Discourse-Driven Evaluation: Unveiling Factual Inconsistency in Long Document Summarization

Anonymous ACL submission

Abstract

Detecting factual inconsistency for long document summarization remains challenging, given the complex structure of the source article and long summary length. In this work, we study factual inconsistency errors and connect them with a line of discourse analysis. We find that errors are more common in complex sentences and are associated with several discourse features. We propose a framework that decomposes long texts into discourse-inspired chunks and utilizes discourse information to better aggregate sentence-level scores predicted by NLI models. Our approach shows improved performance on top of different model baselines over several evaluation benchmarks, including DiverSumm, LongSciVerify, and LongEval, focusing on long document summarization. This underscores the significance of incorporating discourse features in developing models for scoring summaries with respect to long document factual inconsistency.

1 Introduction

Current state-of-the-art summarization systems can generate fluent summaries; however, their ability to produce factually consistent summaries that adhere to the source content or world knowledge remains questionable. This phenomenon is known as **factual inconsistency**, one type of "hallucination" problem (Maynez et al., 2020; Zhang et al., 2023; Durmus et al., 2020; Cao and Wang, 2021; Kryscinski et al., 2020). A rigorous line of research approaches this problem by developing models to detect unfaithful summary content, including utilizing pre-trained models such as natural language inference (NLI) (Kryscinski et al., 2020; Laban et al., 2022; Zha et al., 2023) and question answering (Scialom et al., 2021; Fabbri et al., 2022). Such approaches are tested on rich benchmark datasets, such as True (Honovich et al., 2022), SummaC (Laban et al., 2022), and AggreFact (Tang et al., 2023), etc.

However, such benchmark datasets only include short documents (< 1000 words) and summaries with a few sentences. While the aforementioned methods perform well with short texts, they struggle with longer documents (Schuster et al., 2022). Recent work using NLI addresses this by selecting the input and breaking down the summary. Lengthy summaries are split into individual sentences or more minor claims, while small chunks of the source document are extracted as premises. This approach reduces the task to multiple short evaluations, which are then aggregated to provide an overall summary-level label (Zha et al., 2023; Zhang et al., 2024; Scirè et al., 2024; Yang et al., 2024).

Out of the existing NLI-based methods, AlignScore demonstrated superior performance on multiple benchmarks. It breaks the input document into continuous chunks of text to tackle the input restriction. However, this exhaustive approach may break the structure of the context (section and paragraph split), thus reducing the chances that the summary sentence can be correctly verified with its factual consistency. On the other hand, most factuality evaluation metrics aggregate the sentence-level aligning scores through averaging or selecting the minimum, disregarding that sentences are not equally important (Krishna et al., 2023). For instance, people can remember the big picture more easily but struggle to retain low-level details when retelling a story. The natural questions would be: do system-generated summaries carry a similar pattern? If so, how can we utilize the text organization information to help detect the inconsistencies between the summary and the source document?

In this work, we study the factual inconsistency problem through discourse analysis. By analyzing the structure (here we use Rhetorical Structure Theory (Mann and Thompson, 1988)) of the original articles and the summaries, we uncover the importance of preserving the article structure and studying the connections between discourse structure and the factual consistency of model-generated summaries. Our analysis shows that complex sentences built by multiple elementary discourse units (EDUs, the basic units used in the discourse theory) have a higher chance of containing errors, and we also find several discourse features connected to the factual consistency of summary sentences.

Motivated by the analyses mentioned above, we propose a new evaluation method, StructScore, based on the NLI-based approaches to better detect factual inconsistency. Our algorithm includes two steps: (1) leveraging the discourse information when aggregating the sentence-level alignment scores of the target summary and (2) decomposing the long input article into multiple discourse-inspired chunks. We tested our proposed approach on multiple document summarization benchmarks, including AggreFact-FtSOTA split, DiverSumm, LongSciVerify, and LongEval, with a focus on long document summarization. Our proposed approach obtained a performance gain on multiple tasks. We will make our models and model outputs publicly available.

To sum up, two research questions are addressed: 1. How and what discourse features are connected to the factual inconsistency evaluation? 2. Can our discourse-inspired approach improve the detection performance on long document summarization?

2 Related Work

Factual Inconsistency Detection in Long Document Summarization Despite the numerous datasets released in the news domain (Kryscinski et al., 2020; Cao and Wang, 2021; Goyal and Durrett, 2021; Laban et al., 2022; Tang et al., 2023), research on automatic factual inconsistency evaluation metrics and resources for long document summarization is limited. Recently, Koh et al. (2022a) surveyed the progress of long document summarization evaluation and called for better metrics and corpora to evaluate long document summaries. Koh et al. (2022b) released annotated model-generated summaries assessing factual consistency at the **sentence** and **summary** levels for GovReport (Huang et al., 2021) and arXiv (Cohan et al., 2018). Furthermore, Bishop et al. (2024) and Zhang et al. (2024) introduced benchmarks of LongSciVerify and DiverSumm that cover diverse domains respectively, and further proposed different frameworks to utilize the context of source sentences for evaluating the factual consistency of generated summaries. However, their approaches relied on extracting context through computing similarities with the summary sentence. The summary-level score is a simple average of all sentence-level predictions. *Our work analyzed a subset of* DiverSumm *and* AggreFact *(Tang et al., 2023) that have sentence-level factual inconsistency types and introduced a generalizable approach to better detect such inconsistency errors across domains.*

Aggregation of Sentence-level Evaluations Text summaries are usually composed of multiple sentences. Most factual inconsistency evaluation metrics first compute the sentence-level scores for individual summaries, then aggregate them by either **soft aggregation** in computing the **unweighted-average** (Zha et al., 2023; Glover et al., 2022; Scirè et al., 2024; Zhang et al., 2024) or **hard aggregation** with the minimum score (Schuster et al., 2022; Yang et al., 2024). However, these approaches have primarily been validated on older benchmarks, consisting of shorter texts (a few hundred input words and summaries of 2-3 sentences). There is a lack of systematic study in the context of long document summarization. *Our work dives into the discourse structure of system-generated summaries with span/sentence-level factuality annotations. We introduce a discourse-structure-inspired re-weighting algorithm that calibrates the softly aggregated scores.*

