Document Overlap Is Not Evidence Continuity: Measuring Retrieval Jitter in Citation-Based RAG Evaluation

Punitha Ponnuraj
Independent Researcher / United States
punitha.p.raj@gmail.com

Abstract

RAG evaluations often rely on citations or retrieved evidence traces for correctness checks, provenance claims, and audits, implicitly assuming that evidence remains reproducible under routine retrieval settings. We test this assumption in a controlled diagnostic study where queries, embeddings, and decoding are fixed while retrieval depth, chunk size, and overlap vary. We call the resulting change in attributed evidence *retrieval jitter* and measure evidence identity at two levels: document (`doc_id`) and exact cited span (`doc_id`, `span_hash`). Across BEIR ArguAna and SciFact, we observe a consistent Stability Gap: document overlap remains moderate while span overlap often collapses, including many cases of total span turnover despite non-empty retrieval. We interpret span-level instability as a diagnostic of exact evidence-trace reproducibility, not semantic equivalence. These findings motivate reporting stability diagnostics alongside citation-based evaluation metrics for more reproducible evaluation practice.

1 Introduction

Evaluation of generative AI systems increasingly uses retrieval-augmented generation (RAG) pipelines that provide citations or retrieved evidence as evaluation artifacts for correctness verification, provenance analysis, regression testing, and third-party auditing. In many such workflows, these evidence traces are implicitly assumed to remain stable across runs. This raises a measurement question that is not usually reported explicitly: how stable is the attributed evidence trace under routine retrieval perturbations when the query, embedding model, and decoding behavior are otherwise fixed? In this work, we conduct a controlled empirical study of *retrieval jitter*: systematic variation in attributed evidence caused by minor retrieval configuration perturbations, even when queries, embeddings, and decoding behavior are held constant.

Many current RAG evaluation practices emphasize document-level retrieval metrics, document-level citation support, or passage-level relevance (Es et al., 2024; Caspari et al., 2024; Thakur et al., 2025), but they do not explicitly test whether the exact supporting evidence spans remain stable under routine retrieval configuration changes. To operationalize this phenomenon, we define a hierarchical evidence identity protocol at two levels - document-level (`doc_id`) and span-level (`doc_id`, `span_hash`), where span hashes are computed over normalized cited text. On BEIR ArguAna and SciFact, varying retrieval depth, chunk size, and chunk overlap yields a consistent Stability Gap: document overlap remains moderate while span-level overlap often collapses, including cases of zero span overlap despite non-empty retrieval. As a supplementary exploratory analysis, we also probe whether evidence instability is necessarily mirrored by answer-level change. This probe is not part of the paper's primary claim, but it helps clarify why evidence stability matters in practice: if answer similarity remains high while attributed evidence changes substantially, routine system updates can create a form of silent jitter in which outputs appear stable even though the supporting evidence trace has shifted. We argue that routine reporting of span-level stability diagnostics alongside citation-based metrics can improve transparency in settings where reproducibility or auditability matters.

Contributions. We introduce retrieval jitter and define a hierarchical evidence-identity protocol at document and span levels for measuring evidence stability under routine retrieval-configuration changes; show across ArguAna and SciFact that document-level overlap can mask substantial span-level turnover, making document-only stability checks insufficient for some citation-based evaluation and auditing workflows; and release RagCiteCheck, an open-source Python/CLI har-

Proceedings of the Workshop on Evaluating Evaluations (EvalEval), pages 221–226

ness for logging retrieval evidence and computing document- and span-level stability diagnostics.

2 Related Work

Prior work has shown that correctness does not guarantee faithful attribution in retrieval-augmented generation, and cited evidence may fail to fully justify model outputs (Wallat et al., 2025; Liu et al., 2023). Evaluation frameworks such as RAGAS assess answer relevance and faithfulness, but do not explicitly test whether the same evidence is retrieved consistently across routine system updates (Es et al., 2024). Related work on long-context use, hallucination, and retrieval stability further suggests that answer quality and evidence grounding can diverge under system variation (Wang et al., 2025; Hsia et al., 2025; Zhang et al., 2026). Chunking and segmentation are also known to affect retrieval precision, context coverage, and grounding quality (Bhat et al., 2025; Schreieder et al., 2025; Stäbler et al., 2025). Our work builds on these observations by focusing specifically on the construct validity of document-level provenance as a proxy for evidence continuity under routine configuration drift.

