Stress Testing Factual Consistency Metrics for Long-Document Summarization

Anonymous ACL submission

Abstract

Evaluating the factual consistency of abstractive text summarization remains a significant challenge, particularly for long documents, where conventional metrics struggle with input length limitations and long-range dependencies. In this work, we systematically evaluate the reliability of six widely used reference-free factuality metrics, originally proposed for short-form summarization, in the long-document setting. We probe metric robustness through seven factuality-preserving perturbations applied to summaries, namely paraphrasing, simplification, synonym replacement, logically equivalent negations, vocabulary reduction, compression, and source text insertion, and further analyze their sensitivity to retrieval context and claim information density. Across three long-form benchmark datasets spanning science fiction, legal, and scientific domains, our results reveal that existing short-form metrics produce inconsistent scores for semantically equivalent summaries and exhibit declining reliability for information-dense claims whose content is semantically similar to many parts of the source document. While expanding the retrieval context improves stability in some domains, no metric consistently maintains factual alignment under long-context conditions. Finally, our results highlight concrete directions for improving factuality evaluation, including multi-span reasoning, context-aware calibration, and training on meaning-preserving variations to enhance robustness in long-form summarization.[1]

1 Introduction

Abstractive summarization has seen rapid advances with the advent of large language models (LLMs), but ensuring that generated summaries faithfully reflect the source content remains a persistent challenge (Laban et al., 2024; Wright et al., 2025). Summaries that read fluently can nonetheless introduce hallucinated details or omit critical facts (Belém et al., 2025a), undermining their reliability for downstream tasks in domains such as medicine, law, and scientific review (Asai et al., 2024). Traditional evaluation measures like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), which rely on n-gram overlap with reference summaries, are useful for measuring surface similarity but fail to capture factual consistency, since two summaries can overlap heavily in wording while still differ in correctness (Maynez et al., 2020). This gap has motivated the development of reference-free factuality evaluation metrics, which assess whether the statements in a summary are supported by the source document itself rather than by comparison with a human-written reference, using techniques such as question answering (Wang et al., 2020; Scialom et al., 2021; Fabbri et al., 2022), natural language inference (Laban et al., 2022; Chen and Eger, 2023; Zha et al., 2023), or LLM-based scoring (Liu et al., 2023; Fu et al., 2024).

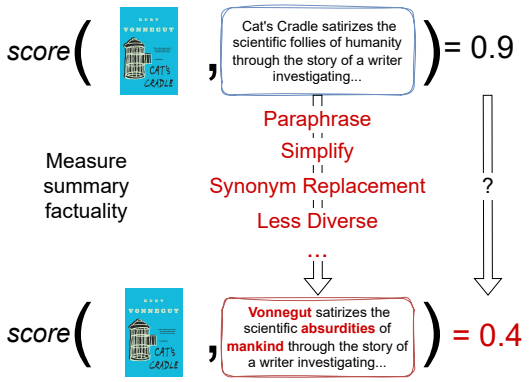

Figure 1: We aim to see how robust summary factuality metrics are for long and multi-document setups by applying meaning-preserving perturbations and comparing metric scores before and after these edits.

While such metrics have demonstrated promise on short-document datasets, their scalability to

[1]We release our experimental code and all perturbed summary outputs as zip files in the submission.

long-form summarization is likely to be hindered by challenges unique to long context lengths (Russak et al., 2024; Sarthi et al., 2024; Edge et al., 2024). Important details may be dispersed across hundreds or thousands of tokens and thus overlooked by metrics that process only truncated inputs (Laban et al., 2024); multi-document summaries must reconcile diverse writing styles and potentially conflicting information (Asai et al., 2024); and the reference-free setting deprives evaluators of gold annotations, necessitating robust intrinsic evaluation protocols. Our goal is to systematically benchmark widely used factual consistency metrics under these long-document conditions, in order to reveal their robustness and limitations.

Building on this foundation, we apply a stress-testing methodology established in previous work (Ramprasad and Wallace, 2024) to six widely-used reference-free metrics (§ 4) across seven factuality-preserving perturbations reflecting realistic long-form summarization phenomena. These perturbations (§ 3.1) broadly cover paraphrasing operations, simplification of complex constructions, synonym replacement, reduced lexical diversity, logically equivalent negations, further compression of the summary, and insertion of unrelated source sentences. They are designed to challenge metrics to distinguish genuine factual consistency from superficial cues and to maintain robustness in the face of lexical and structural variation. On top of this, we investigate the impact of long-document specific phenomena, including retrieval length and evidence diffusion (Goldman et al., 2024).

Through extensive experiments using six evaluation metrics across three benchmark datasets in science fiction, legal, and scientific domains, we expose significant weaknesses in existing metrics, such as inconsistent scoring across semantically equivalent summaries. Our analysis reveals that current factuality metrics vary widely in robustness to meaning-preserving edits, with some highly sensitive to surface changes while others remain more stable. Most metrics benefit from broader retrieval context windows, though with notable domain-specific variation. We also find that metric reliability decreases for information-dense claims that overlap semantically with large portions of the source document, suggesting that current metrics struggle with compressed or globally entangled content. These insights point toward improving evaluation consistency by developing metrics that integrate multi-span reasoning and context-aware calibration.