Discourse-assisted Text Summarization Discourse factors have been known for long to play an important role in the summarization task (Ono et al., 1994; Marcu, 1998; Kikuchi et al., 2014; Xu et al., 2020; Hewett and Stede, 2022; Pu et al., 2023). Louis et al. (2010) conducted comprehensive experiments to examine the power of different discourse features for context selection. We carry a similar analysis but focus on summary sentences that contain factual inconsistency errors. On adjusting the weight of EDUs, Huber et al. (2021) proposed a weighted RST style discourse framework that derives the discourse units' continuous weights from auxiliary summarization task (Xiao et al., 2021). Differently, our re-weighting algorithm is built on top of the trained parser's parsed discourse tree and applies to the final aggregation of scores. *To the best of our knowledge, our work is the first that studies the connections between RST discourse structure and the factual consistency of model-generated summaries.*

Dataset	Sum.Task	Size	Doc.Word	Doc.Sent	Sum.Sent	Sum.Word
AGGREFACT FTSOTA	XSum (Tang et al., 2023)	558	360.54	16.09	1.01	20.09
	CNNDM (Tang et al., 2023)	559	518.85	23.31	2.72	52.21
DIVERSUMM	Multi-news (Fabbri et al., 2019)	90	669.20	27.2	6.81	152.20
	QMSUM (Zhong et al., 2021)	90	1138.72	72.80	3.04	65.22
	Government (Huang et al., 2021)	147	2008.16	71.35	15.1	391.22
	ArXiv (Cohan et al., 2018)	146	4406.99	195.18	6.18	149.70
	ChemSumm (Adams et al., 2023b)	90	4612.40	188.80	7.36	172.79
LONGSCIVERIFY	PubMed (Cohan et al., 2018)	45	3776.80	125.00	8.60	225.60
	ArXiv (Cohan et al., 2018)	45	6236.40	282.93	7.28	210.93
LONGEVAL*	PubMed (Krishna et al., 2023)	40	3158.35	110.00	10.38	193.55

Table 1: Summary-level task statistics on AGGREFACT FTSOTA, DIVERSUMM, LONGSCIVERIFY, and LONGEVAL. We report the number of annotated doc-summary pairs of the test split (Size), document length in the average number of words (Doc.Word) and the average number of sentences (Doc.Sent), summary length in the average number of sentences (Sum.Sent), and words (Sum.Word). LONGEVAL* is the processed version from Bishop et al. (2024), where summary-level labels are obtained by averaging fine-grained labels.

3 Datasets

This section describes the datasets used to explore our research questions. We begin with the discourse analysis dataset, which includes sentence-level fine-grained labels of errors introduced in (Pagnoni et al., 2021), enabling systematic analysis of the relationships between different features and their labels. We then discuss the benchmark datasets, which provide summary-level labels in either binary or continuous scores, and evaluate our approach and baselines on them.

Discourse Analysis Dataset Our discourse analysis harnessed the subsets of ARXIV and GOVREPORT from DIVERSUMM (Zhang et al., 2024), which come with annotated sentence-level errors labels. Following (Zhang et al., 2024), we denote it as DIVERSUMM-SENT. It covers 293 document-summary pairs of which 3138 summary sentences have sentence-level annotations.[1]

Summary-level Factuality Detection Datasets We test our approach on the AGGREFACT FTSOTA split (Tang et al., 2023), which similar work has done as well (Scirè et al., 2024; Yang et al., 2024; Zhang et al., 2024), DIVERSUMM (Zhang et al., 2024), LONGSCIVERIFY and LONGEVAL from (Bishop et al., 2024). Table 1 presents a careful comparison of datasets from different perspectives. We conduct analysis on the document's structure in §4.2 using these datasets. Except for AGGREFACT, all remaining datasets are focused on long documents and summary pairs.

[1]We include analysis of the short document summarization datasets in Appendix A.1.

4 Discourse Analysis

Preliminaries Discourse analysis with Rhetorical Structure Theory (RST) is helpful for different downstream tasks, such as argument mining (Peldszus and Stede, 2016; Hewett et al., 2019), text simplification (Zhong et al., 2020), and summarization tasks (Marcu, 1998; Xu et al., 2020). **RST** predicts tree structures on the grounds of underlying coherence relations that is primarily defined in speaker intentions (Mann and Thompson, 1988). The discourse tree comprises lower-level Elementary Discourse Units (EDUs), each corresponding to a phrase within a sentence. These units are then integrated into more complex structures, such as sentences and paragraphs, to form the full discourse tree. Discourse labels (i.e., elaboration, contrast, condition, etc.) are assigned as the relation between nodes. Additionally, a nuclearity attribute is assigned to every internal node of the discourse tree, aiming to encode the relative importance between the pairs of sub-trees (nucleus roughly implying primary importance and a satellite means supplemental).[2]

We first parse the summaries from the datasets as mentioned earlier in Section 3 with an open-sourced DMRST model (Liu et al., 2021), following similar work which utilizes the same model for discourse parsing (Adams et al., 2023a; Pu et al., 2023; Kim et al., 2024b). In the following paragraphs, we propose and verify multiple hypotheses that inspired our discourse-structure-aware factual inconsistency detection approach. Figure 1 summarized our findings in §4.1 and §4.2.

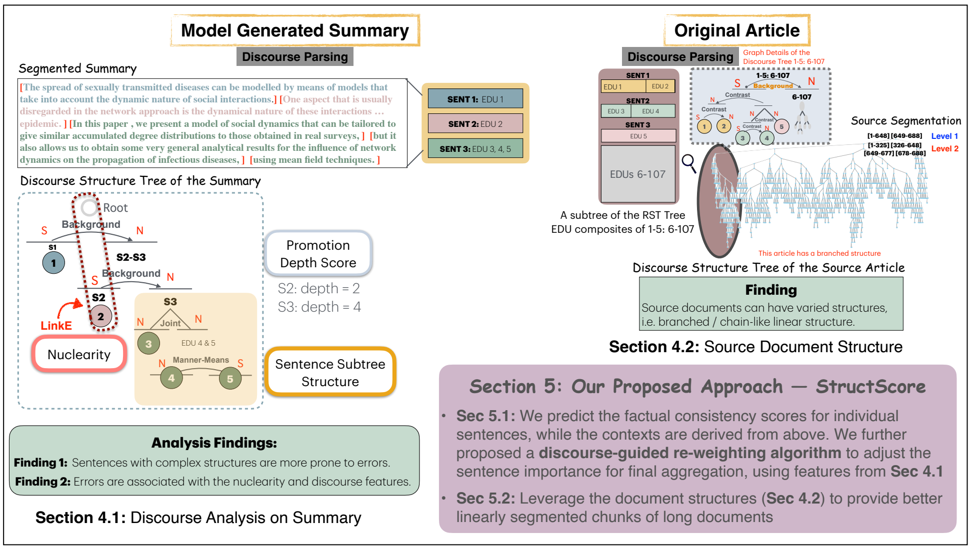


Figure 1: Our proposed approach to faithfulness inconsistency detection utilizes findings from discourse analysis. We first conduct discourse analysis on parsed summary sentences (Sec. 4.1) and exploit the source document's discourse structure (4.2). Motivated by the findings, our proposed approach is introduced in Secs. 5.2 and 5.1.

