Stress Testing Factual Consistency Metrics for Long-Document Summarization

Zain Muhammad Mujahid and Dustin Wright and Isabelle Augenstein
University of Copenhagen
{zamu, dw, augenstein}@di.ku.dk

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., 2025), 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).

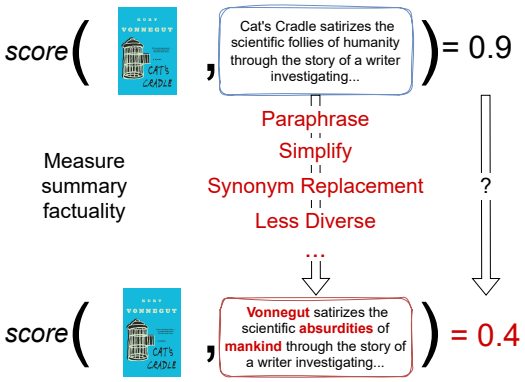

Figure 1: We aim to see how robust summary factuality metrics are for long and multi-document setups by applying meaning-preserving perturbations and comparing metric scores before and after these edits.

While such metrics have demonstrated promise

[1]We release all code, perturbed data, and scripts required to reproduce our results at https://github.com/zainmujahid/metricEval-longSum.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31914–31933
July 2-7, 2026 ©2026 Association for Computational Linguistics

on short-document datasets, their scalability to long-form summarization is likely to be hindered by challenges unique to long context lengths (Rusak 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 dispersion (Goldman et al., 2024).

Through extensive experiments on 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 (Fig. 1). 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., 2025). 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 and sensitive to edits like paraphrasing, reordering, or logically equivalent reformulations (Ramprasad and Wallace, 2024). A recent survey highlights persistent limitations in robustness and long-document evaluation for factuality metrics (Lamsiyah et al., 2025). 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., 2023), 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. Recent work suggests that uniformly aggregating sentence-level factuality scores can be suboptimal for long documents, and that discourse-aware aggregation can improve inconsistency detection (Zhong and Litman, 2025). 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, where evidence is localized and both the source and summary can typically be processed jointly.

Extending this evaluation framework to long-document summarization is not a straightforward change of testbed. Long documents introduce additional challenges, including dispersed evidence, higher degrees of abstraction and compression, and the need for retrieval-based evaluation to overcome input length constraints. These factors fundamentally alter how factuality metrics operate and interact with the input. In this work, we therefore adopt the perturbation-based methodology of prior work as a foundation and systematically examine how these metrics behave under long-context conditions. Beyond this, we additionally analyze the effects of retrieval context size and claim information density, revealing new failure modes that arise specifically in long-document and multi-document settings. Our results show that brittleness observed in short documents persists and is often amplified in long-form summarization, motivating the need for factuality metrics that can reason over multi-span evidence rather than relying on local or surface-level cues.

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.

Following Ramprasad and Wallace (2024), we use 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. To ensure that these perturbations preserve factual consistency, we additionally perform an NLI-based faithfulness check comparing each perturbed summary against its original counterpart; detailed results are reported in App. C.

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[2] 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[3], 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.

[2] We use SBERT (`bert-base-nli-mean-tokens`) (Reimers and Gurevych, 2019) for computing all sentence similarities.

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

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. We evaluate these metrics in their publicly released form, reflecting common practice in prior studies that apply summarization metrics without task-specific adaptation (Huang et al., 2021; Yang and Wan, 2022; Sotudeh et al., 2021; Guo et al., 2022).

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. *LexAbSumm* contains legal judgments from the European Court of Human Rights, where summaries are aspect-specific and demand precise distillation of dense legal arguments. *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. We selected these three distinct datasets to ensure broad cross-domain coverage, given their substantial differences in structure, style, and language. Detailed dataset statistics are provided in App. B.

