Efficiency vs. Verifiability in Evidence-Aware RAG: Does Prompt Compression Preserve Citation Grounding?

Aiyu Li* **Qian Peng*** **Bin Chen†**
Xi'an Jiaotong-Liverpool University, Suzhou, China
Aiyu.li23@student.xjtlu.edu.cn
Qian.Peng23@student.xjtlu.edu.cn
Bin.Chen02@xjtlu.edu.cn
*These authors contributed equally. †Corresponding author.

Abstract

Retrieval-augmented generation (RAG) is widely used in domain-specific and knowledge-intensive applications, where long prompts increase inference cost and may exceed context limits. Prompt compression is therefore appealing, but existing evaluations focus primarily on answer quality, overlooking whether compressed systems remain faithful to the retrieved evidence. In this paper, we ask: **does compression that preserves answers also preserve grounding?** Using Self-RAG and LLMLingua-2 in a controlled setting, we evaluate compressed RAG on ASQA in terms of both answer correctness and citation grounding. Under increasing compression, answer correctness drops by only 2-4%, whereas grounding drops by 40-50%. This stark divergence shows that answer-only evaluation can substantially overestimate the reliability of compressed RAG in evidence-aware scenarios. We further propose a lightweight hierarchical compression strategy that prioritizes evidence-bearing spans. It recovers nearly all grounding loss while maintaining comparable answer quality. Our results reveal a clear trade-off between efficiency and verifiability, and suggest that compression in RAG should be customized to downstream verification needs rather than treated as a one-size-fits-all efficiency intervention.

1 Introduction

Large language models (LLMs) achieve strong generation quality but often lack domain-specific knowledge and remain prone to hallucination (Shuster et al., 2021). Retrieval-augmented generation (RAG) mitigates these issues by grounding outputs in retrieved external evidence (Lewis et al., 2020; Shuster et al., 2021). In practice, however, effective RAG systems often condition on multiple passages to improve recall and support evidence aggregation (Izacard and Grave, 2021; Izacard et al., 2023), leading to long prompts. Such long contexts increase token-based cost, slow inference, and may exceed context limits (Wang et al., 2024), making context reduction techniques such as selection, pruning, and compression increasingly important for real-world deployments (Pan et al., 2024; Verma, 2024).

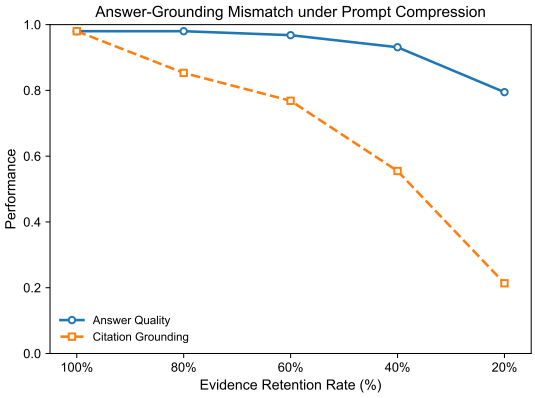

Figure 1: Prompt compression introduces a mismatch between answer quality and evidence grounding. As evidence is increasingly compressed, answer-level performance remains relatively stable, while citation-level grounding deteriorates significantly.

This efficiency challenge is especially relevant in domain-specific and evidence-intensive applications, where retrieved contexts are lengthy, heterogeneous, and difficult to fit within a budget. At the same time, many such applications require not only correct answers but also transparent evidence attribution (Yue et al., 2023; Huang et al., 2024b; Malaviya et al., 2024). However, prompt compression methods are still commonly evaluated using answer-level metrics alone (Pan et al., 2024; Verma, 2024; Min et al., 2023), leaving a critical question open: **does compression that preserves answer quality also preserve evidence grounding?** Figure 1 visually illustrates this trend, showing how compression creates a mismatch between answer quality and citation grounding. This question—which is fundamental to understanding the relia-

bility of compressed RAG—has not been systematically investigated with respect to both answer accuracy and evidence traceability (Li et al., 2024; Saad-Falcon et al., 2024; Es et al., 2024). It is particularly important for **evidence-aware RAG systems** such as Self-RAG (Asai et al., 2024; He et al., 2024), which explicitly rely on retrieved evidence during generation and self-reflection.

We investigate this question in **evidence-grounded long-form question answering** (Malaviya et al., 2024), where responses must synthesize information from multiple passages while remaining verifiable. ASQA is a suitable benchmark for this setting (Stelmakh et al., 2022) because it requires **citation-backed long-form answers** to ambiguous questions and rewards coverage of multiple interpretations. In this scenario, preserving the structure and completeness of supporting evidence can be as important as preserving answer correctness. Consequently, answer-centric evaluation may obscure failures in evidence coverage and citation traceability under compression (Min et al., 2023; Li et al., 2024; Huang et al., 2024b; Saad-Falcon et al., 2024; Es et al., 2024).

To examine this effect, we combine Self-RAG (Asai et al., 2024) with LLMLingua-2 (Pan et al., 2024) in a controlled evaluation on ASQA. We compress only the retrieved evidence passages while keeping the question, instruction, retrieval configuration, generator, and decoding fixed. At multiple evidence retention levels, we evaluate both **answer quality** and **citation grounding**. Our results show a clear answer–grounding mismatch: answer-level performance degrades only mildly, whereas evidence grounding deteriorates much more sharply as compression increases. We further find that a simple citation-aware hierarchical compression strategy can partially recover grounding while maintaining competitive answer quality.

