Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

RAG chunking sensitivity: can boundaries reverse answers?

> 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 · 2026-10-21 · rag chunking sensitivity · rag chunk boundary effects · rag chunking evaluation · rag answer consistency · context engineering · retrieval sensitivity · evidence traceability · chunk overlap · late chunking

Original article: https://reliablecontext.com/articles/rag-chunk-boundaries-supported-answers



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology: source-native support and annotations
4. Boundary-phase design and controls
5. Configurations and budget controls
6. Measurement and the boundary ledger
7. Oracle diagnostics and paired uncertainty
8. Analysis of key segments
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

Retrieval-augmented generation (RAG) chunking sensitivity is the extent to which segmentation choices change retrieved support or the answer produced. A boundary can separate a matched rule from the exception that determines its application. Existing technical coverage already discusses dependent exceptions, source offsets and fixed-budget comparisons; those checks are established practice. This report proposes a narrower addition: a **paired boundary-phase sweep** that keeps source text and questions fixed while moving segmentation boundaries, then reports average and worst-phase outcomes. It supplies a reusable annotation specification, not a newly executed benchmark. (supermemory.ai)

Bennani and Moslonka (2026) evaluate **SPLADE** retrieval with **Ministral-8B-Instruct-2410** on Natural Questions short-answer questions over English Wikipedia. (arxiv.org) Their experiment reports no measurable overlap benefit and excludes rerankers and late-interaction models. (arxiv.org) This finding does not establish that overlap cannot preserve adjacent mandatory conditions elsewhere.

Representation context and visible evidence require separate evaluation. **Late Chunking** contextualizes token embeddings before applying chunk boundaries at pooling time; it does not, by itself, put every contextualizing sentence into the returned quotation. **LlamaIndex** sentence-window retrieval instead illustrates replacing a retrieved sentence with surrounding text before generation. Both can affect a pipeline, but only the actual model-visible evidence can establish whether the required condition survived. (arxiv.org) (developers.llamaindex.ai) Citation quality also needs its own verdict: Gao and colleagues' ALCE benchmark evaluates correctness and citation quality separately. (aclanthology.org)

The proposed release decision therefore has distinct gates for complete retrieved support, correct condition use and resolvable attribution. Compare the candidate and baseline under a **fixed context-token budget**, with fixed top-k reported separately. Accept jointly sufficient evidence across chunks. Include an oracle full-support prompt, preserve failures and successes, and estimate paired uncertainty without treating offsets from the same question as independent observations. SciPy's paired bootstrap preserves correspondence through shared resampling indices; the query and document-cluster design specified here is an analyst

recommendation. (docs.scipy.org) The release threshold must be set before observing results, with uncertainty and benchmark comparisons reported, consistent with the National Institute of Standards and Technology's (NIST) evaluation guidance. (airc.nist.gov)

Introduction and Background

A retrieved sentence can be relevant and still be insufficient. A rule might apply only to an explicitly defined population, a numerical limit might have a nearby exception, or a pronoun might refer to a definition in the preceding paragraph. The evaluation target is therefore **complete support**, not merely a text match. HotpotQA's sentence-level supporting facts provide a research precedent for recording evidence separately from the answer. (aclanthology.org)

A **chunk boundary** is a segmentation cut in a specific representation of a document. A **source span** identifies a passage in an immutable source version. A **mandatory qualifier** is a condition, exception, definition or scope statement whose omission can change the justified answer. A **sufficient support set** is an adjudicated collection of spans that jointly entails the expected answer, including its necessary qualifications. These are the operational definitions used in this proposed protocol.

Pinecone describes chunking as breaking large texts into smaller segments and recommends evaluation on representative data and queries. (www.pinecone.io) For Reliable Context's context-engineering audience, the practical decision is whether an indexing release preserves the evidence needed to defend an answer. This is an assessment of chunking and retrieval policies, with no publisher product included as a candidate.