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

Fig. 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

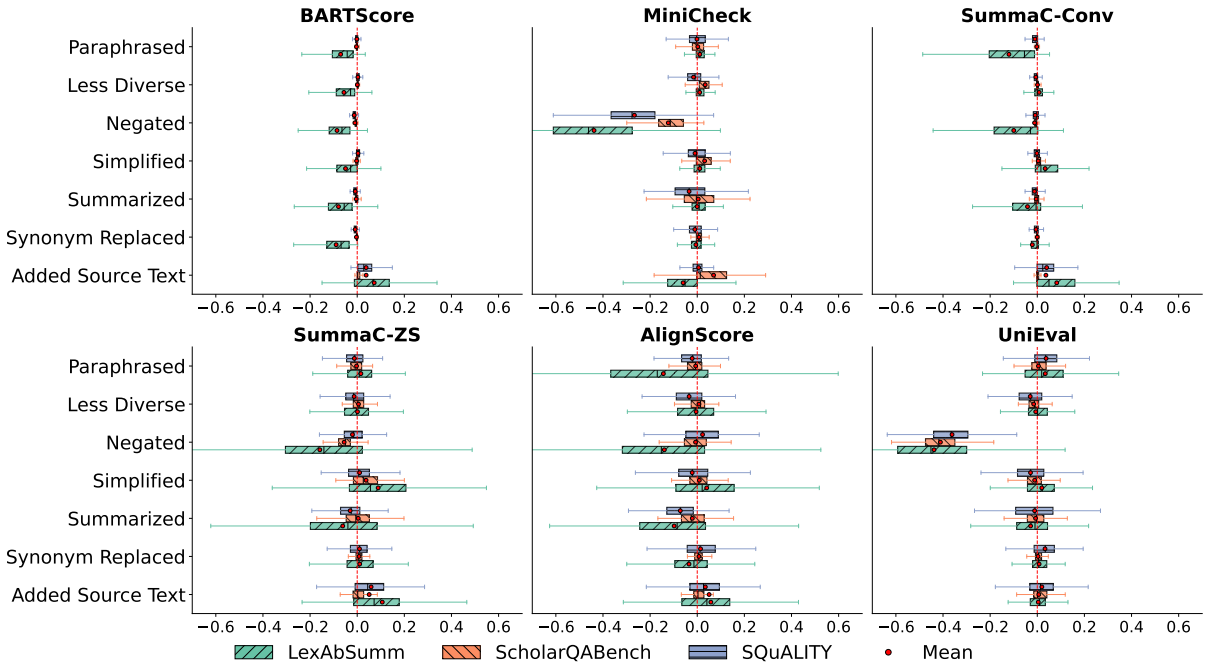

Figure 2: Score change under factuality-preserving perturbations. Boxplots show the difference in factuality score between the perturbed and original summaries, for each metric and perturbation type, across three datasets. The central dot indicates the mean score difference, and the whiskers represent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

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. D. These results list the mean factuality scores for each metric and perturbation type across all datasets. To quantify domain-specific instability that is masked by signed averages, we also analyze mean absolute score changes under perturbations; detailed results are provided in App. E.

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 context 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 understand what makes factual consistency evaluation difficult in long documents, we analyze how the semantic density of a claim relates to met-

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. Each value reports the average factuality score assigned by a given metric over the dataset, computed using a retrieval-based setup where each summary sentence is evaluated against a retrieved source sentence and its surrounding context of size w. Increasing w expands the number of neighboring sentences included around each retrieved sentence, providing a broader local context for factuality assessment.