Our contributions are threefold:

1. We provide a controlled study of **prompt compression** in an evidence-aware RAG pipeline using Self-RAG, LLMLingua-2, and ASQA.

2. We show that **answer-preserving compression** can substantially impair **evidence coverage** and **citation grounding**, meaning that answer-only evaluation can overestimate the reliability of compressed RAG in evidence-aware settings.

3. We provide quantitative analysis and a simple **grounding-aware compression strategy** that highlights the trade-off between efficiency and verifiability, and suggests that compression in RAG should be evaluated in a task-aware manner rather than as a one-size-fits-all efficiency intervention.

2 Background and Related Work

RAG improves language generation by conditioning a parametric language model on retrieved external evidence, thereby reducing reliance on memorized knowledge and improving factuality (Guu et al., 2020; Lewis et al., 2020; Ram et al., 2023). A key development is *multi-passage conditioning*: because relevant information is often distributed across multiple sources, RAG systems commonly retrieve and use several passages to support evidence aggregation and cross-document reasoning (Lewis et al., 2020; Izacard and Grave, 2021; Izacard et al., 2023). Representative architectures such as Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) and ATLAS (Izacard et al., 2023) show that conditioning on multiple retrieved passages can substantially improve performance, but also increases input context length (Liu et al., 2024; Jin et al., 2025).

Evidence-intensive RAG needs long retrieved contexts, especially pronounced in long-form question answering, where answers must synthesize complementary evidence across partially redundant sources (Fan et al., 2019; Stelmakh et al., 2022). Benchmarks such as ELI5 (Fan et al., 2019) emphasize multi-sentence explanatory answers grounded in retrieved documents, while ASQA (Stelmakh et al., 2022) further requires citation-backed long-form answers to ambiguous questions covering multiple interpretations. In these settings, practical RAG pipelines often retrieve a relatively large top-k set of passages to reduce recall errors (Karpukhin et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021). Although this improves evidence coverage, it also increases inference cost, latency, and the risk of exceeding context limits, motivating growing interest in context reduction techniques for long-context RAG (Liu et al., 2024; Jin et al., 2025; Jiang et al., 2024; Xu et al., 2024).

Evidence-aware and self-reflective RAG have moved beyond retrieval quality alone and begun to explicitly model evidence use during genera-

tion (Gao et al., 2023a; Shi et al., 2024; Slobodkin et al., 2024; Huang et al., 2024a). Self-RAG (Asai et al., 2024), for example, introduces an adaptive retrieve–generate–critique framework where the model decides when retrieval is needed and uses retrieved evidence to guide self-reflection. Such evidence-aware RAG methods make grounding and attribution more central to the generation process, especially in citation-sensitive applications such as long-form QA and customized knowledge assistants (Gao et al., 2023a; Slobodkin et al., 2024; Huang et al., 2024a; Patel et al., 2024). However, while previous work has improved retrieval, aggregation, and evidence-aware generation, it has paid limited attention to how efficiency-oriented context reduction affects evidence grounding in these settings (Gao et al., 2023a; Shi et al., 2024; Slobodkin et al., 2024; Huang et al., 2024a).

Prompt compression and context pruning aim to reduce the token length of LLM input while preserving downstream task performance under limited budgets (Li et al., 2023; Jiang et al., 2023; Pan et al., 2024; Jiang et al., 2024; Xu et al., 2024; Jha et al., 2024). These methods are motivated by the observation that prompts often contain substantial redundancy, allowing parts of the input to be removed or condensed with limited effect on answer quality. For example, Selective Context (Li et al., 2023) removes low-information tokens based on surprisal estimates from a smaller language model, while LLMLingua-2 (Pan et al., 2024) formulates prompt compression as a token classification problem and trains a lightweight compressor through data distillation. Other approaches learn compact representations, such as "gist" tokens, to summarize long inputs into shorter prompts (Mu et al., 2023). Despite their methodological differences, these approaches are typically optimized and evaluated with respect to *efficiency* (e.g., token reduction, latency, memory usage) and *answer-level performance* (e.g., Exact Match (EM), accuracy, or ROUGE) (Jiang et al., 2023, 2024; Xu et al., 2024).

This evaluation focus leaves a gap for settings where *which evidence is preserved* matters as much as whether the final answer remains correct. Recent work has started to examine information retention and grounding under prompt compression more directly (Łajewska et al., 2025), but the interaction between compression and retrieval-augmented generation remains relatively underexplored (Xu et al., 2024; Jin et al., 2025). In RAG, a compressed context may still produce a correct answer while failing to preserve the specific supporting passages needed for verification—for example, because the model relies more heavily on parametric memory or on incomplete retained evidence (Zhang et al., 2024; Xu et al., 2025).

This limitation is particularly important in citation-grounded RAG, where evidence coverage and traceability are part of the task objective rather than optional properties of the output (Gao et al., 2023b; Petroni et al., 2021; Huang et al., 2024a; Slobodkin et al., 2024; Patel et al., 2024). However, most prior compression studies do not systematically evaluate evidence-level outcomes such as citation precision, citation recall, or support coverage in multi-document generation settings (Zhang et al., 2024; Xu et al., 2025). Our work addresses this gap by studying prompt compression inside an evidence-aware RAG pipeline and evaluating not only answer quality but also citation-level grounding under controlled evidence retention.