The question is narrower than general answer accuracy: **Can an arbitrary segmentation phase reverse a support or answer verdict when the underlying source has not changed?** Supermemory already discusses exception dependencies and source offsets. The incremental contribution here is the paired estimate across preregistered phases, an explicit support annotation and a release rule. (supermemory.ai)

This report separates published findings from proposed measurements throughout. No original retrieval run, answer-reversal rate, production latency or measured cost is supplied. Research was accessed on the publication date, October 21, 2026. Numerical benchmark findings below remain attached to their original configurations. Evaluation artifacts should document test sets, metrics and tools, as NIST recommends. (airc.nist.gov)

Methodology: source-native support and annotations

Annotate the complete evidentiary unit

Begin with immutable source versions and questions written independently of chunker output. An annotator reads the full source, identifies the justified answer and marks alternative sufficient support sets. A second adjudication pass resolves whether an adjacent sentence is genuinely necessary or merely topical. MultiHop-RAG similarly associates questions and answers with supporting evidence, although its benchmark is not the boundary-phase protocol proposed here. (openreview.net)

Use these annotation fields:

- **Source identity:** record the document identifier, observation date, version and content digest.
- **Answer contract:** specify the permissible answer, its scope and any required qualification.
- **Support alternatives:** enumerate each independently sufficient set of source-native spans.
- **Dependency type:** label exception, definition, referent, numerical scope or cross-reference.
- **Missing-condition behavior:** specify whether incomplete evidence requires abstention or permits a narrower answer.
- **Adjudication state:** preserve disagreements, decisions and the reason for requiring each span.

Hypothetical Example: consider this invented source: “Requests may be retained for 30 days. This limit is 7 days for requests marked temporary.” The question is, “What retention limit applies to requests marked temporary?” The expected answer is at most seven days. The qualifier's “This limit” refers to the preceding retention rule. If only the default sentence is retrieved, the answer must not silently apply its default to temporary requests.

The following JavaScript Object Notation (JSON) is a reusable annotation fixture. Its integers are illustrative coordinates and rule values, not benchmark observations. Offsets are half-open UTF-8 byte positions in the exact ASCII source shown.

```
{
  "fixture_id": "temporary-retention-v1",
  "source_version": "sha256:ad0875163c10c4b7fd2d6efed6efae515cef0f8ea236b45216cb8ef12c8a5563",
  "source_text": "Requests may be retained for 30 days. This limit is 7 days for requests marked
temporary.",
  "coordinate_system": "utf8-byte-half-open",
  "spans": {
    "rule": [0, 37],
    "qualifier": [38, 89]
  },
  "question": "What retention limit applies to requests marked temporary?",
  "sufficient_support_sets": [["rule", "qualifier"]],
  "expected_answer": "At most 7 days.",
  "if_qualifier_missing": "Abstain on the temporary-request limit."
}
```

Coordinates alone are insufficient provenance. The Natural Questions repository uses both byte offsets and token indices. (github.com) Azure's Text Split skill exposes offsets in distinct encodings, while World Wide Web Consortium (W3C) text selectors use Unicode code points and specify normalization rules. These coordinate systems must be explicitly translated rather than treated as interchangeable. (learn.microsoft.com) (www.w3.org)

For every transformation, test that a normalized selection maps back to the intended original quotation. Preserve a mapping through whitespace collapse and tokenizer processing, test repeated passages and non-ASCII text, and verify the source slice exactly. W3C's quote selector uses surrounding text to distinguish repeated occurrences; its position-selector guidance warns about source changes. (www.w3.org)

Boundary-phase design and controls

For a token-window splitter, define a chunk capacity C and overlap O, giving stride S = C minus O. A phase specifies the first interior cut; subsequent cuts follow the same stride. Create core intervals between those cuts and the source edges, then extend each core by up to O following tokens. Interior chunks have capacity C; retain shorter edge chunks so every original token remains represented. Declare this edge policy in the manifest.

Preregister phases spanning the stride, including cuts before, between and after annotated dependent sentences where the splitter permits them. Do not add text to force a favorable or unfavorable cut. Prefix padding, if investigated, is a separate intervention because it changes the input. Structural splitters need a different, explicit perturbation rule: their natural boundaries need not implement arbitrary token phases.

For comparisons across different strides, preregister shared phase labels and their mapping to each configuration, then record the actual cuts. An unshifted structural alternative can serve as the same control across baseline phases. Its repeated appearance does not create additional independent observations. Report within-configuration fragility separately from the paired candidate-versus-baseline effect.