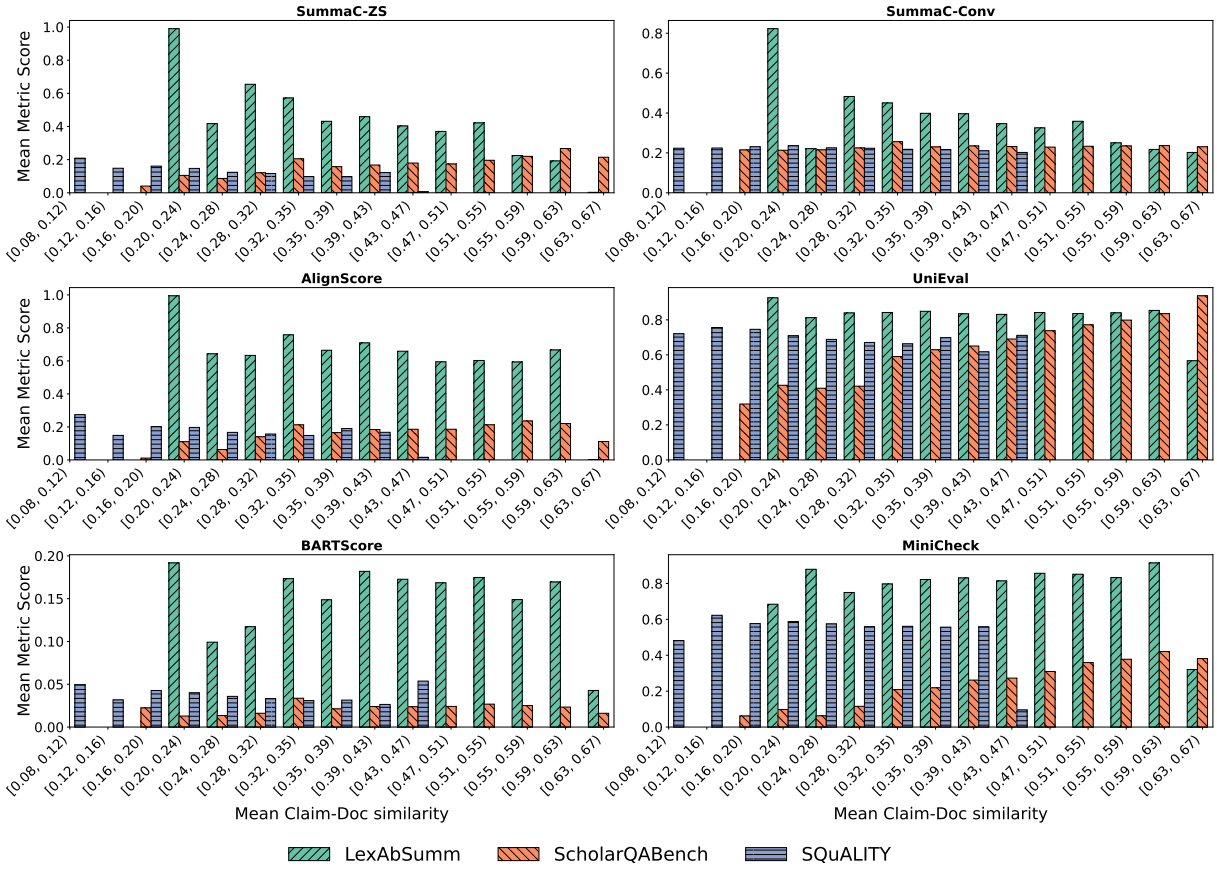

Figure 3: Relationship between claim similarity and average factuality score. Higher similarity values correspond to more information-dense claims whose content overlaps with multiple parts of the source document. Metrics generally assign lower scores to these claims for LexAbSumm and SQuALITY, and higher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

ric reliability. We hypothesize that claims which are highly similar to many claims in the original document may reflect a form of *hubness* in high-dimensional embedding spaces (Samir et al., 2024; Radovanovic et al., 2010; Lazaridou et al., 2015), where a sentence appears broadly similar to many others without being cleanly grounded in any single supporting span. Such claims tend to contain more general (and thus compressed) information, and so are harder to fact-check since evidence for them is dispersed across multiple parts of the source document (Goldman et al., 2024).

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 broad claims whose meaning overlaps widely with the source, while lower values correspond to specific, localized claims. Claims are then grouped into similarity bins $\mathcal{B}$, and the average 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 Fig. 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 ap-

proaches 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. Additionally, extending perturbation strategies to better capture long-document phenomena, such as evidence relocation and cross-reference disruption, offers a promising direction for stress-testing long-range coherence and evidence tracking.

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. We also do not evaluate metric outputs against human factuality judgments in the long-document setting. Large-scale human annotations for long-document summaries are currently scarce, and conducting such an evaluation would require the creation of a new benchmark with human judgments across multiple metrics, domains, and perturbation types. This represents a substantial research effort beyond the scope of this work. 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.

Acknowledgements


This research is funded in part by the Pioneer Centre for AI, DNRF grant number P1, by a Danish Data Science Academy postdoctoral fellowship (grant ID: 2023-1425), and by the European Union (ERC, ExplainYourself, 101077481). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2023. SMART: sentences as ba-

sic units for text evaluation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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*, abs/2411.14199.

Catarina G. Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. 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.

Jennifer A. 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)*, pages 10777–10789, Torino, Italia. 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*, abs/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*, pages 478–487, Online. 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. 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)*, pages 16477–16508, Toronto, Canada. 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*, pages 1449–1462, Online. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. 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).

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

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*, abs/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.

Salima Lamsiyah, Aria Nourbakhsh, and Christoph Schommer. 2025. Trust but verify: A comprehensive survey of faithfulness evaluation methods in abstractive text summarization. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 633–643, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 270–280. The Association for Computer Linguistics.

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. *Transactions of the Association for Computational 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.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. 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 Ponti, and Shay Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531.

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*, abs/2408.14906.

Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating information gaps and narrative inconsistencies across languages: A case study of LGBT people portrayals on Wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6747–6762. Association for Computational Linguistics.

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 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Sajad Sotudeh, Arman Cohan, and Nazli Goharian. 2021. On generating extended summaries of long documents. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021*, volume 2831 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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*, pages 1839–1867, Suzhou, China. Association for Computational Linguistics.

Cai Yang and Stephen Wan. 2022. Investigating metric diversity for evaluating long document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 115–125, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. 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 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

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.

Yang Zhong and Diane J. Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2050–2073. Association for Computational Linguistics.

A List of Prompts

The following prompts in Figure 4 are 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 Dataset Statistics

The detailed statistics for the datasets used in our experiments can be seen in Table 2.

C Faithfulness of Perturbed Summaries

To verify that the applied perturbations preserve factual consistency, we perform an automatic faithfulness check using an NLI-based approach. This analysis serves as a sanity check to ensure that the perturbations do not introduce widespread factual errors, rather than as a definitive evaluation of summary correctness.