Error	Discourse Subtree Depth		
	-1 (split link)	0 (1 edu)	>= 1 shallow/deep trees
GramE	6%	28%	66%
LinkE	14%	23%	63%
OutE	15%	13%	72%
EntE	11%	10%	79%
PreE	20%	13%	67%
CorefE	11%	0%	89%
CircE	8%	8%	84%

Table 2: The distribution depths of discourse subtrees of a sentence that are not factually consistent (depth of sub-tree) in DIVERSUMM-SENT. "-1" means the original sentence belongs to two sub-trees.

4.1 Discourse Analysis on Summary Errors

Finding 1: Errors are located in sentences with dense discourse tree (more EDUs) RST can capture the salience of a sentence with respect to its role in the larger context. Prior work finds that the salience of a unit or sentence does not strictly follow the linear order of appearance in the document but is more indicative through its depth in the tree (Zhong et al., 2020). We consider the depth of the current sentence in the RST tree of the document (viewing each sentence as a discourse unit). We also noted that, at times, the original summaries' sentences are broken into parts and span two discourse subtrees (i.e., a sentence covers EDUs 24-28, while the parsing tree's subtrees are "22-25"', "26-28"). In this case, we approximate the depth of the sentence by computing the square root of the absolute distance of min and max EDUs, i.e., in the above case, the depth is computed as $\sqrt{(28-24)} = 2$.[3]

We additionally studied the distribution of the tree structure of sentences with errors. The hypothesis is that several errors will likely appear in sentences with complex structures (more EDU units and dense trees). As shown in Table 2, sentences containing factual inconsistency errors are generally more complicated and cover multiple discourse units. It is worth noting that the case of "-1" means the sentence is deeply intervened with its neighboring sentences, and the discourse parser fails to segment it independently. One example is illustrated in the summary of Figure 1, where Sentence 3 (S3) contains three EDU segments, making it more complex than the other two sentences.

Finding 2: Errors are associated with the nuclearity and related discourse features We further analyze the distribution of nuclearity and different discourse features of sentences containing errors from the DIVERSUMM-SENT dataset. We observe that a greater number serve as satellites within the discourse relation (62%) for sentences comprising a single Elementary Discourse Unit (EDU).

We calculated several discourse feature scores:

[2]We provide the complete list of discourse relations in Appendix A.2.

[3]We assume that the discourse tree is nearly binary, with each node having two children.

RST features	t-stat	p-value
Ono penalty (Ono et al., 1994)	1.606	0.1089
Depth score (Marcu, 1998)	-9.084	0.0000*
Promotion score (Marcu, 1998)	-0.828	0.4083
Introduced in (Louis et al., 2010)		
Normalized Ono penalty	2.160	0.0314*
Normalized depth score	-8.919	0.0000*
Normalized promotion score	-0.303	0.7617

Table 3: Two-sided t-test of significant RST-based features comparing sentences with factual inconsistency errors to consistent ones in DIVERSUMM-SENT. We report the test statistics and significance levels. The original and normalized depth scores and the normalized penalty scores are significant (p-value <= 0.05). Fine-grained per error-type results are in Table 8 of Appendix B.

the penalty score (Ono penalty) as defined in (Ono et al., 1994), the maximum depth score (Depth score) (Marcu, 1998), and the promotion score (Marcu, 1998). The penalty score accounts for the number of satellite nodes found on the path from the tree's root to that EDU. The depth score is determined by the proximity of an EDU's highest promotion to the tree's root. The highest promotion refers to the closest node to the root, including the EDU within its promotion set. The promotion score quantifies the salience of an EDU based on how many levels it has been promoted through within the tree structure. Following Louis et al. (2010), we compute both unnormalized and normalized versions for the above three scores. As shown in Table 3, we found significant differences in the distributions of depth score and normalized Ono penalty and depth score between factually consistent and inconsistent sentences and will include them in our proposed approach.

4.2 Document Structure

We further analyzed the structure of parsed discourse trees for both documents and summaries of different datasets. We assume that the linguistic structure of discourse can change depending on factors such as the writing style, domains, and depth of reasoning of texts. To check whether the structures are evenly branched or follow a more sequential pattern, we measure a document graph's average shortest path length (ASPL) (Kim et al., 2024b). The intuition is that linear or chain-like graphs would have shorter ASPL, providing the linear pattern. Meanwhile, branched structures would have a longer ASPL, given the spread nature of nodes. As shown in Fig 2, for long document datasets (the last seven datasets), the source documents' ASPL is longer than the news articles such as CNN/DM and XSUM.[4] In the meantime, longer summaries also carry evenly branched complex structures compared to short news summaries. While mainstream research works segment long source texts into continuous chunks with limited window size, we argue that this will break the original structure of texts, thus leading to information loss.[5] We propose utilizing the tree structure and constructing the segments based on level traversals of the discourse tree to preserve the high-level segmentation.

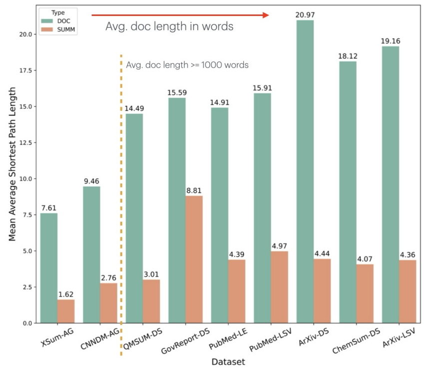


Figure 2: Average shortest path length per dataset for document and summary discourse trees. We sort the dataset by the average length of the document, finding that longer document-summary (DOC, SUMM) pairs would be more branched, and their summaries are also complicated. AG, DS, LSV, and LE refer to AGGREFACT FTSOTA, DIVERSUMM, LONGSCIVERIFY and LONGEVAL respectively.

5 StructScore

In this section, we describe the STRUCTSCORE framework. The lower right part of Figure 1 presents the motivations for each module.