3 Approach

We study how retrieved-context compression affects evidence-grounded long-form question answering in RAG. Our key design principle is to intervene only on the *retrieved evidence context* while keeping all other components of the pipeline fixed as illustrated in Figure 2. Concretely, across experimental conditions, the question, task instruction, retriever, generator, and decoding configuration remain unchanged; only the retrieved passages supplied to the generator are modified through compression. This controlled design allows us to examine whether compression that preserves answer quality also preserves the evidence needed for citation grounding.

3.1 Retrieved-Context Compression

We apply compression exclusively to retrieved document passages before generation. The compressed passages are then passed to the RAG backbone as evidence inputs, while all non-evidence inputs are preserved verbatim. Restricting the intervention to the evidence branch isolates the part of the pipeline most directly tied to grounding, verifiability, and citation support.

To instantiate this intervention, we use LLMLingua-2 (Pan et al., 2024), a task-agnostic prompt compression method designed to reduce input length while retaining information use-

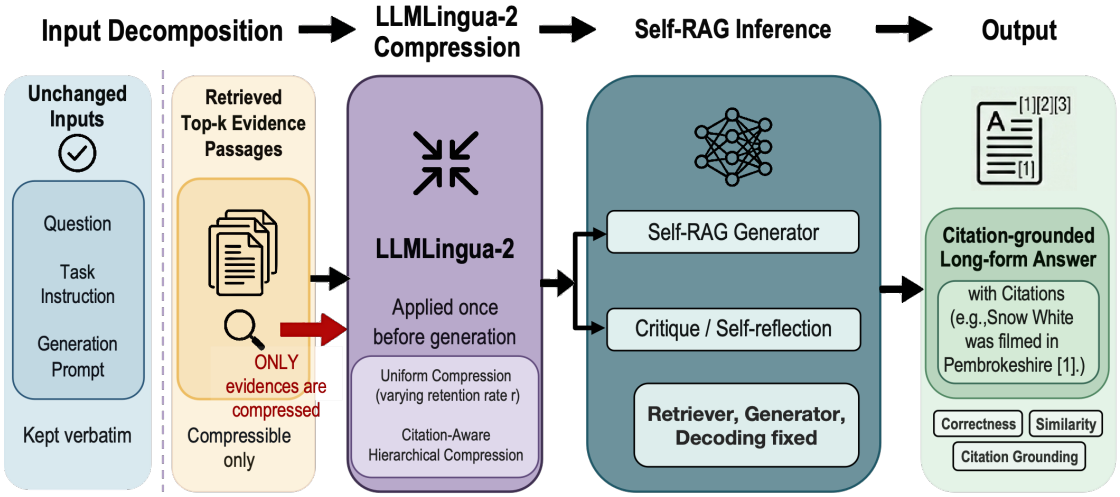

Figure 2: Controlled compression pipeline. Only the retrieved passages are compressed (using uniform or citation-aware hierarchical strategies) before being fed to the Self-RAG generator. All other components (question, instruction, retriever, decoding) remain fixed. The output is a citation-grounded long-form answer, evaluated on answer correctness, distributional similarity, and citation grounding.

ful for downstream generation. In our setting, LLMLingua-2 is used only to compress retrieved evidence passages prior to answer generation.

3.2 Compression Strategies

We compare three retrieved-context conditions.

No Compression. All retrieved passages are used in their original form. This setting serves as the reference condition for both answer quality and citation grounding.

Uniform Compression. We apply a uniform **retention rate** ρ to all retrieved passages. Experiments are conducted with $\rho \in \{0.20, 0.40, 0.60, 0.80, 1.00\}$, where $\rho = 1.00$ corresponds to the uncompressed setting. This strategy provides a controlled way to analyze how progressively reducing retained evidence affects answer quality and citation behavior.

Citation-Aware Hierarchical Compression. Beyond uniform compression, we also study a simple structure-aware strategy intended to better preserve potentially citable evidence. Retrieved passages are divided into a small set of highly relevant passages and the remaining passages, and different retention rates are applied to these two groups. Highly relevant passages are preserved with a higher retention rate, while the remaining passages are compressed more aggressively. In addition, a fixed number of high-priority sentences are retained within each passage to preserve local evidence structure.

In our final configuration, the most relevant passages are kept uncompressed with retention rate $h_i = 1.0$, while the remaining passages are compressed with retention rate $h_o = 0.60$. We additionally retain four sentences per passage. Unless otherwise stated, passage relevance is determined by retriever ranking, and sentence retention does not use gold citations or reference answers. This strategy is intended to balance token efficiency with preservation of the evidence needed for citation-grounded long-form generation.

4 Experimental Setup

We evaluate the effect of retrieved-context compression under controlled conditions using a citation-grounded long-form QA benchmark and a fixed RAG backbone. This section describes the dataset, model configuration, implementation details, and evaluation metrics.

4.1 Dataset: ASQA

We conduct experiments on ASQA, a benchmark for citation-grounded long-form question answering (Stelmakh et al., 2022). ASQA consists of ambiguous factoid questions where a satisfactory response must cover multiple valid interpretations, synthesize information from multiple sources, and provide citations that support verifiable claims. Figure 4 in Appendix A shows the structure of a typical

instance.