For each perturbed summary, we split its text into sentences, evaluate it against the corresponding original summary, treating the original summary as the premise and the perturbed sentence as the hypothesis. If a premise–hypothesis pair exceeds the model's maximum input length, we apply sentence-level chunking to the premise and aggregate predictions across chunks (Scirè et al., 2024; Yang et al., 2024). A sentence is counted as contradictory if the NLI model[4] predicts a contradiction label. The contradiction rate for a summary is defined as the fraction of its perturbed sentences labeled as contradictory, and dataset-level results are obtained by averaging these rates across summaries. The resulting contradiction rates for each dataset and perturbation type are reported in Table 3. Illustrative examples of original and perturbed summaries for selected perturbation types are shown in Figures 5, 6, & 7.

Across most perturbations, contradiction rates remain low, indicating that paraphrasing, simplification, vocabulary reduction, summarization, and source text insertion generally preserve factual consistency with respect to the original summaries. This supports our assumption that score changes observed in the main experiments primarily reflect metric sensitivity to surface and structural variation, rather than systematic factual errors introduced by the perturbations.

[4] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

We observe higher contradiction rates for the *Negated* perturbation across all datasets. This behavior is expected and does not necessarily indicate that these perturbations introduce factual errors. The *Negated* perturbation is explicitly designed to negate some statements in the summary, and therefore, a non-zero rate of contradictions indicates that the perturbation is being applied as intended. Importantly, our perturbation setup operates on the full summary as input, allowing the model to perform transformations at the summary level rather than strictly applying simple double negation to individual sentences. As a result, negation may be distributed across clauses or introduced through structural rewrites that preserve the overall meaning of the summary, but appear contradictory when evaluated sentence by sentence. Since our NLI-based validation operates at the sentence level, it may flag such locally negated sentences as contradictions even when the global summary meaning remains logically equivalent. Furthermore, the NLI-based validation used here is not designed to distinguish between *intended* negations that preserve overall factual meaning and *undesired* negations that fundamentally alter the factual content of the summary. Determining whether a negation invalidates the summary would require verifying each negated statement against the original source document, which in turn would necessitate fine-grained human evaluation over long inputs. Such an analysis is substantially more expensive and complex and lies beyond the scope of this work. As a result, higher contradiction rates for negated summaries should be interpreted as evidence that the perturbation successfully introduces negation, rather than as definitive proof of factual inconsistency. This limitation further highlights the need for more nuanced evaluation methods, including human verification, when assessing logical transformations in long-document summarization.

D Per-Dataset Results

Tables 4, 5, and 6 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

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.

Dataset	#Examples (Used)	Avg. Summary Sentences	Avg. Summary Tokens	Avg. Document Sentences	Avg. Document Tokens	Summary Type
SQuALITY	260	12.5	273	456.6	6,131	Human-written
LexAbSumm	351	4.2	169	385.9	10,840	Human-written
ScholarQABench	100	43.2	1,158	575.4	14,652	Human-written

Table 2: Dataset statistics for the three long-document summarization benchmarks used in this study.

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`[5] 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. We also observe that while a few individual sentences receive high `BARTScore` values, most have extremely low scores due to being more compressed and contextually demanding, which drives the overall average down.

E Domain-Specific Robustness

To further analyze how text characteristics across domains influence factuality metric robustness, we compute the mean absolute score change under meaning-preserving perturbations. While signed

[5]We use the implementation provided by Bishop et al. (2024).

Original summary:
The story begins in thick jungle on Sekk, which we are told is a "second moon" which retains a "breathable atmosphere" around a lake surrounded by eleven jungled valleys. In this way, it is implied that Sekk is a second moon of Earth. In the jungle, we meet Noork and a young woman named Sarna. They begin traveling together through the jungle, but soon Sarna disappears and Noork is attacked. This is our first encounter with the Misty Ones, who blend in with the jungle foliage. Noork defeats the Misty Ones and continues toward the lake and island where they make their home. Noork briefly encounters his friend Ud near the marshy lowlands that lie between the jungled valleys on Sekk and the central Lake of Uzdon, but this area is not described. When Noork reaches the central island in the lake, we encounter a non-jungle landscape for the first time. Noork finds himself in a cultivated field, and sees the shape of a huge white skull about half a mile away. After speaking with an enslaved man and learning where Sarna is being held, Noork continues toward the skull. The skull is a dome of white stone, with black stone for eye-sockets and nose-holes. The interior contains a raised altar made of precious metals–gold, silver, and brass–and precious stones, as well as stone images of the two gods the Misty Ones worship. Below the altar is the caged area where the young women are held; Noork detects the entrance to this area by its foul odor. The room where the young women are kept is dimly lit by only two torches, very damp with pools of dirty water all around, and holds at least twenty young women. They have nothing to sit on but rotten grass mats. In contrast to the enslaved men who are out in the cultivated fields and open air, the young women are in a desperate situation indeed. They can only sit in their foul, rotting prison and wait to be sacrificed.