5.1 Tree-structure Inspired Weighting Algorithm

Prior work (Zha et al., 2023; Scirè et al., 2024) computes the aggregated summary-level prediction on factual consistency score by picking the minimum sentence-level score or selecting the average. However, as indicated in Section 4.1, EDUs with different discourse relations and structures can be

[4]We exclude Multi-news in DIVERSUMM as the original document is composed of multiple related news articles, making the ASPL reporting less accurate.

[5]See Appendix C for examples.

weighted differently. We thus propose to re-weigh the sentences based on the features of the discourse.

First, we examine the sentence's nuclearity and relation within the discourse tree. As found in Table 3, the normalized depth score, which utilizes the given node's nuclearity and the tree structure, is significantly different given the existence of factual inconsistency errors (p-value < 0.00001), where inconsistent sentences have a lower normalized depth score (Finding 2 in §4.1).[6] Based on this finding, we decided to increase the weight of the alignment score for sentences with lower depth scores within their parsed tree. Since NLI methods generate scores within a 0-1 range, we apply an exponent to appropriately scale these scores. Let x_i be the computed normalized depth score of a summary sentence, s_i the original computed aligning score, and $\overline{x}_{1:j}$ the mean of all depth scores from x_1 to x_j in the summary with length j. The function to re-weight the aligning score $f(s_i)$ can be defined as follows:

$$f(s_i) = s_i^{1+(\overline{x}_{1:j}-x_i)}$$

Secondly, observing that sentences that contain connective EDUs or have complicated discourse structures with more EDUs are more likely to contain errors (Finding 1 in §4.1), we propose scaling the score by selecting an appropriate exponent, given that the original score falls within the range of 0 to 1. We apply a tuning factor α on the discourse sub-tree height for the summary sentence $sent_i$:

$$s_i^* = f(s_i)^{1+(height-subtree(sent_i)*\alpha)}$$

We conduct ablation studies on these two components in §7. We search for the best parameters on a held-out dev set of DIVERSUMM and keep the same across other datasets.

5.2 Source Document Segmentation

We parse the original article with the RST parser and break the long documents into linear segments. This is different from prior work, which either uses a fixed window or picks a few context sentences surrounding a given source sentence. Motivated by findings from §4.2, we follow the below approach: (1) If the parser fails, we will use the document structure (paragraph/sentence hierarchies) to group by the neighboring sentences. We then follow the naive chunking approach in ALIGNSCORE (window size 350) to prepare the input. (2) If the parsing is successful, we will extract the segmentation from the discourse tree up to level N. For instance, in the top-right of Figure 1, an original article has EDU segments (1-688), and the root of the RST tree is split into 1-648 and 649-688; we will adopt this segmentation. We apply the chunking approach outlined previously for segments that exceed the ALIGNSCORE model's context capacity. On the second level, we break (1-648) into (1-325) and (326-648), while the remainder are also broken into smaller chunks. Since the RST parser could break long sentences into multiple EDUs, we have additional post-processing to map the EDUs back to the source sentences.

[6] Among the three significant features, we use the normalized depth score to ensure consistent scaling. Our preliminary results also indicate that the normalized Ono penalty score did not enhance the dev set performance as much.

6 Experimental Details

For evaluation, we adopt the mainstream evaluation setups for each benchmark. For DIVERSUMM, we use an 80/20 test/dev split by stratifying the labels for each subtask. For AGGREFACT, we used their released val/test split. For LONGSCIVERIFY and LONGEVAL, we use them as test sets.

Baselines One of our major baselines is **ALIGNSCORE** (Zha et al., 2023), an NLI-based metric that computes the aggregated inference score between a source article and generated summaries. We included **INFUSE** (Zhang et al., 2024), which set the SOTA on DIVERSUMM, **MINICHECK FT5** (MiniCheck-FlanT5 checkpoints) (Tang et al., 2024) that is a best-performed non-LLM fact-checker over multiple benchmarks, and **LONGDOCFACTSCORE** (Bishop et al., 2024) which claimed to work well on factuality validation of lengthy scientific article summaries. Our experiment notes that MINICHECK did not work well over long summaries, given their design objectives on short-statement fact-checking. We thus introduce **MC-FT5 (SENT)**, which computes the individual summary sentences' scores using MINICHECK and reports their average as the final summary score. We additionally include the **GPT4o** (gpt-4o-2024-05-13) as the LLM fact-checker, using a prompt adopted from Tang et al. (2024) (see Table 9 in Appendix D). Given the lengthy summary, we prompted the LLM to assign a binary label (yes/no) to assess individual summary sentences' consistency with the original article. Then, we reported the percentile of "yes"

ID	Evaluation Model	AGGREFACT		DIVERSUMM						LSV		LONGEVAL
		XSM_{AG}	CND_{AG}	MNW	QMS	GOV	AXV	CSM	*Macro-AVG*	PUB	AXV	PUB
	evaluation metric	*AUC*		*AUC*						*Kendal's τ*		*Kendal's τ*
	avg src. len	360.54	518.85	669.20	1138.72	2008.16	4406.99	4612.40	–	3776.80	6236.40	3158.35
Baselines												
1	**LONGDOCFACTSCORE**	50.47	65.27	61.20	40.69	83.52	65.36	60.06	62.17	61.0	61.0	29.0
2	**MINICHECK-FT5**	75.04	72.62	48.68	45.31	70.26	61.77	52.93	55.79	26.5	38.1	17.4
3	**GPT4o**	75.36	70.47	51.11	70.22	86.81	67.78	61.53	67.49	54.7	51.8	51.2
*Apply our approach with different **baselines**(↑ means improved the performance compared to the baseline with significance.)*												
4	**ALIGNSCORE**	75.66	69.50	46.74	56.48	87.02	77.46	61.03	65.75	54.9	53.9	36.9
5	+ re-weighting	75.67	69.20	45.33	53.95	87.29↑	81.15↑	60.55	65.65	53.0	54.3↑	34.8
6	+ LV1 SEGMENT	76.23↑	69.25†	45.86†	61.25↑	86.74†	79.47↑	64.15↑	67.49↑	51.9	52.8	43.6↑
7	STRUCTS-LV1	76.20↑	69.03	46.21†	60.06↑	86.04	82.78↑	64.47↑	67.91↑	50.4	53.9†	43.4↑
8	+ LV2 SEGMENT	74.27	70.30↑	46.03†	55.74	85.10	76.79	63.11↑	65.35	58.1↑	51.1	43.9↑
9	STRUCTS-LV2	74.28	69.85↑	45.33	51.86	85.65	80.00↑	63.59↑	65.29	55.3↑	54.1↑	43.7↑
10	**MC-FT5 (SENT)**	79.62	70.95	57.67	60.66	83.24	78.66	59.74	67.99	55.7	52.7	30.2
11	+ re-weighting	79.73	70.76†	56.79	60.36†	84.75↑	79.38↑	60.06↑	68.27↑	52.8	55.1↑	31.4↑
12	+ LV1 SEGMENT	77.84	73.48↑	44.80	61.10↑	87.50↑	85.22↑	63.59↑	68.44↑	57.5↑	51.4	33.0↑
13	STRUCTS-LV1	76.75	73.40↑	38.45	60.66†	88.05↑	86.32↑	63.11↑	67.31	56.2↑	53.8↑	30.7↑
14	+ LV2 SEGMENT	73.70	72.30↑	47.80	57.53	86.26↑	83.73↑	62.07↑	67.48	56.0↑	52.9↑	35.6↑
15	STRUCTS-LV2	71.31	72.30↑	41.27	59.02	87.16↑	84.78↑	61.75↑	66.80	53.4	54.2↑	33.0↑
16	**INFUSE**	68.48	72.52	54.14	39.64	84.41	68.13	57.82	60.83	59.4	55.9	36.9
17	+ re-weighting	67.30	72.37	53.44	40.54↑	84.68↑	74.31↑	59.82↑	62.56↑	58.3	56.3↑	34.6