ASQA is particularly suitable for our study for three reasons. First, it requires **comprehensive answer coverage** rather than a single short answer. Second, it depends on **multi-document evidence aggregation**, since relevant information is often distributed across partially redundant passages. Third, it evaluates **claim-level verifiability** through explicit citations. These properties make ASQA a natural testbed for studying whether retrieved-context compression preserves not only answer quality but also the evidence needed for grounding.

Following the standard benchmark protocol, we report results on the official ASQA development split, comprising 948 instances. Unless otherwise stated, all model variants are evaluated on the same split with the same retrieval and generation settings.

4.2 Backbone Model: Self-RAG

Our backbone is Self-RAG, a self-reflective retrieval-augmented generation framework that explicitly models evidence use during inference (Asai et al., 2024). Self-RAG conditions generation on retrieved passages and produces long-form outputs with citation-related behaviors, making it a suitable testbed for studying how changes to retrieved context affect both answer quality and grounding.

We use Self-RAG as a fixed backbone throughout all experiments. Specifically, we keep the retriever, number of retrieved passages, generation model, and decoding configuration unchanged across all compression conditions. This design ensures that observed differences can be attributed to retrieved-context compression rather than to changes elsewhere in the pipeline.

4.3 Implementation Details

We use the released Self-RAG model, `selfrag/selfrag_llama2_7b`, with its associated retrieval pipeline. For each query, we retrieve the top-$k = 5$ passages and pass them to the generator as evidence inputs. Unless otherwise stated, the same retrieved set is used across all compression conditions for a given query. All decoding parameters are held fixed across settings. For retrieved-context compression, we use LLMLingua-2 with `microsoft/llmlingua-2-bert-base-multilingual-cased-meetingbank`. Compression is applied once before generation. Each retrieved passage is compressed independently of its target retention rate. In the hierarchical setting, the highly relevant passages are defined as the top-$m = 3$ passages according to retriever rank, while the remaining passages are compressed with a lower retention rate. When sentence preservation is used, sentence priority is determined by lexical overlap with the input question or instruction, with an additional bonus for sentences containing digits, without using reference answers or gold citations.

Unless otherwise noted, all reported results are obtained with the same implementation, prompt template, and random seed configuration across conditions.

4.4 Evaluation Metrics

We evaluate compression effects along three complementary dimensions: **Answer Correctness**, **Distributional Similarity**, and **Evidence Grounding**. Table 3 (see Appendix B) summarizes the metrics used in this section. Our main focus is evidence grounding, measured by `citation_rec` and `citation_prec`, which quantify whether generated answers remain verifiable through citations after evidence compression.

Evidence grounding metrics (main focus). Let G denote the set of gold-supported citation items and P the set of citations produced by the system. Let $C(P, G)$ be the number of correct citations under the evaluation protocol. We compute:

$$\begin{aligned} \texttt{citation\_rec} &= \frac{C(P,G)}{|G|}, \\ \texttt{citation\_prec} &= \frac{C(P,G)}{|P|}. \end{aligned} \tag{1}$$

Comparison Protocol. Under this evaluation framework, we compare (i) the no-compression baseline, (ii) uniform compression at multiple retention rates, and (iii) citation-aware hierarchical compression. All conditions use the same retrieval and generation pipeline, enabling a controlled comparison of the trade-off between efficiency, answer quality, and citation grounding as retrieved evidence is progressively reduced.

5 Main Results

We now present the main empirical results of applying retrieved-evidence compression to Self-RAG on long-form question answering. Across all conditions, we keep retrieval and generation fixed and vary only the *evidence retention rate* applied to retrieved passages. A retention rate of r means that $r \times 100\%$ of the original retrieved evidence

is preserved, with $r = 1.00$ corresponding to the no-compression setting.

Our results show a consistent separation between answer quality and evidence grounding under compression. As retention decreases, answer-level correctness degrades only mildly, distributional similarity shifts more noticeably, and citation grounding deteriorates sharply. This pattern indicates that compressed evidence may still be sufficient to produce plausible answers while no longer preserving the supporting spans required for reliable citation.

5.1 Overall Performance Comparison

Main tables. Table 1 summarizes overall performance across evidence retention rates, grouped by answer metrics, distributional similarity, and grounding metrics. We use the 1.00 condition as the no-compression reference. For clarity, Table 2 reports the citation-aware hierarchical strategy separately. A qualitative case study showing how evidence compression changes the supporting spans available for citation is discussed in Appendix C.

Overall trend. Table 1 reveals a clear answer–grounding mismatch under compression. Relative to the no-compression setting (1.00), answer-level metrics remain fairly stable as evidence is reduced, whereas grounding metrics decline much more sharply. Distributional similarity, measured by `Mauve`, exhibits intermediate sensitivity.

This contrast is especially clear under aggressive compression. From retention 1.00 to 0.20, `RougeLsum` decreases from 35.34 to 33.15 (-2.19), and `QA-F1` decreases from 23.35 to 19.08 (-4.27). In comparison, `Mauve` drops from 73.88 to 59.96 (-13.92), while grounding deteriorates much more severely: `citation_rec` falls from 50.10 to 10.95 (-39.15), and `citation_prec` falls from 63.71 to 12.13 (-51.58). These results show that answer plausibility is substantially more robust to compression than citation verifiability.