Freeze the rest of the experiment:

- **Parsing:** preserve original bytes, extraction output, normalization and document order.
- **Indexing:** fix tokenizer, embedding identity, indexing settings and source versions.
- **Retrieval:** fix query processing, filters, search parameters and reranking policy.
- **Generation:** fix model identity, prompt, decoding settings and available tools.
- **Evaluation:** fix questions, support sets, answer rubric and citation checks.
- **Randomness:** pair seeds where supported and record repeated-run variation separately.

Keep a separate manifest for each index build and link every result to it. Preserve the chosen phase list before running retrieval so an unfavorable offset cannot be quietly removed from the reported experiment.

This intervention estimates sensitivity to the segmentation configuration. It does not prove that every observed change arose solely from a missing sentence: different chunks can also change representations and ranking. The ledger must retain those intermediate stages.

Configurations and budget controls

The primary comparison holds the evidence budget constant. The secondary comparison holds the number of retrieved candidates constant. Supermemory already recommends both views, and Microsoft's guidance says optimal overlap depends on content and use case. (supermemory.ai) (learn.microsoft.com) Neither view should be presented as the other.

Table 1 specifies the candidate arms. These are proposed configurations, and all outcome cells remain unmeasured until the protocol is run.

Configuration	Retrieval and presentation policy	Primary comparison	Required accounting
Baseline splitting	Existing splitter at preregistered phases	Same evidence-token ceiling	Complete support, answer verdict, actual prompt tokens
Overlap variant	Same splitter and capacity, preregistered overlap	Same evidence-token ceiling	Repeated source spans, indexed tokens, displaced evidence
Structural alternative	Preserve declared sentence or section units with an explicit size cap	Same evidence-token ceiling	Actual boundary locations, fallback splits, oversized units
Parent-window recovery	Retrieve child units, expand to authorized source windows	Same evidence-token ceiling after expansion	Child hits, unique parents, truncation, permissions and version

The table specifies mechanisms to compare, not a ranking. **LangChain** documents retrieving small chunks and then their parents; the parent can be a larger chunk or the raw document. (reference.langchain.com) **MongoDB** likewise distinguishes initial child retrieval from returning a full parent. A fixed count of children therefore need not imply a fixed count or size of final context units. (www.mongodb.com)

Account for the final prompt

For the primary view, reserve a fixed allowance for instructions and question text, then apply the same evidence ceiling to every arm. Record both the allowance and actual full-prompt usage under the generator's tokenizer. Report the output-token allowance separately. Microsoft notes that chunk token length varies with contents, reinforcing the need to count actual tokens rather than infer them from chunk labels. (learn.microsoft.com)

Use explicit packing rules:

- **Rank order:** define how eligible retrieved units enter the context.
- **Expansion order:** state whether parent expansion occurs before or after reranking.
- **Deduplication:** merge identical source intervals deterministically and retain their provenance.
- **Overflow:** apply a declared packing or truncation rule and log rejected material.
- **Formatting:** count source labels, citation identifiers and separators in prompt usage.
- **Budget exhaustion:** preserve the run as a failure if required support no longer fits.

For fixed-top-k, state whether k refers to raw candidates, reranked children, unique parents or final prompt units. Report actual tokens alongside every result. Milvus explicitly distinguishes embedding retrieval units from the chunks passed to the model. Its documentation also warns that expanding windows can introduce interfering information. (milvus.io)

Treat recovery as a policy with constraints

LlamaIndex's sentence-window example replaces matched sentences with surrounding sentences, while Pinecone describes query-time neighboring-chunk expansion. These are useful recovery mechanisms to test, not guarantees that every exception will fit. (developers.llamaindex.ai) (www.pinecone.io) Similarly, sentence-oriented splitting is no guarantee of intact oversized sentences: Elastic documents splitting a sentence when it exceeds the configured maximum. (www.elastic.co)

Parent expansion must preserve the child hit's source version and the reader's authorization. Qdrant documents metadata filtering, including access-related filtering, and Elastic distinguishes document-level from field-level read restrictions. (qdrant.tech) (www.elastic.co) The proposed test should reapply the applicable policy to every expanded span. A child being readable is not sufficient evidence that its entire parent is permitted.