3 Methodology

Evaluation setting. Experiments are conducted on two evidence-centric benchmarks from the BEIR suite: ArguAna and SciFact (Thakur et al., 2021; Wadden et al., 2020) using deterministic decoding and fixed embedding model. Routine retrieval configuration such as retrieval depth $k \in \{5, 10, 20\}$, chunk size $c \in \{128, 256\}$ tokens, and chunk overlap $o \in \{0, 32\}$ are varied. To avoid inflating stability with inherently unanswerable cases, we restrict evaluation to queries with at least one relevant qrels document. For ArguAna we use a fixed 400-query subset, and for SciFact we use all answerable evaluation queries.

Hierarchical evidence identity. Document-level identity captures the retrieved source document and is identified by `doc_id`. We use *span* to mean the retrieved textual fragment, i.e., the chunk/node text surfaced as evidence to the model. Span-level identity is represented as `(doc_id, span_hash)`, where `span_hash` is computed by hashing normalized retrieved text. Span-level identity is intentionally defined using a strict exact-match criterion over normalized cited text and is therefore sensitive to boundary effects under chunking changes. We therefore interpret span-hash instability as a measure of exact evidence-trace reproducibility, not as a direct test of semantic evidence equivalence. Notably, span-level turnover is also observed in cases where document identity remains unchanged across configuration pairs, indicating that the Stability Gap is not explained solely by document substitution. The purpose of the span-level metric in this paper is to diagnose whether the same attributed textual evidence is reproduced across routine configuration changes, not to decide whether two different spans are semantically interchangeable.

Stability metrics. We compute pairwise Jaccard overlap across configuration runs at both document and span levels (J) and report mean overlap as well as worst-case per-query stability. For each query, document-level stability measures overlap between retrieved document identifiers, while span-level stability measures overlap between span identities. We report mean pairwise stability (J_{avg}), worst-case per-query stability ($J_{\min}$), and collapse rates defined as the percentage of queries where $J_{\text{span}} = 0$ despite non-empty retrieval. We also compute flip-rates, defined as the fraction of configuration pairs where stability falls below a threshold (e.g., $J < 0.5$). We also report null diagnostics for empty evidence sets and non-empty/empty transitions across configurations.

RagCiteCheck takes JSONL evidence logs as input, extracts document- or span-level evidence identities, and outputs pairwise stability, flip-rate, null-evidence, and worst-case per-query diagnostics; code and an archival snapshot are available online.[1]

Exploratory evidence–answer probe. As a supplementary exploratory check, we generate answers under each configuration and compute semantic similarity between outputs using SBERT cosine similarity (Reimers and Gurevych, 2019). We report drift rates below similarity thresholds and analyze correlation between answer similarity and span-level evidence stability. This analysis is not used to establish the paper's main evidence-stability claim; rather, it is included to test whether substantial evidence-trace turnover can occur without correspondingly large answer-level change, i.e., a potential silent-jitter regime.

[1] https://github.com/ppon1086/ragcitecheck; https://doi.org/10.5281/zenodo.18645598.

4 Results

Aggregate stability patterns. Across both datasets, document-level and span-level stability diverge in a consistent way. Because span-level identity uses exact normalized-text matching, some instability under chunk-size changes may reflect boundary-sensitive resegmentation. Accordingly, we interpret span-level instability as a diagnostic of exact evidence-trace reproducibility rather than semantic equivalence. On ArguAna, mean document-level overlap is 0.604 while mean span-level overlap drops to 0.360. On SciFact, mean document-level overlap is 0.520 and mean span-level overlap is 0.224. The gap is larger on SciFact, where the mean gap ratio reaches 2.321, compared with 1.678 on ArguAna. Figure 1 shows this pattern visually: document overlap often stays in a moderate range while span overlap falls much lower.

Worst-case behavior is more prominent. The median worst-case span stability J_{min}, J_{span} is 0.000 for both datasets. In ArguAna, 70.5% of queries have at least one configuration pair with complete span turnover. In SciFact, that rises to 94.0% (Thakur et al., 2021; Wadden et al., 2020). Put differently, for many queries, there exists some routine configuration change that preserves non-empty retrieval while replacing the exact retrieved span entirely.