2 Factual Consistency in Abstractive Summarization

Factual consistency refers to whether a summary accurately reflects the content of the source document. While abstractive summarization systems have become increasingly fluent, they often produce factually incorrect or hallucinated statements (Huang et al., 2025). These hallucinations can range from minor misstatements to major distortions, particularly when the input is lengthy and semantically dense (Belém et al., 2025b). A number of techniques have been proposed to mitigate this (Gao et al., 2023; Qiu et al., 2023; Zhang et al., 2024; Mündler et al., 2024; Wang et al., 2024). Despite these efforts, most prior work has focused on short-form summarization settings, where the task is relatively controlled. In contrast, long-form summarization requires condensing information spread across thousands of tokens, making reliable factuality evaluation substantially more difficult, especially without reference summaries or human annotations.

2.1 Factuality Evaluation Metrics

To overcome the limitations of traditional reference-based metrics (Maynez et al., 2020), a range of reference-free metrics have emerged that assess the alignment between the summary and source directly. Entailment-based approaches such as FactCC (Kryscinski et al., 2020) and SummaC (Laban et al., 2022) use NLI models to judge whether summary sentences are supported by the source. QA-based methods like QAGS (Wang et al., 2020) and QuestEval (Scialom et al., 2021) evaluate factuality by generating and answering questions derived from the summary. Generation-based metrics, including `BARTScore` (Yuan et al., 2021) and T5Score (Trainin and Abend, 2025), estimate the likelihood of the summary given the source using pretrained sequence-to-sequence models. More recent tools like `AlignScore` (Zha et al., 2023), `MiniCheck` (Tang et al., 2024), and `UniEval` (Zhong et al., 2022) aim to improve efficiency and generalization across tasks. While effective on short-document benchmarks, most of these metrics assume the full source and summary can be jointly encoded, limiting their utility for long-form inputs. Moreover, recent work shows that these

metrics are often brittle, overreacting to edits like paraphrasing, reordering, or logically equivalent reformulations (Ramprasad and Wallace, 2024). In this work, we study the behavior of six popular factuality metrics in long-document summarization using a retrieval-based scoring framework, and systematically evaluate their robustness to a set of controlled meaning-preserving perturbations.

2.2 Challenges in Long-Document Factuality Evaluation

Evaluating factual consistency in long documents introduces challenges that differ fundamentally from those in short texts. Long inputs often contain information that is dispersed, hierarchically structured, and cross-referential, requiring models to link evidence across distant sections or even multiple documents (Asai et al., 2024). This results in long-range dependencies and positional biases such as the "lost in the middle" effect (Liu et al., 2024). Yet, research into robust factuality metrics for long inputs remains limited. Koh et al. (2023) identified a clear gap in the literature for automatic evaluation methods tailored to long-document summarization. LongSciVerify (Bishop et al., 2024) and LongEval (Krishna et al., 2023) are two of the only available datasets with human factuality annotations in this setting. Chunk-based approaches like SMART (Amplayo et al., 2022), partially address this by sequentially processing document segments, but they remain computationally expensive and often inconsistent. A more scalable alternative is retrieval-based scoring, exemplified by LongDocFACTScore (Bishop et al., 2024), which retrieves top-k relevant source passages for each summary sentence, and computes sentence-level factuality scores that can be aggregated into a global metric. However, it remains unclear whether existing metrics, when used within such frameworks, behave consistently and robustly across varying retrieval configurations. Our study addresses this gap by analyzing these behaviors under controlled conditions and across multiple domains.

2.3 Adversarial Robustness of Metrics

Recent work has evaluated the robustness of factuality metrics by applying controlled perturbations to the summary or source (Goyal and Durrett, 2021; Chen et al., 2021; Gabriel et al., 2021). Ramprasad and Wallace (2024) showed that many metrics are brittle when faced with logically equivalent but lexically altered summaries, with even benign transformations such as reordering or simplification causing large score shifts. However, these evaluations have been primarily limited to short-document settings. In long-form summarization, where token selection, compression, and abstraction are all harder, such brittleness could be magnified. In this work, we probe whether widely used metrics remain reliable under minimal changes when operating over long, realistic inputs. Our results highlight major inconsistencies and call for more robust, semantically aware approaches for evaluating factual consistency.

3 Metric Robustness in Long-Form Summarization

To analyze the robustness of existing factuality metrics in long-form summarization, we evaluate six widely used reference-free metrics (§ 4) that span diverse architectures and scoring paradigms.

3.1 Perturbation Strategies

To evaluate the robustness of factuality metrics in a controlled manner, we apply meaning-preserving perturbations to the original summaries, as done in (Ramprasad and Wallace, 2024) in the short document case. These perturbations are designed to vary the summary's surface form (lexical choices, structure, or style) while preserving its factual consistency with the source document. In principle, a robust factuality metric should be invariant to such benign edits, assigning similar scores to the original and perturbed versions.