Retention gradient. Performance generally improves as more evidence is retained. However, the rate of recovery differs by metric group: answer-level metrics recover quickly and remain relatively stable even at moderate compression, whereas grounding metrics recover more gradually and remain substantially below the no-compression reference until high retention levels. This suggests that preserving enough evidence for answer generation is easier than preserving the fine-grained support needed for accurate citation.

Citation-aware hierarchical compression. Table 2 shows that citation-aware hierarchical compression remains close to the no-compression reference on both answer quality and grounding. `RougeLsum` slightly improves from 35.34 to 35.75, `citation_rec` changes from 50.10 to 49.80, and `citation_prec` changes from 63.71 to 62.55. At the same time, the average length decreases from 30.12 to 29.29, indicating a modest reduction in retained content. These results suggest that structure-aware retention can preserve citation quality under a mild compression budget, although the current configuration prioritizes grounding preservation more than aggressive token reduction.

5.2 Metric Group Analysis

To better understand the compression effect, we analyze results by metric group.

Answer correctness: relatively robust. We first examine overlap-based metrics (`str_em`, `str_hit`, `RougeLsum`) and QA-based metrics (`QA-EM`, `QA-F1`, `QA-Hit`). Across retention rates, these metrics change comparatively little relative to the grounding metrics. Even under strong compression, the model often preserves enough high-level semantic cues to generate answers that remain partially correct and reasonably close to the references.

This robustness suggests that long-form answer generation can tolerate substantial loss of retrieved detail so long as core entities and salient facts are still available. However, such answer-level stability should not be interpreted as evidence that grounding is preserved.

Distributional similarity: moderately sensitive. We next consider `Mauve`, which captures how closely the distribution of generated responses matches that of the references. Compared with answer correctness, `Mauve` is more sensitive to aggressive compression: lower retention leads to a clear decline in distributional similarity. This suggests that compression removes not only factual support but also stylistic, contextual, and long-tail content signals that help generated answers resemble reference long-form responses.

As retention increases, `Mauve` approaches the no-compression reference, but its degradation remains more pronounced than that of the answer-level metrics. This places distributional shift between answer robustness and grounding failure in the overall compression trade-off.

Table 1: Self-RAG performance across evidence retention rates. Answer metrics (overlap and QA), distributional similarity (Mauve), and citation grounding (recall/precision) are reported. Higher is better.

(a) Answer

Ret.	len	str_em	str_hit	R-L	QA-EM	QA-F1	QA-Hit
0.20	23.39	23.80	5.70	33.15	12.92	19.08	1.37
0.40	23.92	27.02	7.07	33.76	16.52	22.36	2.00
0.60	25.54	27.63	6.33	34.12	16.69	23.24	2.22
0.80	27.87	29.32	7.38	35.01	17.79	23.53	2.11
1.00	30.12	29.69	8.54	35.34	17.70	23.35	2.74

(b) Distribution

Ret.	Mauve
0.20	59.96
0.40	63.92
0.60	66.91
0.80	73.49
1.00	73.88

(c) **Evidence**

Ret.	**cRec**	**cPrec**
0.20	10.95	12.13
0.40	28.46	32.38
0.60	39.41	46.73
0.80	43.76	54.96
1.00	50.10	63.71

Table 2: Citation-aware hierarchical compression (hi=1.00), reported separately.

Setting	len	R-L	**cRec**	**cPrec**
No compression (1.00)	30.12	35.34	50.10	63.71
Citation-Aware (hi=1.00)	29.29	35.75	49.80	62.55

Note: R-L=RougeLsum; cRec=citation_rec; cPrec=citation_prec.

Evidence grounding: sharply degraded. Grounding is the most compression-sensitive aspect of performance. Both citation_rec and citation_prec decline substantially as retention decreases, with much larger absolute drops than those observed for answer-level metrics. This indicates that compression disproportionately removes or fragments the specific spans needed to support generated claims.

The resulting failure mode is not simply lower answer quality, but weaker verifiability: the model can still produce plausible statements while losing the evidence coverage and citation precision required to justify them. In other words, answer correctness and evidence grounding degrade at markedly different rates under compression.

Implications of citation-aware compression. The citation-aware hierarchical setting provides an initial indication that this mismatch is not inevitable. By preserving the highest-priority passages and maintaining local sentence structure, it keeps grounding metrics close to the no-compression reference while preserving answer quality. At the same time, because the current setting uses only a mild compression budget, it should be viewed as a proof of concept rather than a fully optimized efficiency–grounding trade-off. A more systematic exploration of grounding-aware compression policies remains an important direction for future work.

Takeaway. Overall, the results reveal a consistent three-level effect of evidence compression: *answer correctness is relatively robust*, *distributional similarity is moderately sensitive*, and *evidence grounding degrades sharply*. This answer–grounding mismatch highlights the central limitation of answer-only evaluation in compressed RAG systems and motivates compression strategies that explicitly preserve verifiable supporting evidence.

5.3 Prefix Truncation Baseline

To test whether the answer–grounding mismatch observed above is specific to the learned compressor, we further compare against a simple **prefix truncation** baseline. In this setting, each retrieved passage is truncated by keeping only the first ρ fraction of its tokens before being passed to the original Self-RAG pipeline, while retrieval, generation, and decoding are kept fixed. This provides a simple non-learned control for testing whether the degradation pattern is compressor-specific or reflects a broader consequence of reducing retrieved evidence.