Table 4: Results for all summarization tasks in AGGREFACT-FTSOTA (AGGREFACT), DIVERSUMM, LONGSCIVERIFY (LSV) and LONGEVAL on Pubmed. For AGGREFACT, we report the overall ROCAUC on XSum and CNN/DM. respectively. In DIVERSUMM, CSM, MNW, QMS, AXV, and GOV refer to ChemSum, MultiNews, QMSUM, ArXiv, and GovReport. We also report the macro-average of DIVERSUMM AUC. We highlight the best performed approach where multiple greens indicate systems indistinguishable from the best according to a paired bootstrap test with p-value < 0.05, and the second-best system for each column. The six baseline models are **bolded**. Cells with † mean the result is indistinguishable from the raw baseline according to the bootstrap test. We report the average of 3 runs for GPT4o, given the randomness in LLM inference.

answers as the summary-level rating. Unless especially noticed, we reran the baseline models on our datasets using the original authors' released codebase and checkpoints. Implementation details can be found in Appendix D.

Our Approach We re-utilized baseline models to compute the scores between context chunks and summary sentences, including ALIGNSCORE (Zha et al., 2023), MINICHECK-FT5 (SENT) and INFUSE (Zhang et al., 2024), and experimented with below settings to apply our proposed approaches:

- + re-weighting: we apply the discourse-inspired re-weighting algorithm to adjust the sentence-level scores. We tune the factor α on height-subtree weighting as 1 over the validation set of DIVERSUMM and apply it to other benchmark datasets.
- + LVN. SEGMENT: Instead of using the default chunking approach, we segmented the source documents with the algorithms introduced in Sec. 5.2 with different levels of granularity.
- STRUCTS-LVN: Combining top two methods.

The reweighting and segmentation can not be applied to LONGDOCFACTSCORE, as it produced negative scores on all enumeration of source-target sentence pairs, which does not utilize the structural information. INFUSE utilizes the ranked list of entailment scores for all document sentences associated with each summary sentence. Thus, the segmentation approach does not affect.

Evaluation Metrics For experiments with AGGREFACT-FTSOTA and DIVERSUMM, following (Laban et al., 2022; Zhang et al., 2024), we adopt ROCAUC (Bradley, 1997) which measures classification performance with varied thresholds as our evaluation metric.[7] On LONGSCIVERIFY and LONGEVAL, we report Kendall's Tau τ, following the original paper (Bishop et al., 2024).

7 Results

Overall Performance Table 4 presents our main results with detailed setups. Overall, our pro-

[7]To determine the statistical significance of performance differences, following Zhang et al. (2024), we randomly re-sample 70% of the test instances 100 times and evaluate the models on these sets.

posed approach (with different combinations of re-weighting and segmentation settings) achieves the best or second best across AGGREFACT and DIVERSUMM. On LONGEVAL-PUB, excluding the top-performed GPT4o model, our approaches surpassed the other non-LLM baselines, with a score of 43.9 (row 8) compared to 36.9 (row 4 and row 16). The rest of the section addresses the following research questions: **RQ1:** Can the re-weighting algorithm help improve the models' performance? **RQ2**: How does source document segmentation impact factual inconsistency detection? **RQ3**: How does combining both in STRUCTSCORE perform?

RQ1. *We observe that the re-weighting algorithm improves prediction performance on different baselines (rows 4-5, 10-11, 16-17).* For long source documents, the re-weighting approach consistently improves or closely matches performance on GOV, AXV, CSM, and LSV-AXV. On the other hand, for both XSM and CND, the re-weighting algorithm does not help much. We posit that the short summary length (1-3 sentences) has minimally structured information, so the scores will not change much from the baseline. For MNW and QMS, the short summaries in QMS (averaging 3 sentences) reduce the effectiveness of the re-weighting algorithm. Moreover, MNW's non-factual sentences often receive high prediction scores, which our re-weighting approach tends to amplify, leading to a drop in performance. We also observe a slight performance drop on LSV-PUB and LongEval-PUB for ALIGNSCORE and INFUSE, potentially due to the different document structure of scientific articles from the medical domain. These observations also suggested potential future work for a dynamic weighting algorithm based on the document structure and domain knowledge. In Table 5, we ablate the two discourse factors from the re-weighting algorithm with our best baseline MC-FT5 (SENT) on a subset of long datasets. We noticed that both features are helpful, and the improvement in adding subtree height is greater.[8]

RQ2. *We find that applying document and discourse-structure-inspired approaches enhances performance across different baselines on long document summarization tasks.* We start by applying the level-1 and level-2 segmentation to preserve the document structures while segmenting at higher levels. For example, MC-FT5 (SENT) with LV1 SEGMENT obtains the highest macro-average

[8]We include a more complete table in Appendix E.

Model	GOV	AXV	CSM	LSV-AXV
MC-FT5 (SENT)	83.24	78.66	59.74	52.73
+ *subtree height*	84.55	79.09	60.55	55.08
+ *depth score*	83.65	78.90	59.90	53.80
re-weighting	84.75	79.38	60.06	55.08

Table 5: Ablation results on a subset of datasets from DIVERSUMM and LONGSCIVERIFY, the top and bottom rows are rows 10 and 11 in Table 4 .