We define seven perturbation types, each targeting a different linguistic dimension. These include *Paraphrased*, where the summary is rewritten with alternate phrasings and syntactic structures; *Simplified*, where complex or compound constructions are rewritten into shorter, more readable sentences; *Synonym Replaced*, where content words are substituted with close synonyms to test for lexical invariance. We also generate *Less Diverse* summaries that reduce vocabulary variation, exploring whether metrics implicitly reward stylistic richness. Additional perturbations include *Negated*, which introduces logically equivalent negations to probe sensitivity to syntactic polarity, *Summarized*, which further compresses the summary to test how conciseness is handled, and *Added Source Text*, which inserts a factual sentence directly from the source that is unrelated to the main summary content. All seven perturbed summaries are generated using the

GPT-4o (Hurst et al., 2024) model via the OpenAI API. The detailed prompts used to generate each perturbation are provided in App. A.

While these transformations are meaning-preserving, they pose particular challenges in the long-document setting: summaries must capture information scattered across thousands of tokens, so perturbations that change sentence structure, reduce vocabulary, or alter flow can disrupt long-range dependencies and retrieval alignment. This makes them a rigorous test of whether factuality metrics remain robust. Any significant fluctuation in scores, despite no factual errors being introduced, indicates that a metric is reacting to surface-level edits rather than faithfully assessing factual consistency.

3.2 Retrieval-Based Scoring for Long Documents

Most factuality metrics in existing literature are designed for short inputs and cannot directly process the full content of long documents due to token length limitations. This is particularly problematic in long-form summarization, where summaries may draw on information scattered across multiple sections or even multiple documents. To address this, we follow the retrieval-augmented strategy proposed by Bishop et al. (2024), which enables factuality evaluation at the sentence level without requiring the metric to ingest the entire source document at once.

Let $S = \{s_1, s_2, \ldots, s_m\}$ denote the summary, where each s_j is a sentence, and let $D = \{d_1, d_2, \ldots, d_n\}$ be the set of sentences in the source document. For each summary sentence s_j, we compute a sentence embedding $\mathbf{e}_j$, and similarly obtain embeddings $\{\mathbf{e}_1^D, \ldots, \mathbf{e}_n^D\}$ for the source document using a pre-trained sentence encoder. We compute cosine similarity between s_j and each $d_i \in D$ and retrieve the top-K most similar source sentences. Each retrieved sentence $d_{j,k}$ is then expanded to include the surrounding context within a symmetric window size w, forming a snippet:

$$d_{j,k}^{(w)} = \{d_{j,k-w}, \ldots, d_{j,k}, \ldots, d_{j,k+w}\}. \quad (1)$$

We evaluate the factual consistency of s_j against each of the K context snippets, using any automated metric $\mathcal{M}$, and take the maximum score:

$$\text{score}(s_j) = \max_{k \in \{1,\ldots,K\}} \mathcal{M}(s_j, d_{j,k}^{(w)}). \quad (2)$$

Finally, the summary-level factuality score is computed by averaging these sentence-level scores across all sentences in the summary. In our experiments, we explore how varying w affects metric behavior, shedding light on the sensitivity of different metrics to retrieval context size.

4 Experimental Setup

Metrics We evaluate six reference-free factuality metrics that represent diverse architectures and scoring paradigms. `BARTScore` (Yuan et al., 2021) estimates the log-likelihood of the summary given the source using a pretrained BART model[2], treating factuality as a conditional generation problem. For consistency with other metrics, we exponentiate the log-likelihood scores to obtain normalized values that are directly comparable across metrics. `SummaC-Conv` and `SummaC-ZS` (Laban et al., 2022) represent entailment-based approaches that apply pretrained NLI models to compute consistency between summary and source sentence pairs; the former uses a learned aggregation layer, while the latter relies on zero-shot averaging. `AlignScore` (Zha et al., 2023) leverages contrastive alignment learning across NLI, QA, and summarization tasks and demonstrates strong cross-domain generalization. `UniEval` (Zhong et al., 2022) formulates the evaluation as a multi-dimensional question answering task within a unified T5-based framework, jointly considering factuality, coherence, relevance, and fluency. Finally, `MiniCheck` (Tang et al., 2024) is a lightweight, sentence-level factuality classifier that achieves near GPT-4 performance at a fraction of the cost. We use the `Bespoke-MiniCheck-7B` variant, which ranks highest on the LLM-AggreFact benchmark (Tang et al., 2024), and take advantage of its 32k-token context window to provide full-document context during evaluation.

Datasets We conduct our analysis on three long-document summarization datasets spanning diverse domains: *SQuALITY* (Wang et al., 2022), *LexAbSumm* (T.y.s.s. et al., 2024), and *ScholarQABench* (Asai et al., 2024). *SQuALITY* consists of public domain science fiction stories paired with expert-written summaries that balance narrative abstraction and fine-grained detail, with an average input length of 5,200 tokens. *LexAbSumm* contains legal judgments from the European Court of Human Rights, where summaries are aspect-specific

[2] https://huggingface.co/facebook/bart-large-cnn

and demand precise distillation of dense legal arguments; inputs average 14,357 tokens. *ScholarQABench* is a multi-document benchmark based on open-access computer science papers, where the task is to generate detailed, factual answers to expert-written queries using evidence from multiple documents; inputs average 16,341 tokens.

