Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Long context LLM robustness: Does evidence position matter?

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-11 · long context llm robustness · llm evidence position bias · lost in the middle · long context evaluation · context engineering · citation evaluation · document order · paired bootstrap · source provenance

Original article: https://reliablecontext.com/articles/long-context-llm-evidence-position-robustness



In this article

1. Executive Summary
2. Introduction and Background
3. Evidence position and claim-level verdicts
4. Freeze an eligible, versioned evidence bundle
5. Run a paired position-swap experiment
6. What existing benchmarks establish
7. A source-versioned paired scorecard
8. Analysis of key segments and failures
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

Long context LLM robustness depends on the task and the position of decisive evidence. A context window states what can be supplied; it does not establish that a source-supported answer remains correct when the same passage moves. Liu et al. (2024) documented position sensitivity in multi-document question answering and synthetic key-value retrieval, including degradation when relevant information appeared in the middle. (aclanthology.org) That finding is not a universal verdict on subsequent models: Google Research reports no middle-position effect for **Gemini 2.5 Flash** in its tested simple factoid retrieval setting. (research.google) Context engineers should therefore measure the exact claim, evidence bundle, model version, and output requirements their application uses.

Benchmark results show why retrieval alone is insufficient. **Loong** requires every supplied document to contribute to the answer; **NoLiMa** deliberately minimizes lexical overlap between questions and relevant passages. (aclanthology.org) (proceedings.mlr.press) In the published NoLiMa evaluation, **11 of 13 models** fell below half their short-context baseline at **32K tokens**. (proceedings.mlr.press) **LongBench v2** contains **503 multiple-choice questions**, but selecting an answer does not demonstrate that its supporting source can be recovered. (longbench2.github.io) **ALCE**, Automatic LLMs' Citation Evaluation, explicitly evaluates correctness and citation quality separately. (collaborate.princeton.edu) These results describe particular experimental conditions, rather than a current league table for organizational workloads.

This report proposes a **paired position-swap protocol**: freeze permission-eligible source versions, the claim, total input tokens, prompt, model, and decoding configuration; place decisive evidence early, centrally, and late; then judge every output against the same evidence verdict. Treat answer correctness, recoverable citation, citation completeness, abstention, and contradiction handling as separate outcomes. The provenance requirement follows the distinction between a particular document version and a mutable latest-version address; the citation requirement follows published attribution evaluation. (www.w3.org) (arxiv.org) A correct answer with an unresolved source locator fails the joint supported-answer criterion.

The worked scorecard is explicitly **hypothetical**, not a model measurement. Its constructed **200 paired cases** yield supported-answer pass rates of **88%, 76%, and 87%** at early, middle, and late positions. Middle minus the mean edge rate is **−11.5 percentage points**; a case-level percentile bootstrap produces a **95% interval from −16.50 to −6.50 points**. Those figures illustrate the calculation, not expected performance. Resampling preserves each case's linked positions, consistent with paired-bootstrap practice. (docs.scipy.org) The decision is whether correctness and recoverable support meet predeclared tolerances across relevant positions and length bands, with uncertainty and raw transitions visible, rather than whether a model accepts a larger input.

Introduction and Background

Context engineers often assemble an answer-bearing passage together with related documents, previous observations, and genuine distractors. The operational question is narrow: **does the same eligible evidence support the same defensible answer after its position changes?** The report examines observable outputs. It does not infer a model's internal use of evidence from an answer alone. The separation matters because published citation evaluation measures answer correctness and citation quality independently. (collaborate.princeton.edu)

Evidence position is the location of a decisive source span within the actual serialized input. **Decisive evidence** is the passage, or set of passages, that a reviewer needs to accept the claim under the case's stated rules. **Position robustness** means that required output properties remain within a declared tolerance across controlled placements. These are operational definitions for the proposed test, not claims that an accepted standard already specifies this protocol. The empirical motivation comes from Liu et al.'s controlled position experiments and later application-centered evaluation. (aclanthology.org) (princeton-nlp.github.io)