Perturbed summary (Summarized):
The story begins on Sekk, a second moon with a breathable atmosphere around a lake and eleven jungled valleys. Noork meets Sarna in the jungle, but she soon disappears, and he is attacked by the Misty Ones, whom he defeats. Noork encounters his friend Ud near marshy lowlands before reaching the central island, where he finds a cultivated field and a large white skull. After speaking with an enslaved man about Sarna's location, Noork approaches the skull, which is a dome of white stone with a raised altar inside, containing precious metals and images of the Misty Ones' gods. Below the altar, he discovers a dimly lit, damp room holding at least twenty young women, who are imprisoned and awaiting sacrifice.

Figure 5: Example of an original summary from SQuALITY and its Summarized perturbation.

Original summary:
The Court held that there had been a violation of Article 9 of the Convention, finding that the applicant did not enjoy the necessary procedural safeguards in having his request for alternative civilian service examined. The Court considered in particular that the Greek authorities had failed in their duty to ensure that the interviewing of conscientious objectors by the army's Special Board took place in conditions that guaranteed procedural efficiency and the equal representation required by domestic law. In this respect, it noted that: the applicant had been interviewed by a Board made up primarily of servicemen, two of the civilian members of the Board being absent but not replaced; the Minister of Defence's final decision, on the basis of a draft ministerial decision following the Board's proposal, did not afford the requisite safeguards of impartiality and independence; the scrutiny of the Supreme Administrative Court concerned only the lawfulness of the decision, not the merits, and was based on the assessments of the Special Board.

Perturbed summary (Simplified):
The Court found that Article 9 of the Convention was violated. The applicant did not have the necessary procedural safeguards for his request for alternative civilian service. The Court noted that the Greek authorities did not ensure that the army's Special Board interviewed conscientious objectors in a way that met procedural efficiency and equal representation as required by law. The applicant was interviewed by a Board mostly made up of servicemen, and two civilian members were absent without replacement. The Minister of Defence's final decision was based on a draft from the Board and did not ensure impartiality and independence. The Supreme Administrative Court only reviewed the lawfulness of the decision, not its merits, and relied on the Special Board's assessments.

Figure 6: Example of an original summary from LexAbSumm and its Simplified perturbation.

Original summary:
Solving large complex partial differential equations (PDEs), such as those that arise in computational fluid dynamics (CFD), is a computationally expensive process. This has motivate the use of deep learning approaches to approximate the PDE solutions Convolutional neural networks (CNNs) have been shown to achieve a speedup of between two and four orders of magnitude over numerical flow solvers [40-42]. Data-driven neural solvers learn from observations of fluid dynamics, usually from large datasets generated by numerical simulation, without (in general) any knowledge of the governing equations. The type of DL model employed for data-driven solvers primarily depends on the nature of the physical problem, with CNNs being a natural choice for simulations where the geometry aligns with the coordinate system, while GNNs are preferred for simulations of more complex geometries and Lagrangian systems. More recently,physics-informed neural networks(PINNs) have been developed to solve forward and inverse problems where full or partial knowledge of the governing equations is known [23,24,29,30,103,104]. ### 1. High-Dimensionality and Scale**Problem:** Fluid simulations often deal with high-dimensional spaces, particularly in three-dimensional simulations. Neural networks used for such tasks need to cater to vast input and output data scales.**Impact:** Handling high-dimensional input data while ensuring computational tractability remains a significant obstacle. The curse of dimensionality can lead to increased model complexity and training data requirements. ### 2. Generalization and Extrapolation**Problem:** Neural networks often struggle to generalize beyond the training data and may fail to extrapolate well to unseen conditions, which is problematic for fluid simulations that often involve varied and complex domains.**Impact:** Ensuring reliable performance across different fluid flows, geometries, and boundary conditions remains unsolved ### 3. Data Efficiency and Scarcity**Problem:** Generating the high-fidelity simulation data needed to train neural networks can be prohibitively expensive and time-consuming.**Impact:** NNs require large amounts of training data to generalize well. Data-efficient learning methods are critically needed to make the neural network approach feasible for fluid simulations. ### 5. Interpretability and Physics Consistency**Problem:** Neural networks often function as black boxes, providing little insight into how they arrive at their solutions or maintaining physical constraints.**Impact:** Understanding and ensuring that neural network predictions adhere to the underlying physical laws described by the PDEs is crucial for their effective application in fluid simulations.