Experimental Design To evaluate metric robustness, we construct perturbed versions of the summaries from each dataset using the seven meaning-preserving transformations described in § 3.1. These edits allow us to test whether metrics exhibit sensitivity to benign changes. For each summary, original and perturbed, we use the framework (§ 3.2) to retrieve source evidence and compute factuality scores using all six metrics discussed above. Beyond this robustness analysis, we also investigate how retrieval granularity and claim information density influence metric behavior. We vary the evidence window size w from Eq. 1 to test how broader or narrower retrieval contexts affect metric performance. Additionally, we measure the information density of each summary sentence using the mean pairwise cosine similarity between its embedding and all sentences in the source document. High information density indicates claims that semantically overlap with many parts of the source and are therefore harder to verify, while low-density claims correspond to specific statements with localized evidence. This analysis reveals how metrics respond to different levels of semantic compression, providing insight into their sensitivity to claim complexity in long-form summarization.

5 Results & Analysis

We evaluate the robustness and behavior of six reference-free factuality metrics across a range of semantic-preserving perturbations, varying retrieval context windows, and differences in claim information density. Our results are presented in three parts: (1) robustness under perturbation (§ 5.1), (2) sensitivity to retrieval granularity (§ 5.2), and (3) metric sensitivity to claim information density (§ 5.3).

5.1 Robustness Against Perturbations

Figure 2 presents the change in factuality scores when different perturbations are applied to the summaries across three datasets. Each plot shows the difference in score (perturbed minus original) for a given metric, broken down by perturbation and dataset. A robust metric should remain invariant to these semantic-preserving changes. However, we find that all metrics show varying levels of sensitivity.

On *LexAbSumm*, `BARTScore` shows clear negative shifts across nearly all perturbations, while remaining relatively consistent on all other datasets. These consistent declines on *LexAbSumm* indicate that `BARTScore` is highly sensitive to even mild surface-level edits in the legal domain, where long and complex sentence structures and domain-specific jargon likely amplify generation-based instability. `MiniCheck` shows very small changes across all perturbations and datasets However, it struggles with logically equivalent negations, especially in *LexAbSumm*. This may reflect a domain mismatch, as the metric appears less effective in capturing factual consistency in legal texts. `SummaC-Conv` and `SummaC-ZS`, both based on NLI, show moderate and more balanced behavior. They are somewhat affected by *Summarized* and *Negated* summaries, especially on *SQuALITY* and *LexAbSumm*, showing they are not fully invariant to meaning-preserving rewrites. `UniEval` is sensitive to most of the perturbations. However, it consistently fails to handle logically equivalent *Negated* summaries across all datasets, suggesting a lack of sensitivity to logical form. `AlignScore` mostly struggles with the legal domain, showing large score drops in response to *Paraphrased, Negated* and *Summarized* summaries. This suggests difficulty in tracking sentence order and logical consistency in structured, formal texts. While it performs more reliably on *SQuALITY* and *ScholarQABench*, it remains less robust than other metrics overall.

Detailed per-dataset results are provided in App. B. These results list the mean factuality scores for each metric and perturbation type across all datasets.

5.2 Effect of Retrieval Context Window Size

Table 1 reports average factuality scores on the original summaries for each metric and dataset, using window sizes $w = 0, 1, 2$ in Eq. 1. We find that most metrics show consistent improvements as the window size increases, suggesting that they can effectively leverage broader local context when making sentence-level factuality judgments. This pattern is particularly pronounced on *LexAbSumm*, where understanding legal arguments often requires attending to multi-sentence spans. `SummaC-ZS` and `SummaC-Conv` display little sensitivity to larger con-