The relevant organizational requirement is a value whose supporting source survives review. Reliable Context's public contract describes a verbatim, cited, dated source behind every value; this is a record requirement, not evidence of long-context model performance. Evaluating the source-supported output therefore requires more than asking whether a response resembles a reference answer. A source locator must lead to the observed version and the passage must support the asserted value, with qualifiers intact. W3C's provenance model distinguishes document versions and relationships among usage, generation, and derivation. (www.w3.org)

This report synthesizes research and documentation accessed on **October 10, 2026** and specifies a reusable acceptance test. It reports no proprietary model runs, organizational benchmark results, or measured Reliable Context accuracy. Studies using synthetic needles, all-relevant documents, ranking, multiple-choice questions, and generated citations answer different questions. Their results should guide test construction rather than be combined into a single performance score. HELMET's authors report low correlations among application categories and poor downstream prediction from simple needle tasks. (princeton-nlp.github.io)

Evidence position and claim-level verdicts

A position test needs a stable **claim-level verdict** before any model output is judged. For a source-backed value, the reference should state what the eligible evidence supports, what it leaves unresolved, and which qualifications must appear in an acceptable answer. The verdict may be supported, contradicted,

indeterminate, or unavailable to the intended reader. These are proposed evaluation labels. Their role is to prevent answer matching from silently substituting for evidence assessment; ALCE provides the research precedent for evaluating correctness and citation quality separately. (arxiv.org)

The distinction produces several independent outcomes:

- **Answer correctness:** the asserted value and material qualifications agree with the adjudicated reference.
- **Locator validity:** the returned document identifier and span resolve within the frozen source version.
- **Citation support:** the recovered passage supports the specific assertion rather than merely discussing its subject.
- **Citation completeness:** every decisive part of a compound claim has recoverable support.
- **Appropriate abstention:** the response declares the required unknown when eligible evidence cannot establish the value.
- **Contradiction handling:** the response preserves conflicting evidence and follows the case's stated resolution rule.

This rubric extends existing citation metrics with organizational requirements; it is not presented as an established benchmark. ALCE's original evaluation found incomplete citation support even among its strongest evaluated systems, demonstrating why citation quality deserves an independent outcome. (collaborate.princeton.edu) NIST also advises reviewing and verifying sources and citations during measurement and monitoring. (nvlpubs.nist.gov)

Recoverability is narrower than semantic support. Cohere's citation documentation describes indices into generated response text and referenced sources, while Claude's documentation describes valid pointers into supplied documents. Neither property alone proves that every assertion follows from the cited passage. (docs.cohere.com) (platform.claude.com) A pointer may resolve perfectly to a sentence stating a range when the answer gives an unsupported exact value. Conversely, an answer may be correct while identifying a document that cannot be recovered. The joint pass indicator requires both correctness and complete recoverable support.

An acceptable citation should therefore retain the source identity, observed version, locator, and decisive excerpt. W3C PROV-DM provides concepts for describing derivation and the provenance of provenance records; the proposed rubric adds an explicit entailment judgment. (www.w3.org) That judgment concerns the supported assertion, including units, scope, effective date, and conditions. Exact quotation does not excuse an answer that drops the source's qualification.

Answer consistency across evidence positions should be assessed semantically, not through identical wording. Equivalent paraphrases can pass; repeated unsupported values cannot. Report joint supported-answer consistency alongside correctness consistency and citation consistency. A model that changes its answer appropriately when evidence changes may be useful, but this experiment holds evidence content fixed. The target comparison is therefore whether position alone changes the observed verdict, subject to the serialization controls described below. This framing avoids equating stability with truth. (collaborate.princeton.edu)

Freeze an eligible, versioned evidence bundle

The experimental unit is a **case**, consisting of a claim, a fixed source bundle, and an adjudicated evidence verdict. Freeze what the model and reviewer can inspect, rather than retaining only live web addresses. W3C notes that an identifier for a latest document version can refer to a changing resource; a version-specific identifier answers a different provenance question. (www.w3.org) A subsequent page correction must not silently change the reference for a position experiment already completed.