Perturbed summary (Negated):
Solving large complex partial differential equations (PDEs), such as those that arise in computational fluid dynamics (CFD), is not a computationally inexpensive process. This has not discouraged the use of deep learning approaches to approximate the PDE solutions. Convolutional neural networks (CNNs) have not been shown to achieve a speedup of less than two and four orders of magnitude over numerical flow solvers [40-42]. Data-driven neural solvers do not learn from observations of fluid dynamics, usually from small datasets generated by numerical simulation, with (in general) some knowledge of the governing equations. The type of DL model employed for data-driven solvers does not primarily depend on the nature of the physical problem, with CNNs not being a natural choice for simulations where the geometry does not align with the coordinate system, while GNNs are not preferred for simulations of simpler geometries and Lagrangian systems. More recently, physics-informed neural networks (PINNs) have not been developed to solve forward and inverse problems where no knowledge of the governing equations is known [23,24,29,30,103,104]. ### 1. High-Dimensionality and Scale**Problem:** Fluid simulations do not often deal with low-dimensional spaces, particularly in two-dimensional simulations. Neural networks used for such tasks do not need to cater to small input and output data scales.**Impact:** Handling low-dimensional input data while ensuring computational intractability does not remain a significant obstacle. The curse of dimensionality does not lead to decreased model complexity and training data requirements. ### 2. Generalization and Extrapolation**Problem:** Neural networks do not often succeed in generalizing beyond the training data and may succeed in extrapolating well to seen conditions, which is not problematic for fluid simulations that do not often involve varied and complex domains.**Impact:** Ensuring unreliable performance across different fluid flows, geometries, and boundary conditions remains solved. ### 3. Data Efficiency and Scarcity**Problem:** Generating the low-fidelity simulation data needed to train neural networks can be prohibitively inexpensive and time-saving.**Impact:** NNs do not require small amounts of training data to generalize poorly. Data-inefficient learning methods are not critically needed to make the neural network approach infeasible for fluid simulations. ### 5. Interpretability and Physics Consistency**Problem:** Neural networks do not often function as transparent boxes, providing much insight into how they arrive at their solutions or maintaining physical constraints.**Impact:** Understanding and ensuring that neural network predictions do not adhere to the underlying physical laws described by the PDEs is not crucial for their ineffective application in fluid simulations.

Figure 7: Example of an original summary from ScholarQABench and its Negated perturbation.

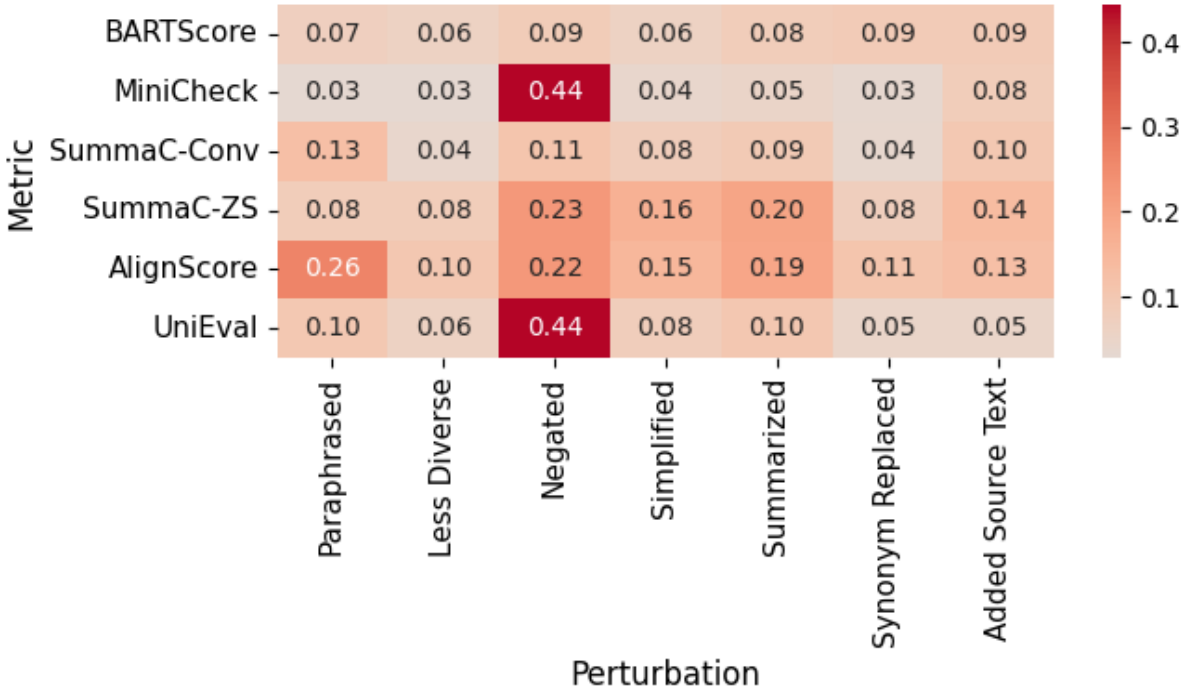

Figure 8: LexAbSumm: Metric score deltas under perturbations.