Representative collapse cases. The aggregate results are not driven by one or two outliers. Representative collapse pairs show the same structure. In ArguAna, one configuration pair yields $J_{doc} = 0.584$ but only $J_{span} = 0.0525$. In SciFact, a comparable pair yields $J_{doc} = 0.419$ and $J_{span} = 0.009$. Similar collapse pairs appear repeatedly across the grid. Importantly, span-level turnover is also observed in many cases where document identity remains unchanged, indicating that the observed instability is not solely attributable to document substitution or retrieval failure but to variation in the specific evidence fragments surfaced.

Null diagnostics. The gap is not explained by retrieval failure. Citation rate remains 1.0 across both datasets and null rate remains 0.0. Pairwise null transitions are also zero in the main experiment. So, the observed instability reflects span substitution rather than missing evidence.

Configuration effects. The strongest instability appears under chunk-size changes. Changing chunk size while keeping the rest of the retrieval setup close to baseline produces the largest drops in span-level overlap. Overlap changes produce intermediate effects. Retrieval-depth changes are milder in the tested grid. This suggests that segmentation choices are a main source of evidence instability, which is relevant because chunking is often treated as a technical setting rather than an evaluation variable.

Pair	Data	$\bar{J}_{span}$	Sim	Drift	ρ
Base vs o0	ArguAna	0.823	0.990	1	0.136
Base vs o0	SciFact	0.637	0.940	6	0.120
Base vs c128	ArguAna	0.060	0.949	6	0.038
Base vs c128	SciFact	0.032	0.767	24	-0.025

Table 1: Exploratory evidence–answer stability probe.

We include a supplementary exploratory probe to test whether evidence instability is always reflected in answer behavior. As shown in Table 1, under chunk-size perturbations, span stability drops sharply (e.g., mean $J_{\text{span}} = 0.032$ on SciFact), while answer similarity remains relatively high (mean 0.767). Spearman correlations are weak ($\rho \in [-0.025, 0.136]$), indicating that output similarity does not reliably track evidence continuity. Although this probe is not designed to support a broad answer-level claim, it highlights a practically important implication of the main findings: retrieval updates can preserve superficially similar outputs while altering the cited evidence trace underneath them, complicating regression interpretation and auditability.

5 Discussion

Our findings suggest that document-level overlap can be an incomplete proxy for exact evidence-trace reproducibility under routine retrieval changes. Many RAG workflows use document overlap or document identity as a coarse indicator of evidence continuity, but our results show that such coarse checks can mask the turnover in specific text spans presented to the model. This creates a diagnostic gap between source-document continuity and exact evidence-trace continuity. The main implication of this study is a measurement recommendation: when reproducibility, regression comparison, or auditability matters, document overlap alone may provide an incomplete picture of evidence stability. This instability has practical implications for evaluation practice, as retrieval configurations such as chunk size, overlap, and retrieval depth are routinely adjusted to manage la-