The source manifest should retain these fields:

- **Source identity:** a stable identifier for the document and its supplying origin.
- **Observed version:** the retained bytes or text, their digest, and the version identifier when available.
- **Observation time:** when the exact source was obtained, separate from publication or effective date.
- **Eligibility:** the relevant source permission and reader-clearance decision at the test cutoff.
- **Decisive spans:** the exact quoted evidence, its source-relative locator, and any alternatives that also support the verdict.
- **Reference verdict:** the expected value, qualifiers, conflicts, permitted citations, and abstention conditions.

These are proposed manifest requirements. They operationalize versioned provenance and reproducibility without prescribing how a production system selects values or resolves identity. W3C's model links activities, usage, generation, and derivation; it also permits provenance descriptions themselves to have provenance. (www.w3.org) The case manifest and adjudication history should consequently remain inspectable alongside the final label.

Build **genuine distractor cases** from material that might realistically enter the same assembly, not merely repeated neutral filler. Similar names, neighboring periods, superseded observations, and plausible but ineligible evidence expose different output risks. Keep their eligibility labels outside the position treatment. The classic Lost in the Middle setup used NaturalQuestions queries with Wikipedia distractors and a designated answer-bearing passage, a useful controlled design whose corpus differs from an organization's documents. (cs.stanford.edu)

Include **all-relevant multi-document cases** separately. In Loong, every document is relevant to the final answer, and tasks include locating, comparison, clustering, and chains of reasoning. (aclanthology.org) Such cases cannot be reduced to moving one magic sentence through dispensable filler. A claim may require a premise, a later qualification, and a conflicting observation. Record all decisive spans, including evidence that changes which value is supportable.

For these cases, define the treatment precisely. Moving the whole evidence-bearing document tests document-order sensitivity. Moving a passage inside a document tests intradocument position and may disrupt discourse. Moving several premises together differs from separating them. A multi-hop study by Zhang et al. (2026) found an absolute-position weakest-link pattern in its tested configuration; that result makes the least visible decisive span a useful additional diagnostic, rather than establishing a universal law about interspan distance. (aclanthology.org)

Finally, freeze source availability. If a cited source becomes externally inaccessible after capture, assess recoverability against the retained observed version specified by the protocol. If the evaluator never possessed the eligible evidence, declare the case unavailable rather than infer its content. This maintains a reproducible provenance question instead of mixing source-access failures with model-position effects. (www.w3.org)

Run a paired position-swap experiment

The treatment should alter **where evidence appears**, while holding the eligible information and intended input budget fixed. Start from a serialized baseline, then construct early, middle, and late variants from the same case. Record the decisive span's token coordinates in the final request. Use its midpoint divided by the full serialized input length as the **location percentile**, while retaining start and end coordinates for spans that cross a band boundary. This is the report's proposed positional measure, not a published benchmark convention. (aclanthology.org)

Execute the following sequence:

- **Freeze the request contract:** retain system instructions, question, response format, model identifier, and decoding settings.
- **Serialize every variant:** keep source labels attached to their documents and preserve original source-relative citation locators.
- **Measure final tokens:** use the selected model's counting method on the complete input, including wrappers and the question.
- **Check eligibility and inclusion:** confirm the same source set and decisive spans appear in every variant, without silent truncation.
- **Randomize run order:** balance placement conditions over time instead of always executing early before middle before late.
- **Retain raw outputs:** keep the actual request and response, returned citation objects, usage accounting, and execution status.

Token accounting requires care. Google's documentation distinguishes input-only counting from response usage fields; input length and output or thinking usage must not be conflated. (ai.google.dev) Document permutation may change tokenization at boundaries even when document text is unchanged. Use fixed separators, verify measured totals, and invalidate variants that fail the predeclared token-equality rule. Do not introduce compensating filler casually: that changes the evidence environment. If exact equality is infeasible, report a separate near-fixed-budget experiment rather than describing it as exact control.