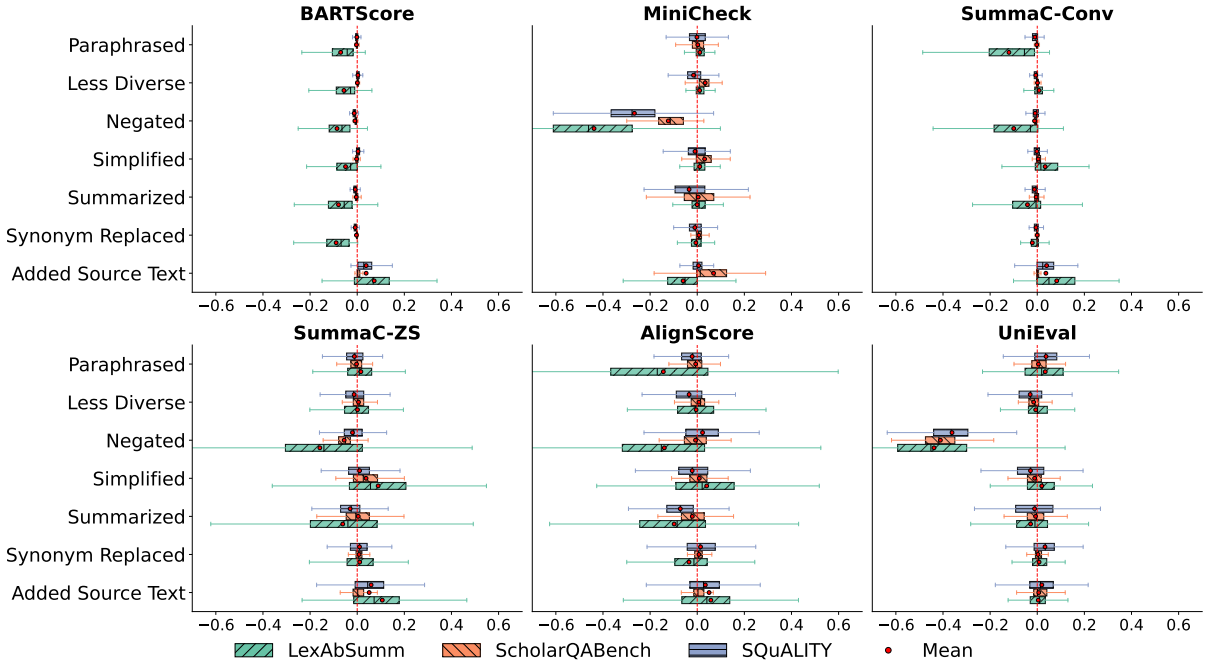

Figure 2: Score change under factuality-preserving perturbations. Boxplots show the difference in factuality score between the perturbed and original summaries, for each metric and perturbation type, across three datasets. The central dot indicates the mean score difference, and the whiskers represent the minimum and maximum values.

Metric	ScholarQABench			SQuALITY			LexAbSumm		
	$w = 0$	$w = 1$	$w = 2$	$w = 0$	$w = 1$	$w = 2$	$w = 0$	$w = 1$	$w = 2$
BARTScore	0.03	0.03	0.02	0.03	0.03	0.03	0.15	0.16	0.16
MiniCheck	0.17	0.15	0.15	0.11	0.15	0.19	0.47	0.53	0.60
SummaC-Conv	0.22	0.23	0.25	0.22	0.23	0.24	0.33	0.33	0.34
SummaC-ZS	0.14	0.18	0.20	0.11	0.13	0.14	0.36	0.38	0.39
AlignScore	0.15	0.21	0.27	0.10	0.18	0.24	0.36	0.52	0.64
UniEval	0.72	0.73	0.74	0.67	0.68	0.70	0.81	0.82	0.84

Table 1: Impact of retrieval context window size w on factuality scores for original summaries. Increasing w expands the snippet around each retrieved sentence.

text windows, implying that their underlying models base judgments on more localized comparisons and are less responsive to extended evidence.

Taken together, these results indicate that retrieval-based scoring can improve factuality assessment in long-document summarization, especially when a broader context is provided. However, NLI-based metrics remain insensitive to increasing context windows.

5.3 Metric Sensitivity to Claim Similarity

To better understand what makes factual consistency evaluation difficult in long documents, we analyze how the semantic density of a claim relates to metric reliability. We hypothesize that claims which are highly similar to many claims in the original document contain more general (and thus compressed) information, and so are harder to fact-check since evidence for them is distributed across multiple parts of the source document.

We approximate this by computing the mean pairwise cosine similarity between each summary sentence s_j (claim) and all n sentences d_i in the source document D,

$$\text{Sim}(s_j, D) = \frac{1}{n} \sum_{i=1}^{n} \cos\left(\mathbf{e}_j, \mathbf{e}_i^D\right), \quad (3)$$

where e_j and e_i^D denote the sentence embeddings of the claim and document sentences, respectively. Higher $\text{Sim}(s_j, D)$ values indicate more information-dense claims whose meaning overlaps broadly with the source, while lower values correspond to specific, localized claims. Claims are then grouped into similarity bins $\mathcal{B}$, and the average