Overall trend. Figure 3 shows that the truncation baseline reproduces the same qualitative pattern as the learned compression results. As the retention rate decreases, both answer-level performance and citation grounding decline, but grounding metrics remain substantially more compression-sensitive than answer-level metrics. This suggests that the answer–grounding mismatch is not unique to LLMLingua-2, but persists even under a very simple positional compression rule.

Answer quality: gradual degradation. Under LLMLingua-2, QA-F1 decreases from 23.35 at the no-compression reference (Ret.=1.00) to 23.53 at Ret.=0.80 (+0.18), 23.24 at Ret.=0.60

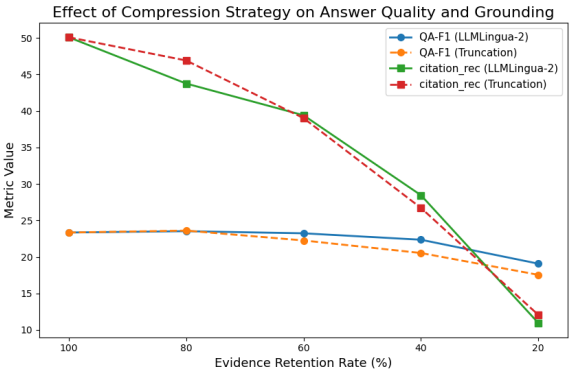

Figure 3: Comparison of learned compression (LLMLingua-2) and prefix truncation across evidence retention rates. We compare the learned compressor (LLMLingua-2) with a simple prefix truncation baseline across evidence retention rates. Both methods exhibit the same qualitative pattern: answer-level performance degrades gradually, while citation grounding collapses sharply under aggressive compression.

(−0.11), 22.36 at Ret.=0.40 (−0.99), and 19.08 at Ret.=0.20 (−4.27). The prefix truncation baseline follows a similar but slightly steeper trajectory, with QA-F1 changing from 23.35 at Ret.=1.00 to 23.61 at Ret.=0.80 (+0.26), 22.26 at Ret.=0.60 (−1.09), 20.54 at Ret.=0.40 (−2.81), and 17.55 at Ret.=0.20 (−5.80). These trends indicate that answer-level quality remains partially preserved even under substantial evidence reduction.

Evidence grounding: much sharper degradation. The contrast is much stronger for citation grounding. Under LLMLingua-2, citation_rec falls from 50.10 at Ret.=1.00 to 43.76 at Ret.=0.80 (−6.34), 39.41 at Ret.=0.60 (−10.69), 28.46 at Ret.=0.40 (−21.64), and 10.95 at Ret.=0.20 (−39.15). The prefix truncation baseline exhibits the same failure mode: citation_rec decreases from 50.10 at Ret.=1.00 to 46.92 at Ret.=0.80 (−3.18), 39.02 at Ret.=0.60 (−11.08), 26.72 at Ret.=0.40 (−23.38), and 12.06 at Ret.=0.20 (−38.04). Thus, while answer-level metrics degrade gradually, grounding coverage collapses much more rapidly as evidence is reduced.

Implication of the baseline. The key implication is that the mismatch between answer quality and grounding preservation is not an artifact of a particular learned compressor. Even a simple prefix truncation baseline—which introduces no learned salience modeling or compression objective—shows the same qualitative separation between relatively stable answer quality and sharply degraded citation grounding. This strengthens our main claim that, in evidence-aware RAG, preserving enough information to generate a plausible answer is easier than preserving the structured supporting evidence needed for reliable citation grounding.

Takeaway. Overall, the baseline comparison reinforces the main conclusion of this section: reducing retrieved evidence produces a consistent answer–grounding mismatch across compression strategies. Learned compression may affect the exact severity of degradation, but the qualitative pattern remains the same, suggesting that grounding deterioration is a general risk of evidence reduction rather than a method-specific artifact.

6 Conclusion

Prompt compression is increasingly used to reduce RAG costs, but its evaluation often focuses on answer-level metrics. In this work, we study retrieved-evidence compression in an evidence-aware RAG pipeline using Self-RAG, LLMLingua-2, and ASQA. By varying only the evidence retention rate, we isolate how compression affects citation-grounded long-form QA.Our experiments reveal a consistent *answer–grounding mismatch*: answer correctness degrades only mildly under compression, while citation grounding declines sharply (e.g., 2–4 points vs. 40–50 points). This shows that answer-only evaluation can substantially overestimate the reliability of compressed RAG.

We further find that a simple **citation-aware hierarchical compression** strategy preserves grounding under a mild budget while maintaining answer quality, suggesting that grounding-aware retention is a promising direction. More broadly, compression methods for evidence-centric QA should be evaluated not only on answer quality and efficiency, but also on their ability to preserve citation-supporting evidence. We discuss limitations in Appendix D and ethical and societal implications in Appendix E.

Acknowledgments

This study is supported by XJTLU RDF Funded Research Project RDF-24-02-084. We are also grateful to the ACL program chairs and reviewers for their valuable comments, which helped us improve this work.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

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, St. Julians, Malta. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393, Miami, Florida, USA. Association for Computational Linguistics.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024a. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024b. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, Bangkok, Thailand. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Kurt Keutzer, and Amir Gholami. 2024. Characterizing prompt compression methods for long context inference. *arXiv preprint arXiv:2407.08892*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. 2025. Long-context llms meet rag: Overcoming challenges for long inputs in rag. In *The Thirteenth International Conference on Learning Representations*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Weronika Łajewska, Momchil Hardalov, Laura Aina, Neha Anna John, Hang Su, and Lluis Marquez. 2025. Understanding and improving information preservation in prompt compression for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17520–17541, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens. *ArXiv preprint arXiv:2304.08467*.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 963–981, Bangkok, Thailand. Association for Computational Linguistics.