Keep the **query position** fixed. Google's guidance recommends placing the query after the context, and Claude's guidance recommends extracting relevant quotations before the task. (ai.google.dev) (platform.claude.com) These are useful candidate prompt choices, but changing them during the position sweep introduces another treatment. Choose the prompt on a development set, freeze it, and evaluate on held-out cases. Test prompt changes in a separate factorial comparison if their interaction with position is the decision of interest.

Position is different from truncation. If a late span was omitted from the request, the model did not receive a late-position treatment. That run diagnoses context assembly or request limits. Explicitly account for tools, hidden service transformations when observable, output reservation, and retrieved material. Avoid declaring a placement effect when the compared inputs differ in which decisive evidence they contain. Microsoft's retrieval FAQ also distinguishes arbitrary-key retrieval from simpler fixed-needle tasks, reinforcing the need to identify the actual workload. (www.microsoft.com)

Use paired case identifiers throughout. For stochastic generation, repeat each placement under a declared replication schedule and distinguish variation across cases from variation across runs. A fixed random seed, when supported, is useful metadata but does not establish that all service behavior is deterministic. The primary analysis should preserve each case's linked observations; paired bootstrap resampling applies the same sampled indices to the linked arrays. (docs.scipy.org)

The protocol estimates an **output sensitivity to a specified permutation**. Changing document order can also change discourse coherence, neighborhood cues, or relationships among premises. It cannot establish which internal mechanism caused an error. Report document-level and span-level treatments separately, with valid construction checks and semantic review of altered discourse. FarRelevant's authors explicitly acknowledge that randomly assembled passages do not resemble natural documents, a limitation relevant to synthetic position permutations. (aclanthology.org)

What existing benchmarks establish

Published benchmarks provide a menu of workload properties, not interchangeable estimates of organizational reliability. The central distinctions are retrieval versus synthesis, single evidence versus multiple premises, lexical matching versus semantic association, and answer selection versus cited generation. HELMET finds synthetic needle results inadequate predictors of downstream performance, while Michelangelo introduces diagnostic tasks that require identifying latent structure across long contexts. (princeton-nlp.github.io) (deepmind.google)

Table 1 compares the evidence each source supplies. The record-contract row identifies an output requirement rather than a model benchmark or a measured competitor result.

Source or evaluation layer	What it tests or requires	What it does not establish
Lost in the Middle, Liu et al.	Position-controlled multi-document question answering and key-value retrieval. (aclanthology.org)	A current universal ranking across models or all organizational tasks.
Loong, Wang et al.	Multi-document tasks in which every document contributes to the answer. (aclanthology.org)	Whether a single retrieved needle predicts all-premise reasoning.
RULER, Hsieh et al.	Synthetic retrieval, tracing, and aggregation tasks. (arxiv.org)	Replacement for realistic application tasks, which its authors explicitly caution against. (github.com)
NoLiMa, Modarressi et al.	Needle retrieval with minimal lexical overlap between question and evidence. (proceedings.mlr.press)	Full source-version recovery or contradiction-aware organizational decisions.

Source or evaluation layer	What it tests or requires	What it does not establish
LongBench v2, Bai et al.	Challenging multiple-choice tasks across long inputs. (longbench2.github.io)	A complete cited-generation acceptance criterion.
ALCE, Gao et al.	Separate fluency, correctness, and citation-quality evaluation. (collaborate.princeton.edu)	The proposed fixed-budget, paired position-swap test of retained organizational sources.
Reliable Context, documented system-of-record contract	A verbatim, cited, dated source behind every value and reproducible past record states.	Published model-position scores or comparative model accuracy; none are asserted here.