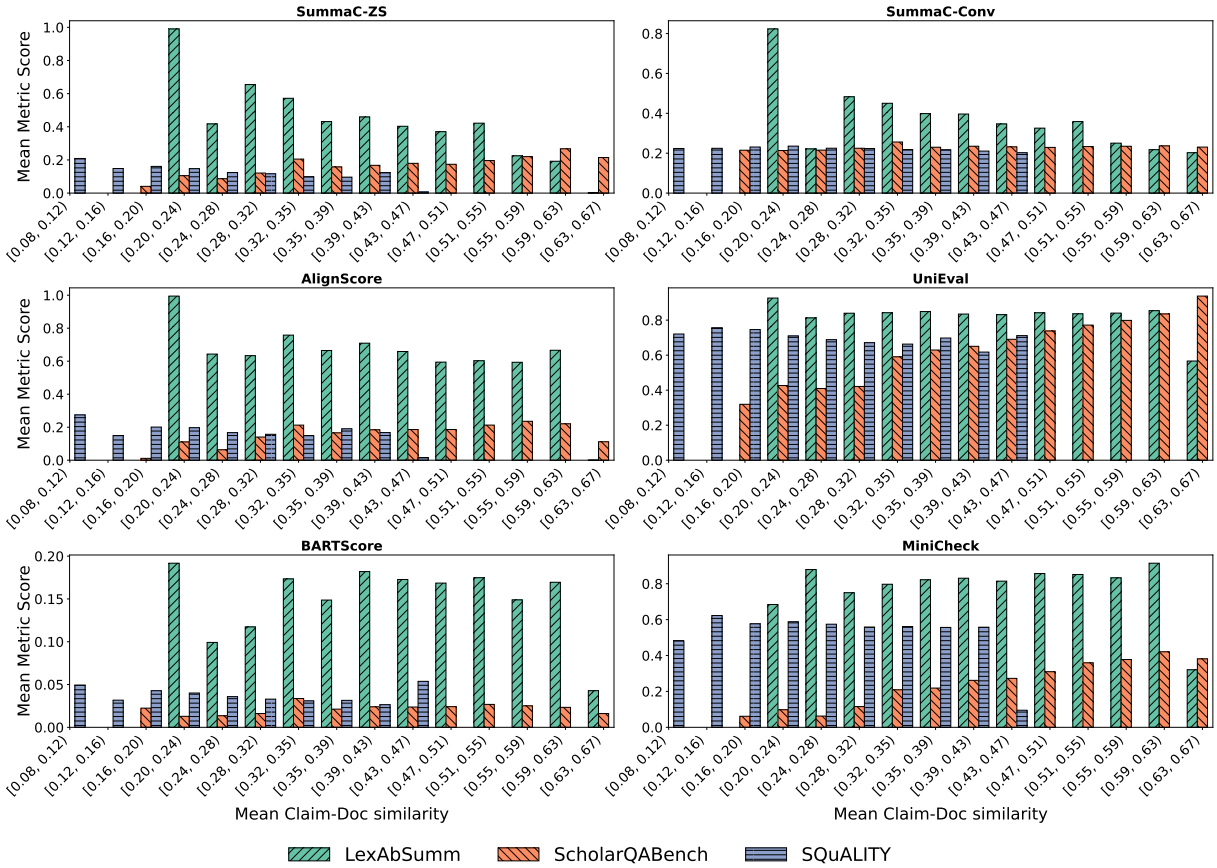

Figure 3: Relationship between claim similarity and average factuality score. Higher similarity values correspond to more information-dense claims whose content overlaps with multiple parts of the source document. Metrics generally assign lower scores to these claims for LexAbSumm and SQuALITY, and higher scores for ScholarQABench, indicating reduced reliability for compressed information.

factuality score for each bin is calculated as

$$\text{Score}_{\text{bin}} = \frac{1}{|\mathcal{B}|} \sum_{s_j \in \mathcal{B}} M(s_j), \quad (4)$$

where $M(s_j)$ is the factuality score for claim s_j.

A few trends emerge between claim similarity and metric sensitivity across all datasets and metrics, as shown in Figure 3. For *LexAbSumm*, metric scores consistently decrease as claim similarity increases, meaning that the more a claim's meaning is entangled with the broader document, the worse the metric becomes at predicting factuality. This is likely because summaries of legal documents may refer to specific aspects of the texts which are easier to fact-check, while more general statements which compress a lot of technical language are more difficult. The same occurs (though less pronounced) for *SQuALITY*, where more general claims are more challenging to fact check. In this case, we are summarizing novels, so more general claims try to compress the story narrative into a compact form, and thus will require retrieving or attending to and reasoning over disparate pieces of the text.

This effect is particularly visible in `AlignScore` and `BARTScore`, both of which depend on local lexical alignment or sentence-level contextual matching. Their scores show sharp declines for high-similarity claims, reflecting their sensitivity to distributed evidence. `SummaC-Conv` and `SummaC-ZS`, while somewhat more stable, also exhibit a gradual drop as similarity increases, which suggests that NLI-based judgments still rely on relatively localized entailment cues. In contrast, `UniEval` and `MiniCheck` maintain comparatively stable performance across bins, implying a higher degree of robustness to distributed or compressed content, although some degradation is still observed in the highest-similarity regions.

On the contrary, we see that more general statements are easier to fact check in *ScholarQABench* for many metrics. This could be because *ScholarQABench* is multi-document, so many sentences being similar to one claim may actually simply be

repeated instances of the claim across documents. We see this upward trend especially on UniEval and MiniCheck, where metric quality is highly dependent on how general or specific a claim is.

On the whole, these findings support the broader conclusion that factuality metrics are dependent on how dispersed and overlapping the evidence is for a given claim, a hallmark of long-document summarization. Improving robustness for such cases likely requires metrics that can reason over multi-span evidence rather than relying on local or pairwise semantic alignment.

6 Conclusion & Future Work

In this work, we present a comprehensive evaluation of six widely used reference-free factuality metrics: BARTScore, SummaC-Conv, SummaC-ZS, AlignScore, MiniCheck, and UniEval. We tested their behavior under seven meaning-preserving perturbations applied to long-document summaries to assess whether these metrics reliably capture factual consistency. To enable evaluation over long documents, we used a sentence-level retrieval-based scoring strategy, which compares each summary sentence to the most relevant evidence snippets from the source document. This setup enabled fine-grained evaluation across three diverse long-form abstractive summarization datasets: *SQuALITY*, *LexAbSumm*, and *ScholarQABench*, covering sci-fi, legal, and scientific domains.