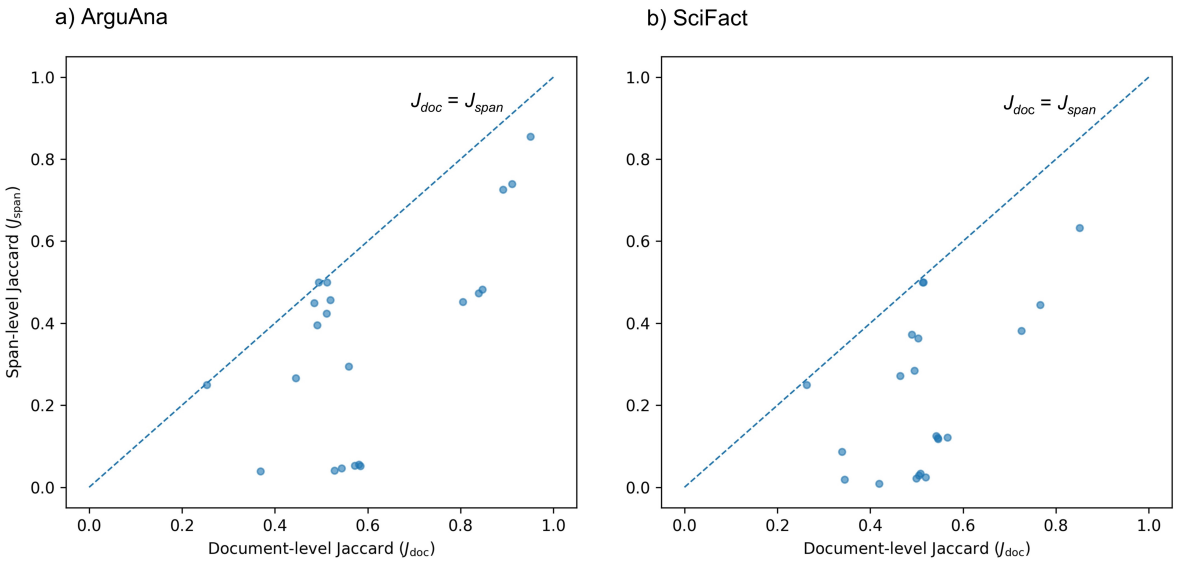

Figure 1: Document-level vs. span-level stability across retrieval configuration pairs on ArguAna and SciFact. Document overlap often remains moderate while span overlap collapses, revealing a Stability Gap.

tency, context budgets, and indexing constraints in production settings. These adjustments are usually treated as engineering tuning rather than evaluation-relevant variation. But our results suggest that such routine configuration changes can alter attributed evidence traces, and hence evaluations conducted before and after deployment updates may become difficult to interpret, even when answer similarity remains high. These findings also reflect a practical conflict between system developers and evaluation researchers. Developers tune retrieval for speed and cost whereas Evaluators and auditors need consistency, traceability, and reproducibility. Hence, span-level diagnostics should be treated as part of evaluation rather than as an optional analysis. Simple reporting practices such as monitoring span-overlap distributions, collapse rates, and null-evidence transitions can provide early indicators of evidence instability. By treating retrieval configuration as an evaluation variable rather than just an implementation detail, practitioners can improve the interpretability and auditability of RAG evaluation pipelines in real-world deployments. As an initial step of mitigation, we encourage incorporating span-level stability diagnostics into citation-based evaluation workflows.

Minimal Evidence Stability Reporting Protocol. Document-level overlap is therefore an incomplete proxy for exact evidence continuity under routine retrieval changes. For citation-based RAG evaluation, we recommend reporting the retrieval configuration, stability at both document and span levels, worst-case per-query span stability, collapse rates, and null-evidence diagnostics. These lightweight checks make regression comparisons and auditability easier to interpret when retrieval settings change.

6 Conclusion

We introduced retrieval jitter as a measurement challenge for citation-based RAG workflows. Through a hierarchical evidence identity framework, we showed that document-level identity checks can hide instability at the level of evidence spans surfaced to downstream reasoning modules. Across retrieval configuration changes on BEIR datasets ArguAna and SciFact, results show a consistent Stability Gap, including many cases of complete span turnover despite non-empty retrieval. Our supplementary exploratory evidence–answer probe suggests that answer similarity does not necessarily track evidence continuity, highlighting a potential silent retrieval jitter that can complicate regression interpretation and independent auditing. Together, these findings suggest that citation-based evaluation workflows relying only on document-level checks may provide an incomplete view of evidence continuity, and that span-level stability diagnostics are a lightweight addition for more interpretable, reproducible, and auditable RAG evaluation.

7 Limitations

We analyze stability under a controlled grid of retrieval configurations (top-k, chunk size, and overlap) using a fixed embedding model and deterministic decoding. Other sources of variability—such as retriever training changes, index refreshes, query reformulation, or stochastic generation—are outside the scope of this study. Because span-level identity is defined using exact normalized cited-text matching, some observed turnover may reflect boundary-sensitive resegmentation under chunking changes; accordingly, our analysis should be interpreted as measuring exact evidence-trace reproducibility rather than full semantic evidence equivalence.

Finally, while we propose span-level stability diagnostics as a reporting practice, we do not evaluate mitigation strategies such as stability-aware chunking, evidence anchoring, or retrieval regularization. Designing and validating such interventions remains future work.

References

Sinchana Ramakanth Bhat, Max Rudat, Jannis Spiekermann, and Nicolas Flores-Herr. 2025. Rethinking chunk size for long-document retrieval: A multi-dataset analysis. *Preprint*, arXiv:2505.21700.

Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. 2024. Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems. In *Proceedings of the 1st Workshop on Information Retrieval's Role in RAG Systems (IR-RAG@SIGIR 2024)*, pages 62–70.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158. Association for Computational Linguistics.

Jennifer Hsia, Afreen Shaikh, Zora Zhiruo Wang, and Graham Neubig. 2025. RAGGED: Towards informed design of scalable and stable RAG systems. In *Proceedings of the 42nd International Conference on Machine Learning*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Tobias Schreieder, Tim Schopf, and Michael Färber. 2025. Attribution, citation, and quotation: A survey of attribution practices in natural language generation. *Preprint*, arXiv:2508.15396.