The table separates a **documented requirement** from measured research results. A record contract can define what must be defensible, but model selection still requires an application test. Multiple-choice accuracy cannot certify source locators, and citation metrics cannot automatically certify the eligibility or version of the source being cited. The proposed protocol combines these concerns as observable acceptance properties. (arxiv.org) (www.w3.org)

Ranking is another distinct setting. Boytsov et al. (2025) found early-position relevance bias in commonly used ranking corpora and designed a farther-evidence diagnostic. Across the evaluated standard datasets, no long-document model exceeded its same-model short-input FirstP baseline by more than **5% on average**. (aclanthology.org) That is a ranking result under the paper's baseline definition, not a question-answering pass rate. Benchmark position artifacts can obscure the capability a context engineer needs to evaluate.

Reasoning benchmarks add complementary stressors. **BABILong** scatters necessary facts through natural text and includes **20 reasoning tasks**, while **InfiniteBench** covers mathematics, code, and novels in English and Chinese, with automatically generated tasks that permit length scaling. (axi.lims.ac.uk) (chen-yingfa.github.io) These workloads help distinguish retrieving a passage from integrating several premises. Google's long-context guidance similarly cautions that multiple-needle retrieval does not preserve the accuracy of a basic single-needle setup. (ai.google.dev)

Results also depend on evaluation procedure. LongBench v2 reports a time-limited human comparison with an abstention option, and HELMET's authors report no clear winner across all application categories. (longbench2.github.io) (huggingface.co) Neither observation supports choosing a model from a single headline percentage. Use the benchmark most similar to the intended workload, then test the actual source bundle and output contract.

A source-versioned paired scorecard

The scorecard should keep input position separate from source position. The former is a treatment variable; the latter identifies supporting text in a retained document. A permutation changes the input percentile while preserving source-relative locators. Citation interfaces also differ: Claude documents character ranges for plain text, page ranges for PDFs, and block ranges for custom content. (platform.claude.com) Preserve those native locators rather than forcing every source into an unqualified sentence number.

Table 2 illustrates the proposed case-level schema using a **Hypothetical Example**. The source, model, and values below are invented for demonstration; the repeated claim concerns a supported threshold of seven units. A production scorecard would retain full serialized requests and complete reviewer records outside this compact display.

Case ID and placement	Source version and observation date	Span midpoint percentile; input tokens	Answer verdict	Exact cited span and locator	Reviewer label	Model and prompt version
H-001 early	Memo A, retained v3; 2026-10-01	5%; 16,000	Correct: seven units	“The approved threshold is seven units.” Section 2, sentence 4	Correct answer; citation complete	Illustrative model M; prompt P1
H-001 middle	Same retained source	50%; 16,000	Correct: seven units	Memo A cited, sentence unresolved	Correct answer; locator fails	Same M and P1
H-001 late	Same retained source	95%; 16,000	Incorrect: nine units	“The approved threshold is seven units.” Section 2, sentence 4	Locator resolves; assertion unsupported	Same M and P1

The middle answer is correct but fails the **joint supported-answer** criterion. The late output demonstrates the reverse distinction: finding a valid source sentence does not make the asserted value correct. The schema follows versioned-provenance requirements and the separation of correctness from citation quality. (www.w3.org) (collaborate.princeton.edu) A real implementation should retain every decisive span when a claim needs several sources, plus the evidence verdict and independent reviewer judgments.

A compact table is a review surface, not the full experimental record. Retain citation objects and exact excerpts before normalization, associate each reviewer judgment with its rubric version, and preserve both initial labels and adjudicated verdicts. Provenance descriptions can themselves have provenance, supporting an inspectable history of the evaluation record. (www.w3.org) This matters when a later reviewer revises an evidence judgment: distinguish a changed reference label from a changed model output.

Analysis of key segments and failures

Report position effects **within relevant segments** before aggregating them. An overall average may mix short single-span cases with long multi-document contradictions, or answerable claims with cases that require abstention. HELMET's category-dependent results provide a reason to keep application categories visible rather than treat them as exchangeable. (huggingface.co) The proposed segments should be established before reviewing final outcomes.

Useful segments and failure labels include:

- **Length band:** the actual serialized input tokens, including request wrappers, rather than a document word-count proxy.