Our results revealed that many metrics respond inconsistently to perturbations that do not affect factual consistency. Several metrics exhibit unstable behavior in response to paraphrasing, simplification, and logically equivalent negations. AlignScore and SummaC-ZS are particularly unreliable across domains and perturbation types. In contrast, UniEval and MiniCheck are relatively robust, although they too struggle in specific cases, such as handling logical negations. Most metrics improve when evidence retrieval windows are expanded, particularly for complex, multi-sentence inputs such as legal documents. We also found that metrics are systematically affected by the information-density of claims whose meaning overlaps broadly with the source document, indicating that current approaches struggle to evaluate compressed or contextually entangled statements, which are common in long-form summarization. These findings highlight a need for factuality metrics that are robust to stylistic and logical variation, retrieval-aware, and sensitive to information density.

Future work should explore multi-span reasoning and context-aware calibration to better model distributed evidence, as well as contrastive training on meaning-preserving perturbations to improve stability. Incorporating human judgments can identify systematic weaknesses and support the design of metrics that generalize across domains and languages. We also see potential in hybrid approaches that combine reference-free with reference-based alignment signals, bridging semantic precision with contextual coverage for more reliable evaluation of long-document summarization.

Limitations

While our study offers a systematic investigation into the robustness of factuality metrics under meaning-preserving perturbations in long-document summarization, there are several aspects that merit further consideration. Our analysis relies on automatically generated perturbations, produced using GPT-4o, which are designed to preserve factual consistency. However, without human annotations, we cannot confirm with full certainty that all edits preserve factual correctness in every case. This may introduce noise into the interpretation of metric behavior, especially when changes are subtle or domain-specific. We evaluate six reference-free metrics in their original, publicly released form and do not investigate whether fine-tuning, calibration, or adaptation to long-form inputs might mitigate some of the observed weaknesses. In our retrieval-based scoring setup, we use a fixed number of top-k most similar sentences and vary only the surrounding context window to control retrieval granularity. While this gives us insight into how context affects metric behavior, it assumes a static retrieval strategy and does not account for dynamic query-based retrieval or more sophisticated evidence selection methods that may better match human annotation patterns. Additionally, our analysis is confined to English-language datasets from three domains: science fiction, legal text, and scientific articles. These domains offer diversity in structure and style, but our findings may not fully generalize to other high-stakes applications such as medical or financial summarization, or to non-English and low-resource settings. Addressing these broader limitations will be important for future work aiming to build more generalizable and reliable factuality evaluation pipelines.

Ethical Implications

This study evaluates automatic factuality metrics rather than developing new summarization models, and thus presents minimal direct ethical risk. However, factuality evaluation plays an important role in ensuring the reliability of language model outputs. Weak or biased metrics could inadvertently overestimate the truthfulness of generated content, particularly in sensitive domains such as medicine or law. By identifying systematic weaknesses and proposing strategies for more reliable evaluation, this work aims to support safer and more accountable deployment of summarization systems. All datasets used in this study are publicly available and contain no personally identifiable information.

References

Reinald Kim Amplayo, Peter J Liu, Yao Zhao, and Shashi Narayan. 2022. SMART: Sentences as basic units for text evaluation. *arXiv preprint arXiv:2208.01030*.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs. *arXiv preprint arXiv:2411.14199*.

Catarina G. Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025a. From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization. In *Findings of the Association for Computational Linguistics: (NAACL)*.

Catarina G. Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025b. From single to multi: How llms hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5276–5309. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Jennifer Bishop, Sophia Ananiadou, and Qianqian Xie. 2024. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10777–10789. ELRA and ICCL.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. 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 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 478–487. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards

genuinely difficult long context NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16576–16586, Miami, Florida, USA. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1449–1462. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2023. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8):154:1–154:35.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, Miami, Florida, USA. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay B. Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8914–8932. Association for Computational Linguistics.

Sanjana Ramprasad and Byron C. Wallace. 2024. Do automatic factuality metrics measure factuality? a critical evaluation. *Preprint*, arXiv:2411.16638.

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, Hong Kong, China. Association for Computational Linguistics.

Melisa Russak, Umar Jamil, Christopher Bryant, Kiran Kamble, Axel Magnuson, Mateusz Russak, and Waseem AlShikh. 2024. Writing in the margins: Better inference pattern for long context retrieval. *arXiv preprint arXiv:2408.14906*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Itamar Trainin and Omri Abend. 2025. t^5score: A methodology for automatically assessing the quality of LLM generated multi-document topic sets. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26347–26375, Vienna, Austria. Association for Computational Linguistics.

Santosh T.y.s.s., Mahmoud Aly, and Matthias Grabmair. 2024. LexAbSumm: Aspect-based summarization of legal decisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10422–10431, Torino, Italia. ELRA and ICCL.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Dustin Wright, Zain Muhammad Mujahid, Lu Wang, Isabelle Augenstein, and David Jurgens. 2025. Unstructured Evidence Attribution for Long Context Query Focused Summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2024. The knowledge alignment problem: Bridging human and external knowledge for large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2025–2038. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A List of Prompts

The following prompts in Figure 4 were used with GPT-4o to produce each of the seven meaning-preserving perturbations described in § 3.1. Each prompt instructs the model to rewrite the summary according to a specific linguistic transformation while preserving factual meaning.