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards improved multi-source attribution for long-form answer generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3906–3919, Mexico City, Mexico. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7339–7353, Bangkok, Thailand. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics.

Sourav Verma. 2024. Contextual compression in retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2409.13385*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17716–17736. Association for Computational Linguistics.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. ALiiCE: Evaluating positional fine-grained citation generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 545–561, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, Tokyo, Japan. Association for Computational Linguistics.

A ASQA

We conduct experiments on ASQA, a benchmark for citation-grounded long-form question answering (Stelmakh et al., 2022). ASQA consists of ambiguous factoid questions, for which a satisfactory response must cover multiple valid interpretations, synthesize information from several sources, and provide citations that support verifiable claims. Figure 4 illustrates the structure of a typical instance.

B Evaluation Metrics

We evaluate compression effects along three complementary dimensions: answer quality, distributional similarity, and evidence grounding.

Answer Quality. We measure the quality of generated long-form answers using [str_em, str_hit, RougeLsum, QA-EM, QA-F1, and QA-Hit], following prior work on ASQA and long-form QA. These metrics assess whether the generated response remains semantically and content-wise close to reference answers under increasing levels of compression.

Distributional Similarity. We use **Mauve** to measure the distributional similarity between generated answers and reference answers.

Citation Grounding. Because ASQA requires explicit evidence attribution, we additionally evaluate citation-related performance using evidence-level grounding metrics such as **citation precision**, **citation recall**. These metrics quantify whether generated claims are supported by retrieved evidence and whether the citations preserve sufficient evidence for verification. They are central to our study because answer-level correctness alone may obscure failures in evidence traceability under compression.

Comparison Protocol. Under this evaluation framework, we compare (i) the no-compression baseline, (ii) uniform compression at multiple retention rates, and (iii) citation-aware hierarchical compression. All conditions use the same retrieval and generation pipeline, enabling a controlled comparison of answer quality and citation grounding as retrieved evidence is progressively reduced.

C Case Study

Motivation. To complement the aggregate trends in Table 1, we present a qualitative case study showing how evidence compression changes the supporting spans available for citation. We compare the no-compression setting (Ret.=1.00) with an aggressive compression setting (Ret.=0.20), while keeping the retrieval and generation pipeline fixed.

Case selection. We select a representative ASQA example (`sample_id: 5992104053523265225`) in which the full retrieved evidence contains concrete, sentence-level support for filming locations and dates, whereas the aggressively compressed

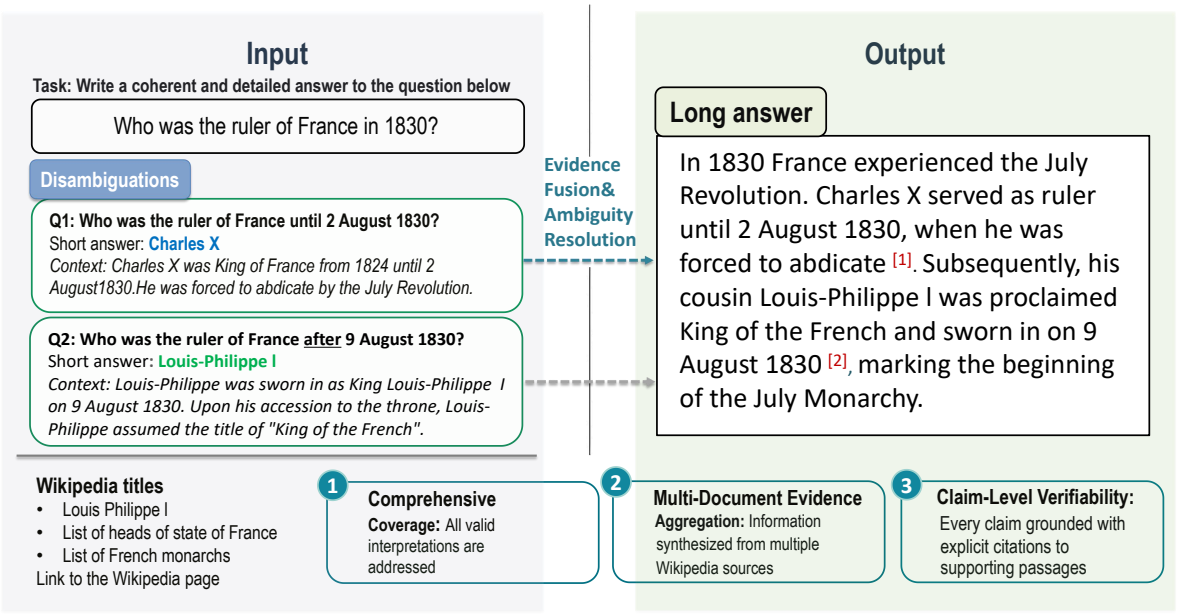

Figure 4: Adapted illustration of an ASQA instance based on the example and annotation interface in Stelmakh et al. (2022). An under-specified (ambiguous) question is paired with multiple disambiguated QA pairs, and the final long-form answer is expected to integrate evidence from multiple sources while providing explicit citations for claim-level verification. The context snippets reflect the original interface content, while the long-form answer is shown schematically for exposition.

evidence becomes fragmentary and keyword-like. This case illustrates how answer plausibility can be preserved even when citation support becomes less precise.