- **Evidence cardinality:** a single decisive span, several complementary spans, or competing observations.
- **Task structure:** lookup, comparison, aggregation, or a chain of dependent premises.
- **Surface overlap:** explicit matching terms versus semantic association between question and evidence.
- **Answer availability:** supportable value, explicit contradiction, unresolved evidence, or permission-limited evidence.
- **Citation failure:** missing source, wrong version, invalid locator, incomplete support, or a recovered passage that does not support the claim.

These are analysis choices, supported by the workload distinctions in Loong, NoLiMa, and ALCE. (aclanthology.org) (proceedings.mlr.press) (collaborate.princeton.edu) They should not become a taxonomy so granular that every estimate rests on a handful of cases. Publish the denominator and the distribution of positions within each segment. Keep incidentally changing source length separate from the planned treatment.

Reviewer labels need their own quality assessment. Use independent initial labeling, preserve disagreement before adjudication, and report how disputes were resolved. **Cohen's kappa** measures agreement between two annotators beyond expected chance agreement; it complements raw agreement rather than replacing the label-confusion table. (scikit-learn.org) A kappa of zero indicates no agreement beyond the measure's chance expectation. (scikit-learn.org) Do not declare an arbitrary universal acceptance threshold without considering class distribution and the consequences of the disputed labels.

For more annotators or incomplete ratings, **Krippendorff's alpha** offers an applicable alternative under a specified measurement level. (www.asc.upenn.edu) Report the rating design, which labels were missing, and whether the coefficient was calculated before adjudication. Agreement after reviewers resolve every disagreement measures a different process from agreement among independent first judgments.

If a model judge assists scoring, validate it against human ratings for the particular task. Google Cloud's evaluation guidance recommends human ratings as ground truth for that comparison, and its groundedness rubric asks whether all aspects of a response are attributable to context. (docs.cloud.google.com) A groundedness score is helpful, but it is not a substitute for checking a retained version and exact citation locator. Preserve human review of decisive disagreements and position-dependent reversals.

Finally, examine transitions. A middle-position failure following an early-position pass is a different operational pattern from a failure at every position. Abstention may avoid an unsupported assertion, yet remain incorrect on a case with adequate eligible evidence. Conversely, abstention can be the correct output when the reference verdict is unavailable or unresolved. The report's rubric therefore retains abstentions as labeled outcomes, with separate denominators, rather than silently dropping them from answer accuracy. This extends the independent-outcome approach of citation evaluation. (arxiv.org)

Data Analysis and Evidence

The research supports **task-specific sensitivity**, not a universal failure rate. In Liu et al.'s appendix, the **Llama-2 13B** base model showed a **20-point** best-to-worst accuracy disparity in a **20-document** setting; the reported degradation after additional tuning was **10 points**. (aclanthology.org) In the published NoLiMa study,

11 of 13 evaluated models fell below half their short-context baseline at **32K tokens**; the reported GPT-4o result declined from **99.3% to 69.7%**. (proceedings.mlr.press) These are historical study conditions, not estimates for an untested current model.

Other figures describe benchmark construction. Loong selected **1,600 instances** with context intervals spanning **10K to 250K tokens**. (aclanthology.org) The revised RULER study evaluated **17 models** across **13 representative tasks**. (arxiv.org) LongBench v2 contains **503 questions**, and its project specifies contexts in **words**, not tokens. (longbench2.github.io) Converting these lengths into a common axis without the model's tokenizer would create an unsupported comparison. The available studies establish why the proposed test needs explicit length, task, and source labels.

For case i, let E_i, M_i, and L_i indicate joint passes at early, middle, and late positions. Define the edge score as $(E_i + L_i)/2$, and the paired effect as the mean of $M_i - (E_i + L_i)/2$. Also report middle minus early and middle minus late separately. The equal weighting is a proposed estimand: it does not imply that operational evidence occurs equally often at both edges. Paired resampling must retain the linked observations for each selected case. (docs.scipy.org)

Table 3 supplies all joint transition counts for a second **Hypothetical Example**. It represents a constructed set of **200 cases**, not observed model performance. A value of one means joint pass; zero means joint failure.

