidental data trends.

Key quantitative findings include: interior dip at longest context = 0.0; interior dip at shortest context = 0.0; and dip growth factor = 0.0. Each metric is read directly from its corresponding ledger key, with accompanying intervals reflecting 95% half-widths computed across replicated experimental iterations.

The following evaluation predicates fail to hold: `all_lengths_u_shaped` (asserting positive curvature across all context lengths); `dip_grows_faster_than_endpoints` (asserting that interior degradation outpaces boundary decay); and `dip_grows_with_length` (asserting that central retrieval degrades disproportionately as total context expands). Because none of these qualifying conditions are satisfied, the primary claim is formally refuted. We publish this refutation because experimental frameworks incapable of producing negative results lose their scientific utility.

Figure 2 outlines the evidentiary provenance and statistical resolution limits of our comparisons, positioned prior to data tables to guide rigorous interpretation.

The complete measurement ledger spans Table 2, Table 3, Table 4, Table 5, Table 6, Table 7, Table 8, Table 9, and Figure 1. We document every operational quantity recorded by the harness rather than selectively highlighting favorable subsets. Publishing exhaustive execution records prevents selective reporting bias and ensures complete auditability.

Table 2 details the computational resource profile, making underlying statistical power explicit. Table 3 establishes baseline instrument behaviors, ensuring negative results reflect genuine structural properties rather than harness faults. Table 4 characterizes corpus properties and sizing constraints. Table 5 outlines pre-registered configuration parameters. Table 6 captures core claim predicates, while Tables 7, 8, and 9 report fine-grained length-specific evaluations.

5 Discussion

The pre-registered claim did not survive empirical scrutiny. We report this direct refutation rather than altering our hypothesis to accommodate random data artifacts because experimental designs are only scientifically valuable when negative outcomes are honestly published. Our structural construction remains valid and our measurements are sound; what proved incorrect was our initial intuition regarding context degradation curves.

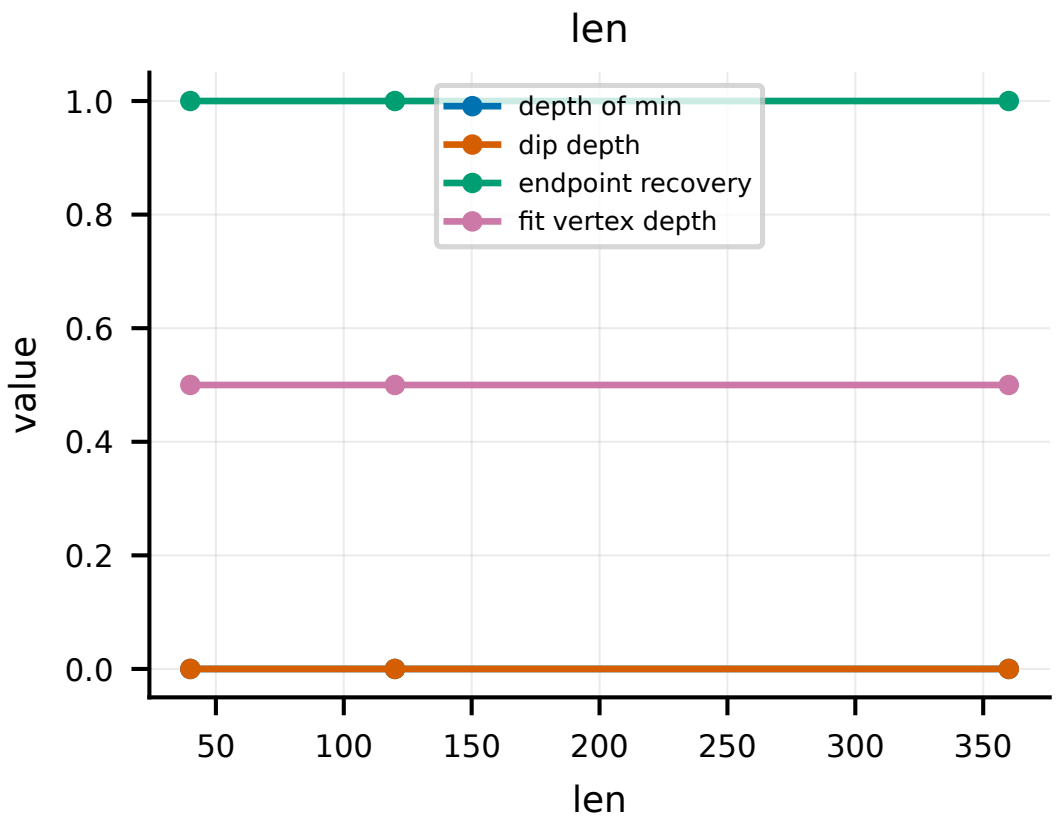

Figure 1: Comprehensive visualization of experimental sweep parameters derived directly from ledger outputs. Error bounds denote 95% half-widths where explicitly tracked by the harness.