AUC on DIVERSUMM, a trend also observed with ALIGNSCORE. Specifically, comparing row 10 and row 12, the Lv1 SEGMENT improved the model's performance on 6 of 7 long datasets from QMS to LongEval-PUB (i.e. 78.66 -> 85.22 and 83.24 -> 87.50 on AXV and GOV from DIVERSUMM). However, the effect of fine-grained segmentation can vary depending on the document's length and structure. For instance, ALIGNSCORE in row 8 with Lv2 segment obtained better performance than Lv1 on LSV-PUB but was the worst on QMS.

RQ3. *Combining both approaches is not universally beneficial across all scenarios.* When both individual approaches contribute positively, the combined STRUCTS generally achieves better performance, as seen in row 13 and row 7 on AXV and CSM. However, when one component causes a performance drop, combining both often leads to weaker overall performance than the stronger component alone. For instance, on GOV, row 7 performs worse than row 4, likely due to the segmentation in row 6, making the model less accurate. Similarly, row 13 performs slightly better than row 10 on LSV-PUB, but row 12's improvement does not translate into better performance gains when combined with row 11. Differences in evaluation metrics (AUC vs. correlation) and dataset sizes may also have influenced these outcomes (i.e., row 13 does not improve much on LE-PUB while both rows 11 and 12 have larger gains).

8 Conclusion

In this work, we approach the factual inconsistency detection of long document summarization through the lens of discourse analysis. We find that discourse factors, with regard to sentence structure, are related to the factual level of sentences. We further propose a framework that leverages the source document structure and introduces re-weighting the sentence-level predictions on top of different NLI-based models to obtain performance gains on multiple long document summarization datasets.

Limitations

Our work contributed to understanding the unfaithful errors in machine-generated summaries from the lens of discourse analysis. Our experiments' validity and subsequent findings rely on the parsed discourse trees generated by an existing parser, following prior work (Adams et al., 2023a; Pu et al., 2023; Kim et al., 2024b). It is important to note that parsed results may also be suboptimal given the challenges of complex hierarchical structures of long documents and the differences between the model's training corpora and our tested domains. We call for more robust RST parsers that can leverage recently contributed annotated discourse corpora with the help of advances in LLM modeling.

Our current approach leaves discourse-relation information unused on the system level; it would be interesting to utilize it to detect and resolve inconsistency errors. We also acknowledge the choices of our current re-weighting algorithm (exponential) can be further studied with more motivation.

In our analysis section, discourse analyses were carried out using the annotated portion of the released dataset, which is limited by the annotation quality and the dataset sizes. Yet, this is by far the only dataset that provides the sentence-level annotations on long document summarizations (i.e., Krishna et al. (2023) released the fine-grained scores, but did not clarify how the spans annotations are collected in their document). We verify the effectiveness of portions of our linguistic-inspired method on other benchmarks, including LONGSCIVERIFY and LONGEVAL. Future work would be to analyze and examine the discourse patterns in other domains, such as story summarization or further book-length summarization tasks (Chang et al., 2024; Kim et al., 2024a).

Ethical Statement

Throughout the paper, we have referenced datasets and models used in our analyses and experiments, ensuring that they are openly available and do not pose concerns with the public release or usage of this paper. We acknowledge the use of Grammarly and ChatGPT-4o for correcting sentences that are less fluent but not for generating or drafting new content.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad, and Kathleen McKeown. 2023a. Generating EDU extracts for plan-guided summary re-ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2680–2697, Toronto, Canada. Association for Computational Linguistics.

Griffin Adams, Bichlien Nguyen, Jake Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan Naumann, and Noémie Elhadad. 2023b. What are the desired characteristics of calibration sets? identifying correlates on long form scientific summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10520–10542, Toronto, Canada. 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.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

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.

Vanessa Wei Feng. 2015. *RST-style Discourse Parsing and Its Applications in Discourse Analysis*. Phd thesis, University of Toronto, Toronto, Canada.

John Glover, Federico Fancellu, Vasudevan Jagannathan, Matthew R. Gormley, and Thomas Schaaf. 2022. Revisiting text decomposition methods for NLI-based factuality scoring of summaries. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–105, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Freya Hewett and Manfred Stede. 2022. Extractive summarisation for German-language data: A text-level approach with discourse features. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 756–765, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

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.

Patrick Huber, Wen Xiao, and Giuseppe Carenini. 2021. W-RST: Towards a weighted RST-style discourse framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3908–3918, Online. Association for Computational Linguistics.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland. Association for Computational Linguistics.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024a. Fables: Evaluating faithfulness and content selection in book-length summarization. *arXiv preprint arXiv:2404.01261*.

Zae Myung Kim, Kwang Hee Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024b. Threads of subtlety: Detecting machine-generated texts through discourse motifs. *ACL2024*.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022a. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55:1 – 35.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022b. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, 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.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 1998. To build text summaries of high quality, nuclearity is not sufficient. *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization.*

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kenji Ono, Kazuo Sumita, and Seiji Miike. 1994. Abstract generation based on rhetorical structure extraction. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action - Proceedings of the 1st European Conference on Argumentation*, volume 2, pages 801–816.

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 and virtual meeting. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. 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. *arXiv preprint arXiv:2404.11184*.

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.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9709–9716.

A Discourse Analyses

A.1 Short Summary Analysis

Dataset	Size	Gran	Error Tag
AGU_CLIFF	300	word	intrin./extrin./other/wld. knowl.
AGU_Goyal'22	150	span	intrins./extrin./other

Table 6: Statistics of Sent/Span-level factual inconsistency datasets AGGREFACT-UNIFIED (AGU) (Tang et al., 2023). We report the size of doc-summary pairs (Size), the granularity of annotation (Gran), and the error labels (Error Tag).

We also conduct a discourse analysis on AGGREFAC-UNITED (Tang et al., 2023), as shown in Table 6. This dataset includes BART and Pegasus summaries from CLIFF (Cao and Wang, 2021) and Goyal'21 (Goyal and Durrett, 2021).[9] In the Goyal22 split of AGGREFACT-UNITED, a total of 61 errors were detected. Intrinsic errors are found to appear more often in satellite EDUs (18/31) with the attribution relation. Regarding extrinsic errors, the nucleus EDUs take the majority. We further analyzed the CLIFF dataset (Cao and Wang, 2021), where span-level annotations of faithful errors are available. Out of 600 sentences, the parser failed to parse 131 summaries, likely due to their short lengths and simplistic structures. Therefore, our analysis focused on the 469 summaries that were successfully parsed. We observed that Elementary Discourse Units (EDUs) containing errors are more likely to appear at the bottom of the discourse tree. These findings are similar to the long summary analysis in §4.