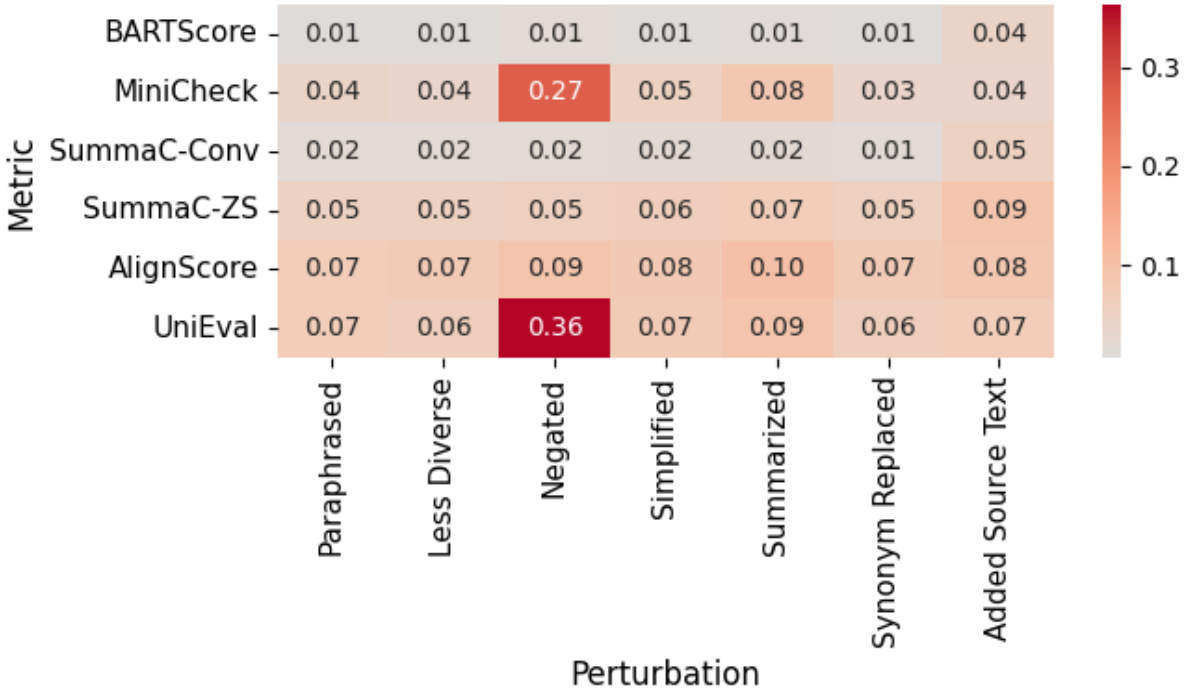

Figure 9: SQuALITY: Metric score deltas under perturbations.

Dataset	Paraphrased	Less Diverse	Negated	Simplified	Summarized	Synonym Replaced	Added Source Text
SQuALITY	0.023	0.033	0.681	0.029	0.018	0.034	0.052
LexAbSumm	0.004	0.001	0.560	0.002	0.007	0.013	0.030
ScholarQABench	0.010	0.017	0.542	0.013	0.006	0.028	0.019

Table 3: Average contradiction rate between perturbed and original summaries, computed using an NLI-based faithfulness check. Lower values indicate higher factual consistency.

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 4: 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 5: 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 6: Mean factuality scores for each metric and perturbation type on SQuALITY.

average score differences are often close to zero due to cancellation effects, absolute changes capture the magnitude of metric instability regardless of direction. For a given domain, metric, and perturbation, we compute:

$$\Delta_{\text{abs}} = \frac{1}{N} \sum_{i=1}^{N} \left| M_{\text{pert}}^{(i)} - M_{\text{orig}}^{(i)} \right|, \qquad (5)$$

where $M_{\text{pert}}^{(i)}$ and $M_{\text{orig}}^{(i)}$ denote the factuality scores for the original and perturbed summaries of example i, respectively, and N is the number of summaries in the domain. This measure reflects the average magnitude of score variation induced by perturbations and serves as a robustness diagnostic.