Maximilian Stäbler, Steffen Turnbull, Tobias Müller, Chris Langdon, Jorge Marx-Gómez, and Frank Köster. 2025. The impact of chunking strategies on domain-specific information retrieval in retrieval-augmented generation systems. In *IEEE COINS*.

Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Assessing support for the TREC 2024 RAG track: A large-scale comparative study of LLM and human evaluations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2759–2763. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, pages 22–32. Association for Computing Machinery.

Baiqiang Wang, Dongfang Zhao, Nathan R. Tallent, and Luanzheng Guo. 2025. On the reproducibility limitations of rag systems. *Preprint*, arXiv:2509.18869.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *Preprint*, arXiv:2601.02993.

8 Appendices

A Retrieval Configuration Grid

Parameter	Values
Top-k	$\{5, 10, 20\}$
Chunk size	$\{128, 256\}$
Chunk overlap	$\{0, 32\}$

Table 2: Retrieval configuration grid used in stability experiments.

These settings reflect common engineering trade-offs between recall, latency, and segmentation granularity in production RAG pipelines.

B Representative Stability Gap Examples

Table 3 shows configuration pairs where span-level evidence collapses despite moderate document-level overlap.

Dataset	Config Pair	J_{doc}	J_{span}
ArguAna	base vs c128	0.584	0.052
SciFact	base vs c128	0.419	0.009

Table 3: Illustrative configuration pairs showing span-level collapse despite moderate document overlap.

C Exploratory Evidence–Answer Probe Details

Answers were generated under deterministic decoding for each retrieval configuration. Semantic similarity between answers was computed using Sentence-BERT cosine similarity. Drift rates were defined as the proportion of configuration pairs where similarity fell below predefined thresholds (0.9 and 0.8).

Correlation between span-level stability and answer similarity was evaluated using Spearman rank correlation. Figures 2 and 3 show the probe analysis done for selected configuration perturbations (baseline vs chunk-size change and baseline vs overlap change) to illustrate how retrieval instability propagates to output behavior.

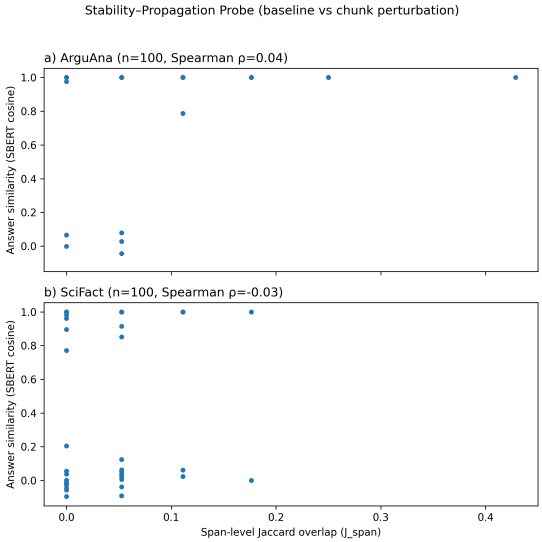

Figure 2: Span-level stability–answer similarity relationship for configuration comparison base vs c128

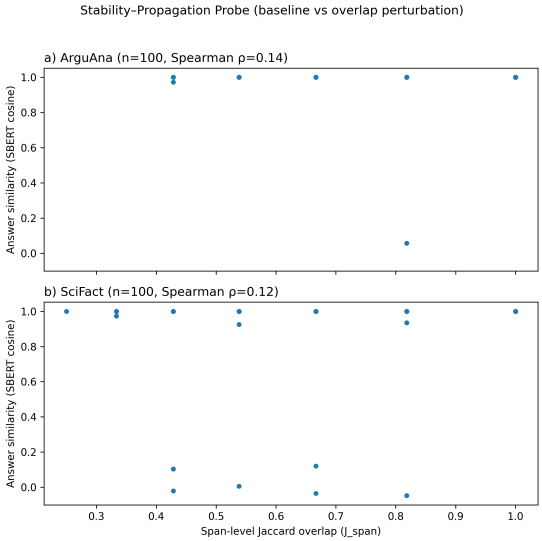

Figure 3: Exploratory scatter plot of span-level evidence stability versus answer similarity for configuration pair base vs o0.