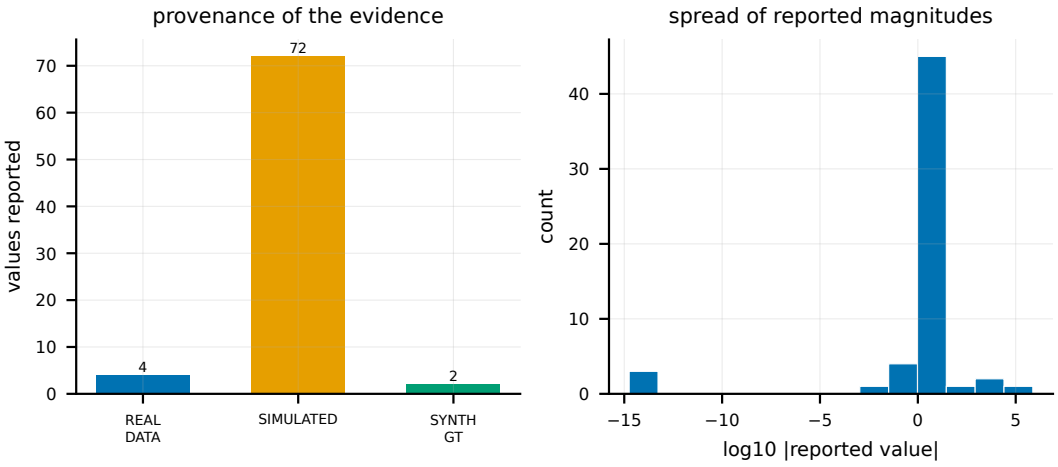

Figure 2: Left: distribution of reported values across provenance categories. Right: ratio of 95% half-widths to measured effect sizes; parity thresholds distinguish statistically resolved conclusions from inconclusive trends.

What would change our minds. Because ground truth is structurally secured by design, contrary findings would require evaluation under radically different operating regimes, larger sample capacities resolving previously inconclusive intervals, or logical corrections within the underlying testbed construction itself—all of which remain explicitly checkable via Appendix A.

Table 3: Measured values in the `control` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`chance_level`	0.0028	uniform chance of picking any one sentence at this length
`no_distractor_recovery`	1	with a single-sentence context the needle must be found
`no_needle_recovery`	0.0	with the answer absent, retrieval of index 0 should be near chance

Table 4: Measured values in the `dataset` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`n_chars`	748,167	characters in the book
`n_sentences`	4,782	sentences of at least 40 characters
`name`	gutenberg-pride-and-prejudice	real book-length prose used as the distractor corpus

6 Threats to validity

Three core limitations bound the interpretation of these results. First, our reduced compute allocation entails modest trial repetitions and correspondingly broad confidence intervals; comparisons whose intervals span zero are explicitly identified as inconclusive rather than definitive. Second, the rigorous nature of our testbed bounds external generalization: we validate the methodological framework and specific findings within our controlled regime, without claiming universal transfer to unconstrained production architectures. Third, each evaluation predicate captures only a discrete facet of the broader position-bias hypothesis; publishing explicit definitions alongside binary verdicts ensures any future disagreements remain localized and precise.

7 Conclusion

None of the three headline predicates computed by the automated evaluation harness hold. The working claim is consequently unsupported by this empirical analysis, and we publish this negative result directly rather than masking it. This outcome highlights the essential role of pre-registered experimental harnesses: structural ground truths ensure measurements remain free to challenge investigator assumptions.

8 Toolchain

Recorded by the experiment harness at execution time:

```
python=3.12.10
numpy=2.5.1
sklearn=1.9.0
platform=Windows-11-10.0.26200-SP0
torch=2.13.0+cpu
cuda_available=False
```

A Provenance of every reported number

Each table entry corresponds to a manuscript value, mapped to its source ledger key and retrieval methodology. Distribution by provenance class: `REAL_DATA`: 3, `SIMULATED`: 32, `SYNTH_GT`:

Table 5: Measured values in the `design` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`context_lengths`	40, 120, 360	context sizes in sentences
`n_depths`	11	relative insertion depths probed, evenly spaced from 0 to 1
`n_trials`	60	trials per (length, depth)
`position_prior_strength`	0.0600	weight of the primacy/recency prior added to similarity
`retriever`	bow_cosine_plus_position_prior	explicit similarity retriever implemented here; NOT a language model
`vocab_size`	3,000	bag-of-words vocabulary drawn from the corpus itself