A.2 Discourse Relations in RST

We include the complete list of coarse-grained and fine-grained relation classes in the RST Discourse Treebank in Table 7, as summarized in (Feng, 2015).

B Discourse Analysis on Fine-grained Error Types

Error Types Relation Error (PreE) is when the predicate in a summary sentence is inconsistent with respect to the document. Entity Error (EntE) is when the primary arguments of the predicate are incorrect. Circumstance Error (CircE) is when the predicate's circumstantial information (i.e., name

[9] AGGREFACT-UNIFIED (AGU_CLIFF) include additional error types such as *comments*, *other errors: noise, grammar* and *world knowledge* (wld. knowl.)

Relation class	Relation type list
ATTRIBUTION	*attribution, attribution-negative*
BACKGROUND	*background, circumstance*
CAUSE	*cause, result, consequence*
COMPARISON	*comparison, preference, analogy, proportion*
CONDITION	*condition, hypothetical, contingency, otherwise*
CONTRAST	*contrast, concession, antithesis*
ELABORATION	*elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition*
ENABLEMENT PURPOSE	*purpose, enablement*
EVALUATION	*evaluation, interpretation, conclusion, comment*
EXPLANATION	*evidence, explanation-argumentative, reason*
JOINT MANNER-MEANS	*disjunction* *manner, means*
TOPIC-COMMENT	*problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question*
SUMMARY	*summary, restatement*
TEMPORAL	*temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence*
ELABORATION TOPIC-CHANGE	*elaboration-additional, elaboration-general-specific,* *topic-shift, topic-drift*

Table 7: The 17 coarse-grained relation classes and the corresponding 78 fine-grained relation types (53 mononuclear and 23 multi-nuclear) in the RST Discourse Treebank. Relation types which differ by nuclearity only, e.g., contrast (mononuclear) and contrast (multi-nuclear), are combined into one single type name here. Table replicated from (Feng, 2015).

RST features Count	GramE (83)	LinkE (35)	OutE (48)	EntE (117)	PredE (15)	CorefE (9)	CircE (13)	ALL Errors (320)
Ono penalty	-1.166	1.855	0.621	1.647	0.730	0.215	1.627	1.606 (0.1089)
Depth score	-5.218**	-7.381**	-4.628**	-3.252**	-2.002	0.214	-0.565	-8.249 (0.0000)
Promotion score	-6.519**	-0.971	-0.440	1.734	-0.195	2.613*	0.629	-0.828 (0.4083)
Normalized penalty	-1.742	3.051**	0.695	1.990*	0.673	-0.002	0.493	2.160 (0.0314)
Normalized depth score	-6.689**	-6.043**	-4.823**	-3.307**	-1.731	-0.153	-1.986	-9.084 (0.0000)
Normalized promotion score	-5.754**	0.487	-0.322	1.796	-0.087	2.206	-0.218	-0.303 (0.7617)

Table 8: Two-sided t-test statistic of significant RST-based features comparing unfaithful sentences to faithful ones in DIVERSUMM annotated split. We report the test statistics and significance levels. For fine-grained errors, we report the significant level in * (0.01 <= p-value <=0.05) and ** (p-value <=0.01). For All errors, we report the p-value in parenthesis.

or time) is wrong. Co-reference error (CorefE) is when there is a pronoun or reference with an incorrect or non-existing antecedent. Discourse Link Error (LinkE) is when multiple sentences are incorrectly linked. Out of Article Error (OutE) is when the piece of summary contains information not present in the document. Grammatical Error (GramE) indicates the existence of unreadable sentences due to grammatical errors.

Fine-grained Error Analysis In Table 8, we demonstrate the breakdowns of fine-grained error types and report the t-test results on different discourse features.

C Example of Segmentation Failures

This section includes one example of the ALIGNSCORE's chunking method that failed to preserve the document structure, while our discourse-inspired chunk addresses it.

For example, as shown in Figure 3a, the original document contains two consecutive sentences: "To determine the extent ..." and "To develop the SMS" (highlighted in the orange box). These sentences are meant to be read together and should not be separated. However, the default chunking approach in ALIGNSCORE and MINICHECK breaks this continuity by placing them in two separate chunks, given the former chunk is large enough. On the contrary, our approach maintains the structural integrity of the documents, keeping the sentences connected as intended. Similarly, in Figure 3b, the conclusion section is separated into two chunks by the default chunking approach, while our method maintains them in a single chunk.

D Implementation Details

D.1 GPT4o Prompts

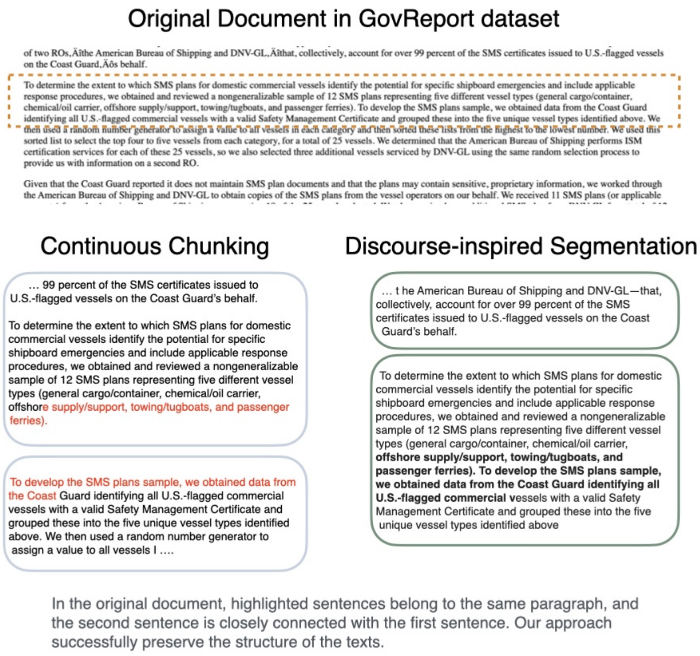
We include our prompt for zero-shot factual consistency evaluation in Table 9.