Measurement and the boundary ledger

Separate support, answer and attribution

A relevant hit does not establish complete support, and a correct answer does not establish a valid citation. ALCE treats fluency, correctness and citation quality as separate dimensions. (aclanthology.org) LlamaIndex's example also permits inspection of the actual sentence window sent to the model, making that window a more appropriate evidence object than the initial matched sentence alone. (developers.llamaindex.ai)

Table 2 defines the proposed metrics. “Answerable” refers to the authorized, frozen source collection, not the model's background knowledge.

Metric	Definition and denominator	Interpretation
Segmentation integrity	Required source spans reconstructible from indexed units and their source mappings / required spans	Detects ingestion or mapping loss before retrieval
Complete-support recall	Answerable questions retrieving every member of at least one adjudicated sufficient support set / answerable questions	Measures complete retrieved evidence, including jointly sufficient chunks
Critical-condition error rate	Questions whose answer contradicts or omits an adjudicated mandatory condition / condition-dependent questions tested	Separates consequential scope errors from wording differences
Support boundary fragility	Questions whose complete-support verdict changes across phases / questions tested	Depends on the disclosed phase set
Answer boundary fragility	Questions whose substantive answer-verdict category changes across phases / questions tested	Categories distinguish correct, incorrect and abstaining answers
Boundary-induced regression	Matched query-phase pairs passing under the baseline and failing under the candidate / matched pairs tested	Must be reported separately for support and answers
Resolvable attribution	Material answer claims with accessible citations that identify sufficient supporting source spans / material answer claims	Measures support and resolvability, not identifier formatting alone

These denominators prevent a large volume of easy questions from hiding errors on the condition-dependent subset. A wording change alone does not change the substantive answer verdict; a changed scope, value or qualification can. Report abstention separately from a wrong conditional answer. An abstention lowers answer coverage but is not itself a critical-condition error. Also report false assertions on genuinely unanswerable questions; the official Natural Questions dataset description includes articles that may not contain an answer. (huggingface.co)

An evidence set can span multiple chunks. Accept their union when the cited, versioned passages jointly support the answer. A single-unit-complete-support flag is diagnostically useful, but requiring every dependency to fit into one large chunk would unfairly reject valid joint evidence. HotpotQA's supporting-fact annotations and MultiHop-RAG's evidence-associated questions provide relevant precedents for evidence beyond an isolated answer string. (aclanthology.org) (openreview.net)

The boundary ledger should contain:

- **Query record:** stable question identifier, expected answer and condition category.
- **Source record:** document/version identifiers, digest and observation date.
- **Support record:** rule, qualifier and alternative sufficient-set identifiers.
- **Boundary record:** phase, stride, actual cut coordinates and edge policy.
- **Retrieval record:** ranked hits, scores, source intervals and expansion lineage.
- **Budget record:** evidence ceiling, actual prompt tokens and repeated source tokens.
- **Answer record:** complete generated text, substantive verdict and condition use.
- **Citation record:** cited version and spans, resolution result and adjudicated verdict.

Store the literal evidence supplied to the model alongside reconstructible source references. Azure provides positional and ordinal outputs, and Qdrant documents attached JSON metadata; neither fact alone establishes that an application's subsequent transformations preserve those references. (learn.microsoft.com) (qdrant.tech)

Oracle diagnostics and paired uncertainty

Run an **oracle full-support prompt** using an adjudicated sufficient set, the same question and the same answer rubric. Keep a production-budget oracle and, if necessary, a separately labeled diagnostic oracle with more space. A diagnostic pass outside the production budget cannot establish deployable performance. Keep budget-infeasible answerable questions in the headline support-recall denominator, flag them, and report a feasible-only diagnostic separately.

Preserve every completed run, including apparently easy successes, and label infrastructure failures separately. Do not substitute a successful retry for an unsuccessful answer without retaining both records. Repeated generations should estimate model variation within a configuration, while the phase sweep estimates the indexing intervention.