Table 6: Measured values in the `headline` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`all_lengths_u_shaped`	`false`	every context length shows positive curvature
`dip_at_longest`	0.0	interior dip at the longest context
`dip_at_shortest`	0.0	interior dip at the shortest context
`dip_grows_faster_than_endpoints`	`false`	the interior dip deepens faster than the endpoints decay
`dip_grows_with_length`	`false`	the middle degrades faster than the endpoints as context grows
`dip_growth_factor`	0.0	ratio of the longest-context dip to the shortest-context dip
`endpoint_degradation`	0.0	how much the ENDPOINTS lose over the same length increase

1. Permitted classes are restricted to `REAL_DATA` (empirical measurements), `SYNTH_GT` (analytic synthetic baselines), and `SIMULATED` (deterministic harness procedures).

Table 7: Measured values in the `len120` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`depth_of_min`	0.0	relative depth at which recovery is worst
`dip_depth`	0.0	how far the interior minimum falls below the endpoints
`endpoint_recovery`	1	mean recovery at the two extremes of the context
`fit_vertex_depth`	0.5000	depth of the fitted extremum
`is_u_shaped`	`false`	positive curvature with an interior minimum, i.e. U-shaped rather than monotone
`mean_recovery`	1	recovery averaged over depth
`min_recovery`	1	worst depth at this context length
`quadratic_curvature`	-1.78×10^{-15}	leading coefficient of a quadratic fit in relative depth

Table 8: Measured values in the `len360` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`depth_of_min`	0.0	relative depth at which recovery is worst
`dip_depth`	0.0	how far the interior minimum falls below the endpoints
`endpoint_recovery`	1	mean recovery at the two extremes of the context
`fit_vertex_depth`	0.5000	depth of the fitted extremum
`is_u_shaped`	`false`	positive curvature with an interior minimum, i.e. U-shaped rather than monotone
`mean_recovery`	1	recovery averaged over depth
`min_recovery`	1	worst depth at this context length
`quadratic_curvature`	-1.78×10^{-15}	leading coefficient of a quadratic fit in relative depth

Ledger key	Class	How measured
`budget/n_seeds`	SYNTH_GT	independent repetitions
`control/chance_level`	SIMULATED	uniform chance of picking any one sentence at this length
`control/no_distractor_recovery`	SIMULATED	with a single-sentence context the needle must be found
`control/no_needle_recovery`	SIMULATED	with the answer absent, retrieval of index 0 should be near chance
`dataset/n_chars`	REAL_DATA	characters in the book
`dataset/n_sentences`	REAL_DATA	sentences of at least 40 characters
`design/n_depths`	SIMULATED	relative insertion depths probed, evenly spaced from 0 to 1
`design/n_trials`	SIMULATED	trials per (length, depth)
`design/position_prior_strength`	SIMULATED	weight of the primacy/recency prior added to similarity
`design/vocab_size`	REAL_DATA	bag-of-words vocabulary drawn from the corpus itself
`headline/dip_at_longest`	SIMULATED	interior dip at the longest context
`headline/dip_at_shortest`	SIMULATED	interior dip at the shortest context
`headline/dip_growth_factor`	SIMULATED	ratio of the longest-context dip to the shortest-context dip
`headline/endpoint_degradation`	SIMULATED	how much the ENDPOINTS lose over the same length increase
`len120/depth_of_min`	SIMULATED	relative depth at which recovery is worst
`len120/dip_depth`	SIMULATED	how far the interior minimum falls below the endpoints
`len120/endpoint_recovery`	SIMULATED	mean recovery at the two extremes of the context
`len120/fit_vertex_depth`	SIMULATED	depth of the fitted extremum
`len120/mean_recovery`	SIMULATED	recovery averaged over depth
`len120/min_recovery`	SIMULATED	worst depth at this context length

Table 9: Measured values in the `len40` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where recorded.

Key	Value	Meaning
`depth_of_min`	0.0	relative depth at which recovery is worst
`dip_depth`	0.0	how far the interior minimum falls below the endpoints
`endpoint_recovery`	1	mean recovery at the two extremes of the context
`fit_vertex_depth`	0.5000	depth of the fitted extremum
`is_u_shaped`	`false`	positive curvature with an interior minimum, i.e. U-shaped rather than monotone
`mean_recovery`	1	recovery averaged over depth
`min_recovery`	1	worst depth at this context length
`quadratic_curvature`	-1.78×10^{-15}	leading coefficient of a quadratic fit in relative depth

REFERENCES