D.2 Baselines

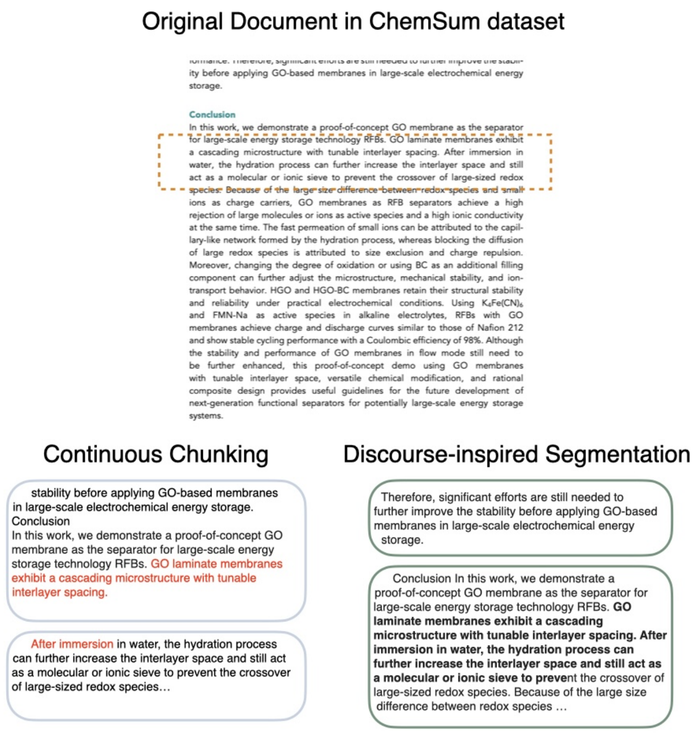
AlignScore (model size 355M) (Zha et al., 2023) is an entailment-based model that has been trained on data from a wide range of tasks such as NLI, QA, and fact verification tasks. It divides the source document into a set of sequential chunks at sentence boundaries. For a multi-sentence summary, it predicts the max scoring value of all combinations of source chunk and target sentence, then returns the unweighted average of all sentences as the summary prediction. We follow the original setting by setting chunk size at 350 tokens and use the default model alingsocre_large ckpt. The model outputs a score between 0 and 1. We conduct experiments on top of their released codebase `https://github.com/yuh-zha/AlignScore`.

MiniCheck-FT5 (model size 770M) (Tang et al., 2024) is an entailment-based fact checker built on flan-t5-large. It has been further fine-tuned on 21K datapoints from the ANLI dataset (Nie et al., 2020) and 35k synthesized data points generated in (Tang et al., 2024) on the tasks to predict whether a given claim is supported by a document. We follow the authors's setting and set the chunk size to 500 tokens using white space splitting. The output score is between 0 and 1. We use the released code repo from `https://github.com/Liyan06/MiniCheck`.

LongDocFactScore (Bishop et al., 2024) is a reference-free framework for assessing factual consistency. It splits source documents and the generated summary into sentences, then computes the pair-wise similarities by computing the cosine



(a) Example from GovReport of DIVERSUMM.



(b) Example from ArXiv of DIVERSUMM.

Figure 3: Example of segmentation failures, left is the output of chunking method used in ALIGNSCORE and MINICHECK, right is the segments produced by our segmentation method.

similarities of sentences (they use the sentence-transformers library initialized with the bert-base-nmli-mean-tokens model). Afterward, for each individual summary sentence, K most similar source sentences are picked. The method extracts the neighboring source document sentences of the selected sentences as context, then applies a metric BARTScore to evaluate the score between source context and summary sentences. The overall summary score is an unweighted average of all sen-

```
Determine whether the provided claims are consistent with the corresponding document. Consistency in this context
implies that all information presented in the claim is substantiated by the document. If not, it should be
considered inconsistent.

Document: [DOCUMENT]
Claims: [CLAIMS]
Please assess the claim's consistency with the document by responding with either "yes" or "no".
The CLAIMs are ordered in the format of a dictionary, with { index: CLAIM }. You will need to return the result in JSON format.
For instance, for a CLAIMs list of 4 items, you should return {0:yes/no, 1:yes/no, ...., 3:yes/no}.

ANSWER:
```

Table 9: Zero-shot factual consistency evaluation prompt for GPT4o.

tences. We follow the authors' parameters setting and utilize their released code repo from https://github.com/jbshp/LongDocFACTScore.

InfUsE (model size 60M) Zhang et al. (2024) uses a variable premise size and breaks the summary into sentences or shorter hypotheses. Instead of fixing the source context, it retrieves the best possible context to assess the faithfulness of an individual summary sentence by applying an NLI model to successive expansions of the document sentences. Similar to prior approaches, it outputs an entailment score for each summary sentence, and the summary-level score is the unweighted average. We follow their settings on INFUSE with summary sentences instead of $\text{INFUSE}_{\text{SUB}}$ as the authors only released the code for the former model. INFUSE outputs scores in the range 0-1. We use the author's released codebase from https://github.com/HJZnlp/Infuse.

GPT4o We used the version of gpt-4o-2024-05-13; we set max_tokens 100, sampling temperature at 0.7, and top_p as 1.0. We call the OpenAI API from https://openai.com/api.

D.3 Machine Configuration for Models

We use up to 4 NVIDIA RTX 5000 GPUs, each equipped with 16 GB VRAM, for model inferences on our hardware. According to Lambda[10] (RTX5000 is depreciated), a single NVIDIA Quadro RTX 6000 (the closest to our setting) GPU costs $0.5 per hour and has 24 GB VRAM.

E Ablation Study

Table 10 presents the ablation results of different discourse features on our baselines. We cover the long document summarization tasks starting from QMS in Table 4.

[10]https://lambdalabs.com/service/gpu-cloud

Model	**QMS**	**GOV**	**AXV**	**CSM**	**LSV-PUB**	**LSV-AXV**	**LE-PUB**
MC-FT5 (SENT)	60.66	83.24	78.66	59.74	55.7	52.7	30.2
+ *subtree height*	60.21	84.55	79.09	60.55	53.6	55.1	30.4
+ *depth score*	60.51	83.65	78.90	59.90	55.7	53.8	33.3
re-weighting	60.36	84.75	79.38	60.06	52.8	55.1	31.4
AlignScore	56.48	87.02	77.46	61.03	54.9	53.9	36.9
+ *subtree height*	52.91	87.29	81.15	60.47	51.7	55.4	34.1
+ *depth score*	56.63	87.29	77.66	60.30	54.3	52.4	36.6
re-weighting	53.95	87.29	81.15	60.55	53.0	54.3	34.8

Table 10: Ablation results on long document datasets from DIVERSUMM, LONGSCIVERIFY and LONGEVAL.