Compare the outcomes carefully. Candidate retrieval failure with oracle success supports a retrieval-stage explanation. Failure in both leaves a generator, prompt, ambiguity or annotation problem unresolved. Success without retrieved support may reflect background knowledge; it still fails the evidence requirement. MultiHop-RAG separately evaluates evidence retrieval and answering from evidence, supporting this separation of test objects. (openreview.net)

An oracle failure should trigger inspection of the full-source annotation before any boundary remedy is credited. Distinguish an infeasible evidence budget, an unresolved referent and a model that ignores visible support. Each diagnosis calls for a different next experiment.

For uncertainty, resample paired query records and carry their phase vectors together. SciPy's paired option uses identical resampling indices across arrays, and Koehn's paired-bootstrap discussion compares systems on the same test set. (docs.scipy.org) (aclanthology.org) If multiple questions share a document, cluster by document so their dependence is retained. Report both the matched average difference and worst-phase performance. Define worst-phase support as the minimum, across phases, of aggregate query support. Taking each query's worst phase first gives a different, more adversarial statistic. Recompute the declared minimum within each bootstrap replicate. Worst-phase performance is a stress-test statistic, not an estimated production frequency.

Analysis of key segments

An average over heterogeneous questions can obscure precisely the dependencies the release decision concerns. Define segments before inspecting candidate results and report their denominators, coverage and uncertainty. Pinecone's representative-data recommendation is consistent with testing the deployed question distribution, although the segment taxonomy below is proposed here. (www.pinecone.io)

Use these segments:

- **Rule and exception:** the default sentence is relevant, but an exception changes its application.
- **Definition and referent:** a term, abbreviation or pronoun requires an earlier definition.
- **Number and scope:** a value is meaningful only with its population, period, unit or denominator.
- **Cross-reference:** a local sentence delegates its operative condition to another section.
- **Joint evidence:** support requires multiple independently located passages.
- **Unanswerable control:** the authorized source collection does not justify the requested answer.

For each segment, preserve easy successes as well as failures. Selecting only known boundary failures can test repair capability, but it cannot estimate corpus-wide fragility. Conversely, a representative sample may contain too few mandatory conditions to assess a high-consequence release. Report a representative estimate and an enriched stress set separately, with no pooled prevalence claim.

The Natural Questions repository says approximately **90%** of its short answers contain a single span. That describes answer annotation, not proof that those answers require no surrounding qualifications. (github.com) It nevertheless cautions against treating one short-answer benchmark as a complete model of dependency-heavy organizational questions. HotpotQA explicitly supplies supporting sentences, while MultiHop-RAG supplies associated evidence for retrieval and reasoning evaluation. (aclanthology.org) (openreview.net)

Distinguish representation from quotable support

Late Chunking's full-input contextualization occurs before chunk-level pooling, and the method still needs boundary cues. Its evaluation includes ranking chunks and deriving document rankings. (arxiv.org) Consequently, an embedding may represent a sentence in light of an exception even when the returned source interval omits that exception. The proposed test must examine the returned interval and final citation, not infer visible support from embedding quality.

Anthropic's contextual retrieval also adds chunk-specific explanatory context before retrieval indexing. (www.anthropic.com) Generated explanatory context should be tagged as such in the ledger. It can help retrieval, but an uncited generated explanation cannot replace the original condition as source evidence. This distinction is an evaluation requirement of the proposed protocol, not a claim about a measured failure rate of either method.

Keep neighboring experiments separate

A boundary-phase sweep fixes source bytes and parsing. Changing table serialization would change representation. Compressing selected context would change the evidence after selection. Reordering evidence in the prompt would change position. Those are different interventions and need their own controls.

Liu and colleagues' *Lost in the Middle* reports sensitivity to relevant information's prompt position. (arxiv.org) A later Google Research report finds that effect absent in its simple factoid tests with **Gemini 2.5 Flash**. (research.google) The contrast is a reason to disclose task and model identity, not a basis for extrapolating either result to arbitrary boundaries. Keep evidence ordering constant in the primary phase sweep and label any factorial extension separately.

Data Analysis and Evidence

What the published numbers establish

Table 3 separates empirical findings from the measurements still needed for this protocol. Metrics from different experiments are not directly comparable.