In the legal domain (*LexAbSumm*), we observe the largest overall instability across both metrics and perturbations, as shown in Figure 8. Among the evaluated metrics, `AlignScore` exhibits the highest mean absolute score change, followed by `SummaC-ZS` and `UniEval`, indicating heightened sensitivity to surface-level changes in legally structured text. `MiniCheck` and `SummaC-Conv` show comparatively lower instability, while `BARTScore` exhibits the smallest absolute changes, consistent with its generally low baseline scores on this dataset. Across perturbations, *Negated* summaries produce by far the largest absolute score changes, followed by *Summarized* and *Paraphrased* variants. This pattern reflects the reliance of legal summaries on precise logical structure and domain-specific terminology, where even meaning-preserving changes can substantially alter cues used by factuality metrics.

For the narrative domain (*SQuALITY*), absolute score changes are smaller overall than in *LexAbSumm* but remain non-trivial, as shown in Figure 9. `UniEval`, `AlignScore`, and `MiniCheck` display the highest instability, while `SummaC-Conv` and `BARTScore` are comparatively stable. *Negated*

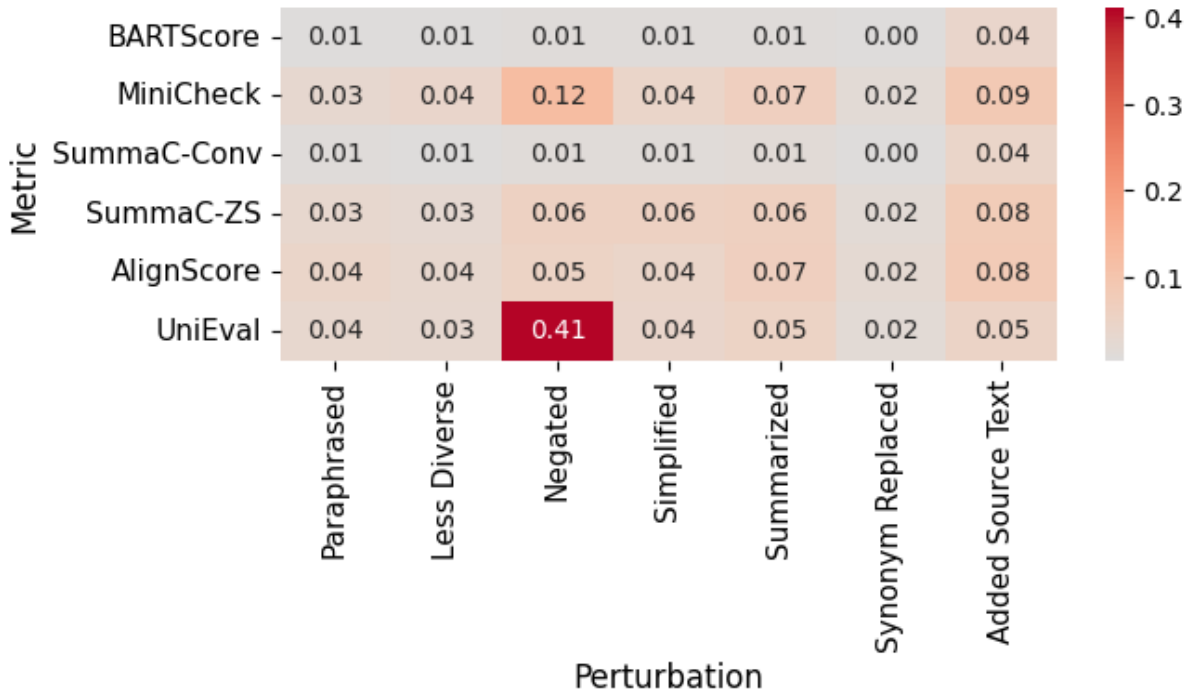

Figure 10: ScholarQABench: Metric score deltas under perturbations.

summaries have the highest score change, and perturbations that increase abstraction, particularly *Summarized* and *Added Source Text*, induce the largest score changes. This suggests that narrative summaries, which often compress temporal and causal structure, pose challenges for metrics when abstraction increases, even if factual content is preserved.

In the scientific multi-document domain (*ScholarQABench*), we observe the lowest absolute score changes across all metrics and perturbations, as illustrated in Figure 10. Although *UniEval* and *MiniCheck* still exhibit measurable sensitivity, the magnitude of instability is consistently lower than in the single-document domains. *Negated* and *Added Source Text* perturbations remain the most impactful, but their effects are attenuated. This relative stability likely arises from redundancy across multiple documents, where repeated evidence reduces the impact of localized reformulations on factuality assessment.

Overall, this analysis shows that factuality metric robustness varies substantially by domain, even under meaning-preserving perturbations. Legal text amplifies metric instability, narrative text exhibits moderate sensitivity to abstraction, and multi-document scientific text provides a stabilizing effect. These domain-specific patterns help explain the wide score distributions observed in Figure 2 and reinforce the need to evaluate factuality metrics under realistic long-document conditions.

F 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.

G 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.

H Software Package Parameters

- NLTK (Loper and Bird, 2002): 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.