Observation. Table 4 shows that under Ret.=1.00, the retrieved passage preserves complete supporting spans that directly justify fine-grained claims such as location and filming dates. Under Ret.=0.20, many of these details remain only as isolated entities or keywords, with much of the local sentence structure removed. As a result, the model can still generate a plausible answer, but the citation becomes less informative and less precise for verification.

Takeaway. This example mirrors the quantitative pattern observed above: aggressive compression may retain enough surface cues to support answer generation, yet fail to preserve the structured evidence needed for reliable citation grounding. It therefore illustrates the core tension of compressed RAG in evidence-centric long-form QA: plausible answers can survive even when verifiable support does not.

D Limitations

Our study has several limitations. First, experiments are conducted with a single evidence-aware RAG backbone (Self-RAG) and a single compression method (LLMLingua-2), so the observed answer–grounding mismatch may not transfer uniformly to other architectures or compressors. Second, evaluation is limited to ASQA, a citation-grounded long-form QA benchmark; although this makes it well suited to our research question, further validation is needed on other evidence-intensive tasks. Third, our citation-aware hierarchical compression setting is only an initial proof of concept under a relatively mild compression budget, rather than a fully optimized grounding-aware compression policy.

E Ethical and Societal Implications

The deployment of compressed evidence-aware RAG systems raises important ethical and societal concerns, particularly regarding reliability, verifiability, and the risk of misplaced user trust. Our findings show that prompt compression can preserve answer-level performance while substantially degrading citation grounding, which means a system may still produce plausible responses even

Table 3: Metric definitions used in §5 (higher is better).

Group	Metric	Meaning (high-level)
Answer (overlap)	`str_em`	Exact-match style score between the generated long answer and reference (normalized).
Answer (overlap)	`str_hit`	Hit-style score indicating whether key reference answer content is covered.
Answer (overlap)	`RougeLsum`	ROUGE-L overlap with the reference long answer.
Answer (QA)	`QA-EM`	Exact match on QA sub-questions (averaged).
Answer (QA)	`QA-F1`	Token-level F1 on QA sub-questions (averaged).
Answer (QA)	`QA-Hit`	Hit rate indicating whether QA-relevant facts are covered.
Distribution	`Mauve`	Distributional similarity between generations and references.
Evidence (main)	`citation_rec`	Coverage of required/supporting citations (recall).
Evidence (main)	`citation_prec`	Correctness of produced citations (precision).

when the supporting evidence becomes incomplete or difficult to verify. This risk is especially important in customized knowledge assistants and other evidence-sensitive settings, where users may rely not only on the final answer but also on the traceability of its supporting sources. We therefore emphasize that efficiency-oriented interventions in RAG should not be evaluated solely by answer quality, but also by whether they preserve verifiable evidence for downstream use. At the same time, prompt compression may reduce token usage and computational cost, which can improve accessibility and lower environmental burden when applied responsibly. In this work, we study these issues in a controlled research setting using a public benchmark, and our results are intended to improve understanding of the trade-off between efficiency and verifiability rather than to justify immediate deployment in high-stakes domains. Systems based on compressed evidence-aware RAG should therefore be used with appropriate human oversight, especially in settings where incomplete grounding could lead to harmful or overconfident decisions.

Table 4: ASQA case study showing how evidence compression changes the *supporting spans* available for citation. Ret.=1.00 preserves sentence-level evidence, while Ret.=0.20 collapses it into sparse keywords, weakening verifiability even when the answer remains plausible.

	No compression (Ret.=1.00)	**Aggressive compression (Ret.=0.20)**
Question	**Q:** Where was *Snow White and the Huntsman* filmed?	
Retrieved evidence	**Evidence excerpt (doc: *Snow White and the Huntsman*):** "Photography took place in the United Kingdom. The beach scenes were predominantly filmed in **Pembrokeshire**, on the **Marloes Sands** beach near the village of Marloes **between September 26 and 29, 2011**. . . . A computer-generated castle was set on nearby **Gateholm** island. . . . The film used academic consultants from the **University of Chichester**."	**Compressed evidence excerpt (same doc, Ret.=0.20):** "United Kingdom Pembrokeshire Marloes Sands September 26 29 2011 castle Gateholm field ramp University Chichester"
Generated answer	**Answer excerpt:** "The beach scenes were filmed in Pembrokeshire, on Marloes Sands near the village of Marloes (Sep. 26–29, 2011) . . . " **Citation behavior:** sentence-level evidence directly supports fine-grained details (location + date).	**Answer excerpt:** "*Snow White and the Huntsman* was filmed in the United Kingdom in Pembrokeshire, Marloes Sands, and Chichester University [1]." **Citation behavior:** evidence becomes keyword-like; it is harder to ground specific claims to precise supporting spans.
Observation	**What this case shows.** Under Ret.=1.00, the model can (in principle) attach citations to complete supporting spans that contain the *exact* filming details (locations, date range, and related context). Under Ret.=0.20, the retained evidence largely preserves *entities* but loses sentence structure and surrounding context, making citations less informative and less precise for verification (consistent with the larger drop in `cRec/cPrec` than in answer-level metrics).	