Source and configuration	Reported quantitative evidence	Boundary-phase interpretation
Bennani and Moslonka, Natural Questions with SPLADE and Ministral	Target chunk sizes **50 to 500**, step **50**; context budgets **500, 1,000, 2,500, 5,000 and 10,000 tokens**. (arxiv.org)	A specified chunking experiment, not a phase-sweep estimate
Same study, overlap analysis	BERTScore differences at most **0.004**; exact-match differences at most **0.001**; **95%** question-bootstrap intervals. (arxiv.org)	Evidence for negligible measured benefit in that setup; no universal overlap guarantee
Anthropic, contextual embeddings plus contextual BM25	Top-20 retrieval failure fell from **5.7% to 2.9%**, reported as a **49%** relative reduction. (www.anthropic.com)	Retrieval evidence under its configuration; not complete-condition recall
Proposed boundary-phase protocol	Complete-support recall, fragility, condition errors, tokens, latency and cost: **not measured**	Requires execution on the reader's frozen corpus and question set

The table supports a bounded conclusion: retrieval design can matter, but the required support-survival estimate is absent. There is no defensible numeric answer here to “How often do boundaries reverse supported answers?” A deployment must measure that frequency over its declared questions and phases.

The experiment specifies **Ministral-8B-Instruct-2410**, temperature **0.1**, while the abstract says “Mistral-8B.” (arxiv.org) Setup overlap is **0% or 20%**; findings discuss **10% to 20%**. (arxiv.org) No exact evaluated question count or reproducible subset-selection recipe appears in the fetched text. The authors call for validation

beyond general text-centric corpora. (arxiv.org)

Anthropic's aggregate graphs use **Gemini Text 004**, top-**20** retrieval and failure defined as one minus recall@20. (www.anthropic.com) It reports **5.7% to 3.7%** failure with contextual embeddings and **5.7% to 1.9%** with the combined contextual methods and reranking. (www.anthropic.com) These are vendor-reported experiment results, not measurements of mandatory-condition use or final citation resolution.

To study **chunk size and answer accuracy**, repeat the phase design at each preregistered capacity. Compare matched questions and inspect size-by-phase interactions. A size that improves average correctness could still worsen condition errors in a particular phase; this is a hypothesis to test. Keep the context ceiling fixed so larger chunks do not receive an undeclared evidence allowance.

Calculate the tradeoffs rather than assume them

Hypothetical Example (calculation): for long, regularly tiled input, the overlap-only token expansion approaches C divided by C minus O, ignoring document edges and metadata. An overlap fraction of **20%** therefore gives **1.25 times** the indexed token volume, or **25%** more. This is a geometric calculation, not a measured storage or billing result. Qdrant describes overlap's storage tradeoff, but real index expansion must be measured. (qdrant.tech)

Record indexed units, unique and repeated source tokens, stored bytes, embedding usage, query latency and actual billed cost. Separate one-time indexing from per-query cost. Compare latency under the same load and report the measurement method. Parent expansion can lower duplicate retrieval while increasing visible tokens; neither effect should be assumed from the mechanism alone.

Report each phase, the phase average and the worst phase, followed by the paired candidate-minus-baseline change. SciPy supports bootstrap confidence intervals, but the declared resampling unit and question mix determine what those intervals describe. (docs.scipy.org) Do not turn a stress-set worst case into an expected production error rate.

Implications and Future Directions

Preregister the release decision

Before evaluating a candidate, choose tolerances for complete-support loss, critical-condition errors, resource usage and unresolved attribution. Define how intervals crossing a tolerance affect the decision. A candidate should not be called safe merely because a small sample produces a statistically inconclusive comparison. NIST recommends uncertainty measures and comparison with performance benchmarks. (airc.nist.gov)

A practical release record includes:

- **Population:** eligible documents, reader clearance and intended question distribution.
- **Intervention:** candidate settings, baseline settings and phase construction.
- **Acceptance rule:** permitted support loss and condition-error tolerance by segment.
- **Uncertainty rule:** resampling units, interval method and insufficient-evidence handling.
- **Resource rule:** token ceiling, latency allowance and measured cost boundary.
- **Disposition:** accept, reject or collect more evidence, with reproducible reasons.