Early E	Middle M	Late L	Number of constructed cases
0	0	0	18
0	1	0	2
0	1	1	4
1	0	0	6
1	0	1	24
1	1	1	146
Total			**200**

The two absent joint patterns have zero cases. These counts give **176 early passes, 152 middle passes, and 174 late passes**, or **88%, 76%, and 87%**. The paired middle-minus-early difference is **−12 points**, middle minus late is **−11 points**, and middle minus mean edge is **−11.5 points**. Early-to-middle discordances are **30 losses and six gains**; late-to-middle discordances are **24 losses and two gains**. These are arithmetic consequences of the table, not literature estimates. Exact McNemar testing uses paired discordances in a contingency table, whereas the three-position mean-edge estimand calls for its own uncertainty analysis. (www.statsmodels.org)

The constructed outputs additionally designate **three early, seven middle, and four late abstentions**, all drawn from joint failures on answerable cases; these counts are hypothetical annotations, not inferred from binary transitions. Resampling **200 whole cases with replacement**, **20,000 times**, using NumPy's default random generator with seed **20261010**, gives a percentile **95% interval of −16.50 to −6.50 percentage**

points for the mean-edge difference. Expand the table in its displayed row order, compute each case's difference, and sample linked case indices to reproduce this calculation. The method preserves the pairing described in bootstrap documentation. (docs.scipy.org)

For a real study, choose the uncertainty method before seeing results and disclose clustered cases, repetitions, exclusions, and sparse segments. SciPy cautions that bias-corrected and accelerated bootstrap endpoints can be undefined for degenerate distributions. (docs.scipy.org) A narrow interval for this illustrative sample would not validate the case generator, label quality, or transfer to another corpus.

Implications and Future Directions

Model selection should begin with an **output contract and tolerance**, followed by a corpus-specific evaluation. Require an absolute supported-answer pass rate at each relevant position and a maximum acceptable paired loss. A small positional gap is insufficient if every placement performs poorly. A strong average is insufficient if a relevant middle segment falls below the acceptance floor. NIST recommends empirically validated evaluation of model-capability claims, while HELMET's research cautions against relying on synthetic retrieval as a proxy for application performance. (nvlpubs.nist.gov) (princeton-nlp.github.io)

Before approving a model or an assembly change, retain these release criteria:

- **Absolute adequacy:** each relevant position and length band meets its declared supported-answer target.
- **Position tolerance:** uncertainty around paired loss is compatible with the maximum acceptable degradation.
- **Citation integrity:** correct answers with unresolved, incomplete, or unsupported citations remain visible as failures.
- **Unknown handling:** appropriate abstention is scored separately from unnecessary abstention on supportable cases.
- **Reference quality:** reviewer disagreements, adjudication rules, and label agreement are documented.
- **Reproducibility:** source versions, prompts, model identifiers, raw outputs, and case exclusions can be inspected.

These criteria are proposed acceptance rules, not vendor performance guarantees. Their citation requirement is compatible with explicit source review and the distinction between response pointers and semantic support. (nvlpubs.nist.gov) (platform.claude.com) A change to prompt structure, document ordering, model version, or eligible source set should trigger the corresponding controlled re-evaluation rather than inherit an earlier pass automatically.

Evidence placement in prompts can be an intervention, but it should be measured as such. Anthropic's historical long-context experiment explicitly varied answer-passage location and prompting strategies; Google's guidance recommends query placement after context. (www.anthropic.com) (ai.google.dev) A favorable edge placement may help one workload while leaving multi-span reasoning or citation recovery unresolved. Do not infer that moving known evidence eliminates the need to assess its support.