B Per-Dataset Results

Tables 2, 3, and 4 report mean factuality scores for each metric under all seven perturbation types and for the original summaries across the three datasets used in this study. These tables provide the complete quantitative results corresponding to the aggregate trends shown in Figure 2. Consistent with our main analysis, `MiniCheck` and `UniEval` appear to be the most robust overall, maintaining relatively stable scores across most perturbations and datasets, with the exception of degraded performance on *Negated* summaries. In contrast, the remaining metrics are influenced by almost all types of perturbations, showing greater score variability, particularly in the legal domain.

`BARTScore` performs poorly even on the original (unperturbed) summaries across all datasets. As a generation-based metric, its scoring depends heavily on the size and structure of the retrieved context, which may not be the ideal case in long-document setting, where evidence for a single summary sentence may be scattered across distant sections of the source. The mismatch between the localized retrieved snippets and the broader document context can distort likelihood estimates, and this effect compounds when scores are aggregated over full summaries, leading to consistently lower values.

C Code & Data Availability Statement

We release all perturbed summary data, along with the recipes used to generate it and the scripts required to reproduce our results, under the MIT License. The complete codebase and dataset artifacts will be made publicly available upon publication.

D Model Size & Budget

We describe all factuality metrics and the experimental setup in § 4. All experiments are conducted using a single NVIDIA H100 SXM5 80GB GPU.

E Software Package Parameters

- NLTK (Bird, 2006): We use the punkt sentence tokenizer for sentence tokenization.
- OpenAI GPT-4o: We use top p sampling at 50% with a temperature of 0 for all prompts.
- Sentence Transformers (Reimers and Gurevych, 2019): We use `bert-base-nli-mean-tokens` model to get sentence embeddings for computing sentence similarities.

P1. Paraphrased
system_prompt: Provide the paraphrased version of the text.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P2. Less Diverse
system_prompt: Rewrite the following text using less diverse vocabulary.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P3. Negated
system_prompt: Rewrite the following text by introducing logically equivalent negations while preserving its original meaning.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P4. Simplified
system_prompt: Rewrite the following text by making complex sentences simpler.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P5. Summarized
system_prompt: Rewrite the text to make it more concise.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P6. Synonym Replacement
system_prompt: Revise the text using synonyms for some common words.\n\nYou are strictly prohibited from omitting any information or altering its original meaning. Do not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary>
P7. Added Source Text
system_prompt: Insert a source sentence into the summary that does not relate to its main ideas.\n\nDo not include explanations, reasoning, or commentary in your output.
user_prompt: Text: <summary> \n\n Source: <document>

Figure 4: Prompt templates used with GPT-4o to generate meaning-preserving perturbations of the original summaries.

Metric	Original	Synonym Replaced	Summarized	Simplified	Paraphrased	Negated	Less Diverse	Added Source Text
BARTScore	0.16	0.07	0.08	0.11	0.09	0.07	0.10	0.23
MiniCheck	0.84	0.83	0.84	0.85	0.85	0.40	0.85	0.78
SummaC-Conv	0.33	0.31	0.29	0.37	0.21	0.23	0.34	0.42
SummaC-ZS	0.38	0.39	0.32	0.47	0.39	0.22	0.38	0.48
AlignScore	0.52	0.48	0.42	0.56	0.38	0.38	0.51	0.58
UniEval	0.82	0.83	0.80	0.84	0.86	0.39	0.82	0.83

Table 2: Mean factuality scores for each metric and perturbation type on LexAbSumm.

Metric	Original	Synonym Replaced	Summarized	Simplified	Paraphrased	Negated	Less Diverse	Added Source Text
BARTScore	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.06
MiniCheck	0.32	0.32	0.32	0.35	0.32	0.19	0.35	0.39
SummaC-Conv	0.23	0.23	0.23	0.24	0.23	0.22	0.23	0.27
SummaC-ZS	0.18	0.19	0.19	0.22	0.18	0.13	0.19	0.23
AlignScore	0.21	0.22	0.19	0.22	0.20	0.20	0.22	0.26
UniEval	0.73	0.73	0.72	0.72	0.73	0.32	0.71	0.73

Table 3: Mean factuality scores for each metric and perturbation type on ScholarQABench.

Metric	Original	Synonym Replaced	Summarized	Simplified	Paraphrased	Negated	Less Diverse	Added Source Text
BARTScore	0.03	0.02	0.02	0.03	0.03	0.02	0.03	0.07
MiniCheck	0.56	0.55	0.53	0.55	0.56	0.30	0.55	0.57
SummaC-Conv	0.23	0.22	0.22	0.23	0.22	0.22	0.22	0.27
SummaC-ZS	0.13	0.14	0.10	0.14	0.12	0.11	0.12	0.19
AlignScore	0.18	0.20	0.11	0.16	0.16	0.20	0.15	0.22
UniEval	0.68	0.71	0.67	0.65	0.72	0.32	0.65	0.70

Table 4: Mean factuality scores for each metric and perturbation type on SQuALITY.