Keep authorization and version consistency as hard constraints. Elastic's document and field security describe distinct read restrictions; an expansion policy must respect the restrictions relevant to the actual returned spans. (www.elastic.co) A new parent fetched from a changed source version is also a different evidentiary object. W3C's warning about brittle position selectors reinforces retaining a versioned representation with the selection. (www.w3.org)

Preserve the artifact after the decision

The reusable deliverable is the annotation file plus the boundary ledger, not a single leaderboard score. Export the question set, source digests, permitted support sets, actual boundaries, retrieved intervals, model-visible context, generated answer and citation verdict. Where redistribution is prohibited, retain the fixture schema and reproducible identifiers inside the authorized environment.

LangChain exposes a parent splitter, and Milvus describes parent-chunk recovery; both illustrate why a manifest needs to distinguish retrieval and presentation policy. (reference.langchain.com) (milvus.io) Source ordinals can assist reconstruction, but the original source/version identity remains necessary. (learn.microsoft.com)

For future work, estimate the prevalence of genuine qualifier dependencies before claiming broad operational impact. Evaluate whether boundary sensitivity changes with query type, model family, retrieval architecture and available context. Treat those as planned investigations. NIST's recommendation to test before deployment and during operation supports revisiting the decision when the pipeline or corpus changes. (airc.nist.gov) Neither a successful fixture nor one favorable benchmark establishes permanent robustness.

Conclusion

RAG chunking sensitivity is a question about preserving sufficient evidence under a controlled indexing change. The relevant outcome is not whether a retrieved fragment resembles the question. It is whether the final context retains the conditions needed for the answer, whether the model applies them correctly and whether its citations resolve to the supporting source version.

The proposed boundary-phase sweep makes that question testable. Annotate complete support before splitting; freeze source bytes, parsing, questions and pipeline components; vary the declared segmentation phase; and compare a candidate with its baseline under a fixed evidence budget. Report fixed-top-k results separately. Accept valid joint evidence rather than requiring a single oversized chunk.

Keep the diagnostic layers distinct. A support failure, a condition-use error and an attribution failure require different remedies. An oracle support prompt can help localize the problem, but cannot justify production success outside the declared budget. Pair the analysis at the question level and preserve document dependencies when estimating uncertainty.

Published findings guide the test design without supplying its missing result. Existing guides already discuss boundary cases and budget controls; contextual embedding methods address a different object from visible source completeness. The useful addition is a reproducible estimate of support and answer fragility across disclosed phases, with a preregistered release decision. Until that experiment is run on the intended corpus, the appropriate performance entries remain **not measured**.

Source links

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

1. https://supermemory.ai/blog/text-chunking-strategies-for-rag
2. https://arxiv.org/html/2601.14123v1
3. https://arxiv.org/html/2409.04701v3
4. https://developers.llamaindex.ai/python/examples/node_postprocessor/metadatareplacementdemo
5. https://aclanthology.org/2023.emnlp-main.398
6. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
7. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
8. https://aclanthology.org/D18-1259
9. https://www.pinecone.io/learn/chunking-strategies
10. https://openreview.net/forum?id=t4eB3zYWBK
11. https://github.com/google-research-datasets/natural-questions
12. https://learn.microsoft.com/en-us/azure/search/cognitive-search-skill-textsplit
13. https://www.w3.org/TR/annotation-model
14. https://learn.microsoft.com/en-us/azure/search/vector-search-how-to-chunk-documents
15. https://reference.langchain.com/python/langchain-classic/retrievers/parent_document_retriever/ParentDocumentRetriever
16. https://www.mongodb.com/docs/atlas/ai-integrations/langchain/parent-document-retrieval
17. https://milvus.io/docs/how_to_enhance_your_rag.md
18. https://www.elastic.co/docs/explore-analyze/elastic-inference/inference-api
19. https://qdrant.tech/course/essentials/day-1/chunking-strategies
20. https://www.elastic.co/docs/deploy-manage/users-roles/cluster-or-deployment-auth/controlling-access-at-document-field-level
21. https://huggingface.co/datasets/google-research-datasets/natural_questions
22. https://aclanthology.org/W04-3250.pdf
23. https://www.anthropic.com/research/contextual-retrieval
24. https://arxiv.org/abs/2307.03172
25. https://research.google/pubs/retrieval-quality-at-context-limit

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.