Future work should increase ecological validity while preserving experimental control: coherent distractors, multilingual source bundles, distributed premises, source corrections, and access-limited evidence. BABILong's reasoning tasks and Michelangelo's latent-structure evaluations offer useful design directions, while InfiniteBench illustrates task diversity and scalable generation. (axi.lims.ac.uk) (deepmind.google) (chen-yingfa.github.io) Independent label audits and held-out organizational cases remain necessary. The result should identify the conditions under which a cited value can be relied upon, with observable failure modes and declared limits, rather than claim that context length itself guarantees knowledge use.

Conclusion

Evidence position can matter, but the decision-relevant question is conditional: **does this model, prompt, and frozen eligible context preserve a correct, recoverably supported claim when its decisive evidence moves?** The literature motivates that test without establishing that every long-context model exhibits the same curve. Liu et al. examined position effects in specific retrieval and question-answering tasks; later simple-factoid results and application-centered benchmarks show why those conditions must remain visible. (aclanthology.org) (research.google) (huggingface.co)

The proposed protocol makes the output requirement concrete. Freeze observed sources and verdicts. Keep input tokens, instructions, model configuration, and eligible evidence fixed. Move evidence through predeclared locations, preserve linked case identifiers, and judge correctness independently from citation recovery and completeness. Report appropriate abstention, contradictory evidence, and the raw transitions behind aggregate scores. Versioned provenance and separate attribution evaluation support these requirements. (www.w3.org) (collaborate.princeton.edu)

The hypothetical scorecard demonstrates how a paired effect and uncertainty interval can be calculated without presenting invented model performance as empirical evidence. Its numbers are illustrative; a reader's decision requires actual held-out cases and a tolerance matched to the intended use. Paired resampling preserves the case relationships that the experiment was designed to compare. (docs.scipy.org)

A larger context window expands the material that may be supplied. Reliability requires a stronger observable property: the answer remains acceptable, the source is recoverable, and the supporting passage still justifies the value across the placements and lengths the application encounters. That property should be demonstrated through the application test, with its uncertainty and scope stated explicitly. (arxiv.org) (princeton-nlp.github.io)

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://aclanthology.org/2024.tacl-1.9
2. https://research.google/pubs/retrieval-quality-at-context-limit
3. https://aclanthology.org/2024.emnlp-main.322
4. https://proceedings.mlr.press/v267/modarressi25a.html
5. https://longbench2.github.io

6. https://collaborate.princeton.edu/en/publications/enabling-large-language-models-to-generate-text-with-citations
7. https://www.w3.org/TR/prov-dm
8. https://arxiv.org/abs/2305.14627
9. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
10. https://princeton-nlp.github.io/HELMET
11. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
12. https://docs.cohere.com/docs/rag-citations
13. https://platform.claude.com/docs/en/build-with-claude/citations
14. https://cs.stanford.edu/~nfliu/papers/lost-in-the-middle.tacl2023.poster.pdf
15. https://aclanthology.org/2026.acl-long.1937
16. https://ai.google.dev/gemini-api/docs/tokens
17. https://ai.google.dev/gemini-api/docs/long-context
18. https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices
19. https://www.microsoft.com/en-us/research/project/minference-million-tokens-prompt-inference-for-long-context-llms/faq
20. https://aclanthology.org/2025.ijcnlp-long.46.pdf
21. https://deepmind.google/research/publications/117639
22. https://arxiv.org/abs/2404.06654
23. https://github.com/NVIDIA/RULER/blob/main/README.md
24. https://aclanthology.org/2025.ijcnlp-long.46
25. https://axi.lims.ac.uk/paper/2406.10149
26. https://chen-yingfa.github.io/research_posts/2024-infinitebench
27. https://huggingface.co/blog/helmet
28. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html
29. https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf
30. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/evaluate-judge-model
31. https://aclanthology.org/2024.tacl-1.9.pdf
32. https://aclanthology.org/2024.emnlp-main.322.pdf
33. https://www.statsmodels.org/stable/generated/statsmodels.stats.contingency_tables.mcnemar.html
34. https://www.anthropic.com/news/prompting-long-context

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.